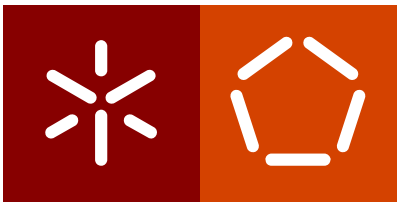
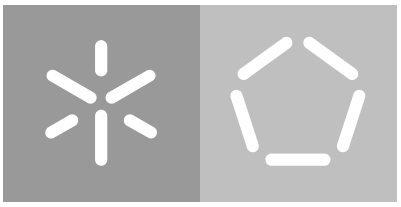


Universidade do Minho
Escola de Engenharia
Departamento de Informática

Maria Inês Alves Faria

Multi-view learning for multiomics data integration for the study of plants

February 2022



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Maria Inês Alves Faria

Multi-view learning for multiomics data integration for the study of plants

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Professor Doutor Oscar Dias

February 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Maria Inês Alves Faria

ACKNOWLEDGEMENTS

Após mais um ano de pandemia, mas também de muito esforço e dedicação, dou por terminada a minha dissertação, à qual não seria capaz de realizar senão fosse pelo contributo e paciência de algumas pessoas. Dessa forma, gostaria de agradecer a todos os que me ajudaram ao longo desta jornada.

Primeiramente, gostaria de agradecer ao Professor Doutor Óscar Dias, que acreditou nas minhas capacidades propondo-me esta oportunidade, e por todos os ensinamentos, e correções ao longo do caminho. Agradeço também à Marta Sampaio, que nunca hesitou em me ajudar em todas as minhas dúvidas e a me guiar ao longo deste processo, motivando-me a fazer um bom trabalho e a exigir mais de mim.

A todos os meus colegas no escritório que me ensinaram a ter um bom equilíbrio entre trabalho árduo, lazer e convívio. Aos meus amigos, que apesar do *covid-19* mantiveram o contacto e partilharam comigo bons momentos, mas também os maus momentos da tese.

Finalmente, aos que me são mais próximos, queria agradecer especialmente aos meus pais, que me apoiaram em tudo, dando-me todas as condições necessárias para fazer um bom trabalho, acreditando sempre em mim. À minha irmã, que sempre disponibilizou o seu tempo para verificar se estava tudo bem comigo e se precisava um ombro amigo com quem falar. A todos os meus familiares, que estão sempre lá para mim. Ao Rafael, um obrigada por toda a alegria deste último ano e por nunca hesitares em me motivar para acabar a dissertação. Por fim, ao Jeremias, que apesar de me roer os cordões, também foi um fofo e aliviou-me o stress.

A todos, muito obrigada!

ABSTRACT

Plants are indispensable for human life and have a significant impact on the economy. Their growth and survival are linked to their metabolism, and its study is important to understand certain mechanisms and responses to different environmental stresses. To enhance plant Genome-Scale Metabolic models, that are used in systems biology to study metabolism, several methods were created to integrate gene expression within the models, resulting in more realistic flux predictions. Therefore, the integration of multiple omics is essential to identify complex biological relationships that may become evident only through the combination of multiple omics data. However, the different sizes, formats and scales of the data being integrated, as well as the different complexities, noisiness, contents, and levels of agreement hinder this task.

Hence, in this work, a pipeline was developed, including Machine Learning (ML) methods to integrate different omics data and extract knowledge on plant behaviour under different environmental conditions. Three different multiomics integration approaches were studied: concatenation-based integration (CBI), transformation-based integration (TBI) and model-based integration (MBI). The models inspected for CBI were DIABLO, SMSPL, Stacked Generalisation, Lasso Regression, Support Vector Machine, Random Forest and Artificial Neural Networks. For TBI, we analyzed SNFtool, Graph-Composite Association Network and Kernel-Relevance Vector Machine. Regarding the MBI, we created an ensemble classifier. All models were tested and cross-validation was executed. The models were created and validated using two different datasets of *Vitis vinifera* and *Arabidopsis thaliana*, for Case Study I and II, respectively. CBI was the most studied strategy, with several models available and easy implementation. DIABLO offered innovative plots to visualize the data correlations, provided the most relevant features to predict the outcome, had a good performance, but takes a considerable running time. SMSPL thanks to its novel strategy offered good performance and the most important features. For the TBI, the SNFtool was the single method capable of identifying the most relevant features, but all were very efficient models and easy to implement. Lastly, MBI was the approach with fewer methods available and harder to implement. Soft voting obtained better results compared to hard voting and obtaining the most relevant features was a difficult task.

The pipeline was successfully created and can be identified in the open-source framework https://insilicoplants.pt/, or the GitHub repository https://GitHub.com/InesFaria-UM/Master_Thesis.git.

Keywords – Multiomics Integration, V. vinifera, Concatenation-Based, Transformation-Based, Model-Based, Machine Learning.

RESUMO

As plantas são indispensáveis à vida humana e têm um impacto significativo na economia. O seu crescimento e sobrevivência estão ligados ao seu metabolismo, cujo estudo é importante para compreender certos mecanismos e respostas metabólicas a diferentes stresses ambientais. A biologia de Sistemas dedica-se a este estudo usando modelos metabólicos à escala genómica (GSM). Para aprimorar os modelos GSM de plantas, vários estudos foram criados para integrar expressão genética nos modelos metabólicos, de modo a obter previsões mais realistas. Desta forma, é fundamental integrar múltiplos dados ómicos para identificar relações biológicas complexas que, até ao momento, não são evidentes. Contudo, os diferentes tamanhos, formatos e escalas dos dados a ser integrados, bem como as diferentes complexidades, barulhos, conteúdos e níveis de concordância dificultam esta tarefa.

Assim, neste trabalho, foi concebida uma pipeline usando métodos de aprendizagem máquina, a fim de integrar diferentes dados ómicos e extrair conhecimento em relação ao comportamento da planta sob diferentes condições ambientais. Três diferentes abordagens de integração multiómica foram estudadas: integração baseada em concatenação (CBI), integração baseada em transformação (TBI) e integração baseada em modelos (MBI). Os métodos discutidos para CBI foram DIABLO, SMSPL, Stack Generalisation, Lasso Regression, Support Vector Machine, Random Forest e Artificial Neural Networks. Em relação a TBI, analisamos o SNFtool, Graph-Composite Association Network e Kernel-Relevance Vector Machine, e para o MBI, criamos um ensemble classifier. Todos os modelos foram testados e submetidos a validação cruzada. Os modelos foram validados usando dois conjuntos de dados diferentes de *Vitis vinifera* e *Arabidopsis thaliana*, como caso de estudo I e II. A CBI foi a estratégia mais estudada, com diversos modelos disponíveis e de fácil implementação. O método DIABLO, apesar de ter um maior tempo de execução, ofereceu formas inovadoras de visualizar as correlações dos dados e as variáveis mais relevantes para prever o resultado, garantindo um bom desempenho. Já o SMSPL obteve um bom desempenho e indicou as features mais importantes. Na TBI, o SNFtool foi o único método capaz de identificar as variáveis mais relevantes. No entanto, todos os métodos TBI foram eficientes e de fácil implementação. Por fim, a MBI foi a abordagem com menos métodos disponíveis e mais desafiante de implementar. A votação *soft* obteve melhores resultados em comparação com a votação *hard*, porém, as variáveis mais relevantes foram difíceis de obter.

A pipeline foi criada com sucesso e pode ser encontrada na *"open-source framework"* https://insilicoplants.pt/, ou no repositório GitHub https://GitHub.com/InesFaria-UM/Master_Thesis.git.

Keywords – Integração Multiómica, V. vinifera, Integração Baseada em Concatenação, Integração Baseada em Transformação, Integração Baseada em Modelos, Aprendizagem Máquina.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

GSM	Genome-scale Metabolic	40
TCA	Tricarboxylic Acid Cycle	5
ATP	Adenosine Triphosphate	5
KEGG	Kyoto Encyclopedia of Genes and Genomes	10
ML	Machine Learning	iii
FBA	Flux Balance Analysis	35
NCBI	National Center for Biotechnology Information	10
UniProt	Universal Protein Resource	10
BRENDA	BRaunschweig Enzyme Database	10
TCDB	Transporter Classification Database	10
TAIR	The Arabidopsis Information Resource	10
PMN	Plant Metabolic Network	10
PGDB	Pathway/Genome Databases	10
EC	Enzyme Commission	11
TC	Transporter Classification	11
SRA	Sequence Read Archive	12
GEO	Gene Expression Omnibus	12
dbGaP	The Database of Genotypes and Phenotypes	12
SAGE	Serial Analysis of Gene Expression	12
MS	mass spectrometry	13
PODC	Plant Omics Data Center	13
PPDB	Plant Proteomics Database	13
PMDB	Plant Metabolome Database	13
PCA	Principal Component Analysis	viii
SVM	Support Vector Machine	viii
CART	Classification And Regression Tree	21
ANN	Artificial Neural Network	ix
RF	Random Forest	viii
miRNA	micro RNA	23

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Plants comprise one of the most prominent groups of living beings. These are multicellular autotrophic eukaryotes and use photosynthesis to obtain energy and food. Plants have an essential role in preserving human life and other living beings since they are responsible for the converting carbon dioxide into oxygen, maintaining the atmospheric balance. Additionally, humans depend on plants as a source of food, such as vegetables, cereals, pulses, fruits, sugar, coffee, spices or oil, but also for medicines, energy, fibre, and building materials [1]. Plants have a significant impact on the economy. A substantial amount of Portugal's economy is related to wine exportation (12th place), olive oil (19th place), fruits (45th place), and half of the world's cork (21st place) [2, 3]. Nonetheless, Portuguese grapevines, categorised as *Vitis vinifera*, in 2020 were expected to decrease by 3% in terms of wine production, mainly due to weather conditions and fungal infections [4].

This way, since survival and growth is virtually connected to metabolism, this area becomes fundamental for further knowledge on fruit production and metabolic responses to diseases and different environmental stresses.

A field that focuses on the study of metabolism is System's Biology, which aims to accomplish a system-level understanding of living systems in a complex and dynamic way with many interactions [5]. Genome-scale Metabolic (GSM) models reflect biological reality and elucidate the genotype-phenotype relation [6], allowing to perform simulations that provide a direct measure (fluxome) of the metabolic phenotype. Additionally, generic databases are also important in systems biology for plant metabolic data information as these allow users to understand the biological system's functions and utilities, for example, genes, metabolic pathways, and metabolites. *Vitis vinifera* genome and metabolic information is available in various databases such as *VitisNet* [7] and *GrapeCyc* [8]. However, knowledge on plant metabolic pathways is still very scarce due to challenges characteristic of higher organisms.For instance, lack of comprehensiveness in plant metabolic networks and missing components, as most proteins' function remains unknown. Secondly, besides photosynthesis and photorespiration that contribute to the complexity of metabolic networks, plants have several unidentified compartments. Additionally, there is also a vast diversity of plant cell and tissue types [9].

Therefore, to further extend our insight into plant metabolism and understand underlying mechanisms leading to an organism phenotype, it is possible to bridge the gap between genotype and phenotype through the integration of context-specific omics data. Thus, it is essential to integrate multiple omics

to identify complex biological relationships that may become evident only through the combination of multiple omics data [10, 11].

The rapid development of high-throughput technologies enabled the generation of large-scale omics data, including plant omics data. Omics data analyse a given biological function, at different levels, including the molecular gene level (genomics), the protein level (proteomics), and the metabolic level (metabolomics). However, omics data are often dispersed and lack standardisation and the different sizes, formats and scales of the data being integrated, and the different complexities, noisiness, contents, and levels of agreement, hinder this task [12].

Hence, the processing and interpretation of omics data requires appropriated tools, such as ML algorithms [13]. ML tools can identify patterns, select relevant features from large datasets and make inferences from the observed data without defining biological assumptions [14, 15]. ML can also be used to analyse the flux data predicted by the context-specific GSM models with other omics from high-throughput technologies to improve the predictions [15].

Finally, the existence of a centralised repository is fundamental to incorporate and organise plant data to ease the study of plant metabolism, develop new bioinformatic tools and integrate information not only from the plant databases but also all relevant omics data to grant the users the ability to compare, analyse and integrate information from several contrasting sources.

1.2 OBJECTIVES

This project's main goal is to develop methodologies to integrate multiple omics data and create new datasets to improve knowledge on plant's metabolic phenotypes and underlying products when facing environmental stresses and diseases. As a primary case study, will use the grapevine, *Vitis vinifera*, to support the procedures development and validation. Neverthless, such methodologies can be used for other species with economic value for our country, like *Quercus suber* (cork).

The following objectives will be pursued to accomplish this main goal:

- Review state-of-art methods for analysing, preprocessing and integrating omics data from different sources and studies employing ML approaches for studying plant metabolism.
- Collect relevant plant omics data that will be organised in an integrated repository.
- Preprocess multi-omics data.
- Develop methods and computational tools based on ML to integrate different omics data and extract knowledge to understand plant behaviour under different environmental conditions.
-Integrate all the collected data, developed tools and algorithms into an open-source computational framework.
- Apply and validate the tools using a case study associated with *Vitis vinifera* metabolism.
- Write dissertation.

1.3 REPORT OUTLINE

This report is structured as follows:

- **Introduction** chapter, where the context and motivation for this project are described, and further objectives and report outlined.
- **State Of The Art** chapter subdivided in the following sections:
  1. *Plant Metabolism* section, providing a resume on plant metabolism;
  2. *Vitis vinifera* section, where we discuss the basics of berry development;
  3. *Sources of Plant Metabolic Data* section, illustrating the principal resources available for metabolic data in the subsection "Databases" and the central plant omics databases and an overview of omics data in the "Omics Data" subsection.
  4. *Machine Learning* section, reviewing the importance of ML approaches for the study of omics data, the basic principles of ML and its application to biological data, in the subsections "Concepts in Machine Learning", "Types of Learning", "Model Evaluation" and "Model Selection", also examples of studies using ML for the study of plants in the "Machine Learning in Plants" subsection.
  5. *Integration of Multiomics Data* section, describing in detail the methods and approaches for multi-omics data integration in the "Dimension Reduction Approaches", "Network-based Approaches", "Bayesian Approach" and "Multiple Kernel Learning Approach" subsections, and applications to plant omics data in the "Integration of Multiomics Data in Plants" subsection. The last subsection "Combination of experimental omics and predicted fluxomics" focuses on this approach and presents corresponding literature.
- **Materials and Methods** chapter divided in:
  1. *Plant Data Collection*, that explains both Case Studies, and where the datasets were taken from;
  2. *Pre-processing*, with all the preprocessing steps executed;
  3. *Feature Selection*, explaining the different filters used in the datasets;
  4. Models used in this project;
  5. *Model Evaluation*, the error metrics used and other forms of validation
  6. *Model Optimization*, selection of hyperparameters
  7. *Computational Framework*, indicating the programs used and where the code is available online.
- **Development**, explaining the pipeline developed, and all the scripts created. Divided in:
  1. *Plant Data Assimilation*
  2. *Pre-processing*
  3. Exploratory Analysis
  4. *Individual Omics Analysis*
  5. *Multiomics Integration*
- **Results and Discussion**, where the results are discussed and explained, by Case Study:

    1. *Case Study I*;
    2. *Case Study II*;
    3. Summary, with a brief explanation of the advantages and disadvantages of the executed models.

- Conclusions and Future Work
- **Supplementary Figures**, subdivided in:
    1. Case Study I;
    2. *Case Study II*.
- **Supplementary Tables**, for each Case Study:
    1. Case Study I;
    2. *Case Study II*.

2

STATE OF THE ART

2.1 PLANT METABOLISM

Metabolism is the sum of all biochemical reactions taking place in a living organism. These biochemical reactions are catalysed by enzymes and compose the metabolic pathways, where various intermediates, named metabolites, are involved. Enzymes connect reactions requiring energy input (converting simpler to more complex molecules, anabolism), with reactions that release energy (transforming complex substances into simpler molecules, catabolism) to biosynthesise new metabolites. Moreover, enzymes can regulate the rate of metabolic reactions according to internal signals and the changing environment [16].

Unlike animals, plants are sessile being exposed to rougher conditions and interacting with various pathogenic or beneficial organisms. Therefore, to defend themselves, plants evolved very complex metabolic networks capable of producing several metabolites essential for growth, development, reproduction and adaptation to the environment [17]. These metabolites differ from plant species and vary accordingly to the organ, tissue or cell at different developmental stages or under certain environmental conditions [18]. Another reason that makes plant metabolism even more complex is their immense compartmentalisation of the interconnected metabolic pathways [19]. Only recently it has become evident that plant metabolites have significant roles for the plants that produce them, with several beneficial aspects for the economy, for example, fragrances, stimulants, insecticides, attractants, antimicrobial, pharmaceutical, dyes, flavours and many more applications [20].

Plant metabolism can be sub-divided into two groups: primary (or central) and secondary (or specialised) metabolism [21]. Primary metabolites normally perform a physiological function in the organism, with fundamental roles related to normal growth, development, and plant's reproduction. This type of metabolite, including ethanol, lactic acid, and particular amino acids [22], is identified in many organism and cells.

Primary metabolism in plants starts with photosynthesis, converting light energy into chemical energy, which forms the sugars used to start cellular respiration. These sugars are then disintegrated in glycolysis and Tricarboxylic Acid Cycle (TCA) pathways, producing energy as Adenosine Triphosphate (ATP) molecules. Glucose can also be oxidised through the pentose phosphate pathway. Shikimate pathway, used for biosynthesis of folates and aromatic amino acids, is also included in the plants' central metabolism. All the products derived from these pathways will serve as precursors for the biosynthesis

of more complex compounds, like secondary metabolites, such as amino acids, fatty acids, starch and structural compounds forming the cell wall and membrane [23, 24].

On the other hand, secondary metabolites (or specialised metabolites) are not directly involved in the development, normal growth or reproduction of plants [16]. Secondary metabolites are usually detected in defined species, particularly tissues/organs at given developmental stages, or under certain environmental conditions, and have essential roles in protecting plants from herbivores and pathogen infection or to attract pollinators or seed dispersal animals [22]. Therefore, secondary metabolites are an integral part of species' interactions in plant and animal communities and their adaptation of plants to their environment.

Since they have a vast chemical diversity, they are used in a lot of pharmaceutical and biotechnological applications. These specialised metabolites are derived from the central metabolic pathways, like the TCA, the isoprenoid pathways, that produces isoprenoids like sterols and the shikimate pathway. Depending on their provenance, they can be divided, as demonstrated in figure 1, into three groups of metabolites: terpenoids, phenylpropanoids and alkaloids [21].

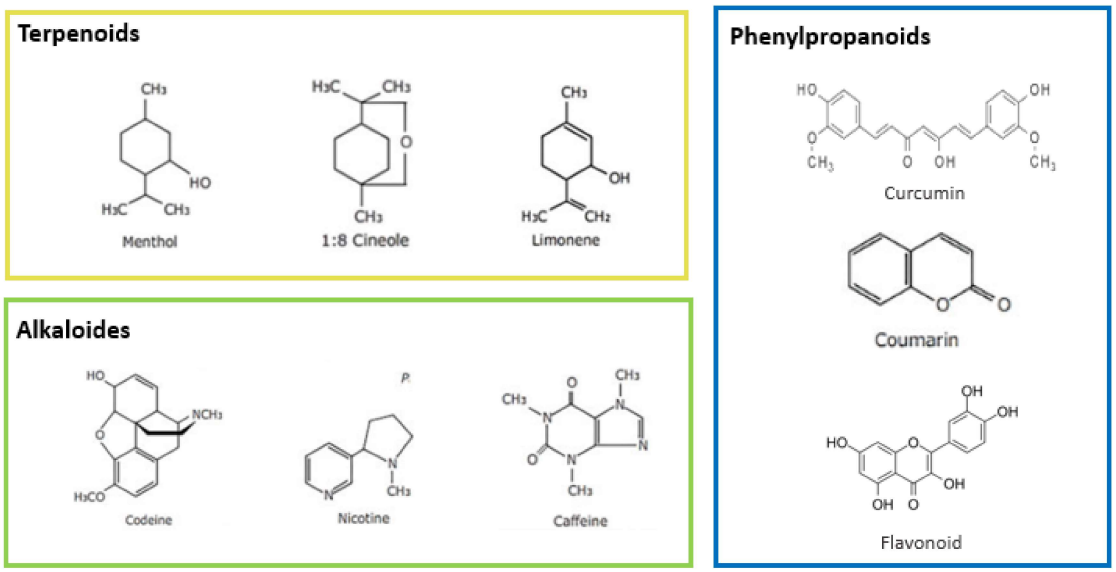


Figure 1: Illustrative compounds of the three groups of secondary metabolites. Terpenoids group: menthol, cinede and limonene; Alkaloids group: codeine, nicotine and caffeine; Phenylpropanoids group: curcumin, coumarin and flavonoid.

Terpenoids are derived from the five-carbon precursor isopentenyl diphosphate from the isoprenoid pathway and have expanded the range of aromas in the perfume industry and flavours of food additives. Phenylpropanoids are biosynthesised from amino acids produced in the shikimate pathway, like phenylalanine and tyrosine. One of the major classes of phenolic natural products are the flavonoids. These products are important in many aspects, not only being in charge of the colours of flowers and fruits, which often function to attract pollinators and seed disperses, but they can also protect plants against ultraviolet-B irradiation. On the other hand, alkaloids are derived from different amino acids

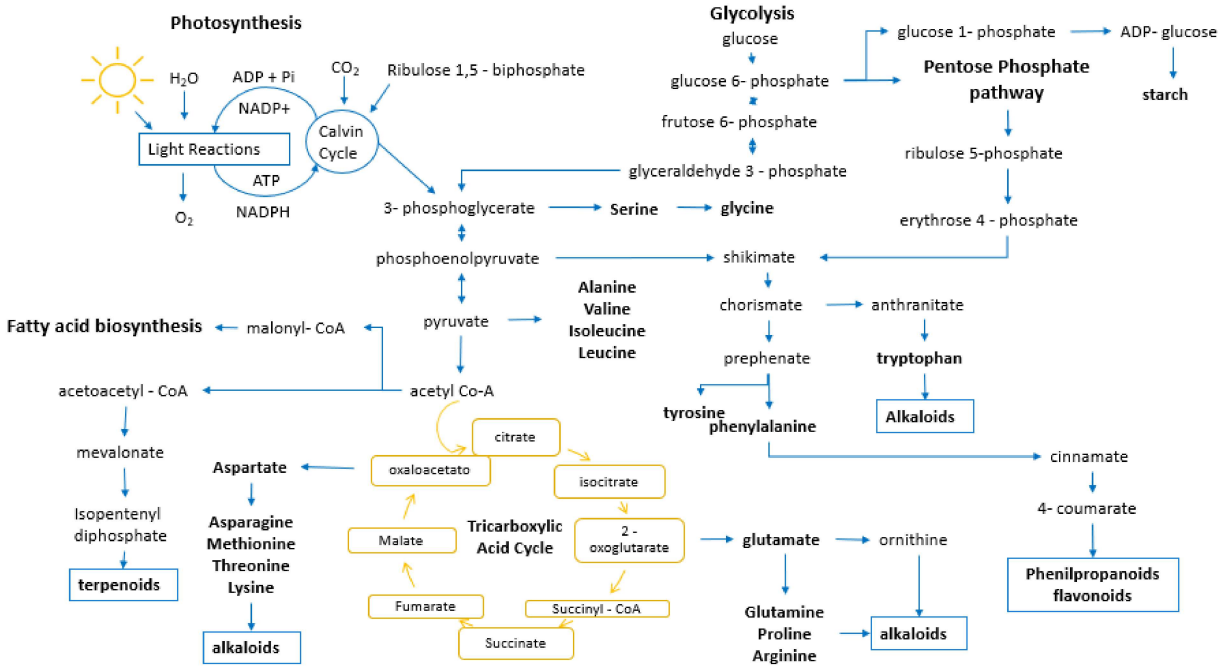


Figure 2: *Illustration depicting primary and secondary metabolism in plants.* Primary metabolism starts when light energy is transformed into chemical energy through photosynthesis and kept as sugar molecules, that will be used to start cellular respiration. The same molecules are then broken down during glycolysis and TCA cycle pathways to form ATP molecules. Furthermore, glucose can also be oxidised during the pentose phosphate pathway to generate reducing equivalents and the precursors for the biosynthesis of nucleotides and aromatic amino acids produced in the shikimate pathway. The shikimate pathway will allow the production of the two groups of secondary metabolites: alkaloids and phenylpropanoids. Lastly, acetyl-CoA can go into the TCA cycle, the fatty acid biosynthesis pathways or the isoprenoid pathway to produce terpenoids and other complex metabolites.

and are still used to this day as prescription drugs, including purgatives, antitussives, sedatives and treatments for a wide range of ailments.

Higher plants produce many secondary metabolites via complex pathways, which are regulated in highly sophisticated manners. In most cases, these bioactive natural compounds are located in particular organs, and their contents in such organs are seasonally regulated. However, they can also be translocated among various plant organs [25]. Figure 2 shows an overview of the central metabolism, which produces the precursors for secondary metabolite's biosynthesis.

Therefore, plants are essential beings that produce a vast diversity of bioactive compounds, with biotechnological and pharmaceutical importance, that promote the country's economy. Since they alter their metabolism to face the diverse environmental stresses around them, the study of plant metabolism is of uttermost importance to understand phenotypes of disease resistance and survival under extreme environmental conditions and improve the production of fruits or metabolites of interest.

2.2 Vitis vinifera

Vitis vinifera belongs to *Vitaceae*, a family of flowering plants and to the genus *Vitis*, that makes up the majority of species from Northern hemisphere. The *Vitis* genus is composed of two sub-genera: Muscadinia and *Euvitis*. *Vitis vinifera* belongs to the *Euvitis* sub-genera, where most of the cultivated grapevines belong. More specifically, to the Euroasian group, since it is native to the Mediterranean region, Central Europe and Southwestern Asia [26]. This plant species has high economic value since their grapes are used for many aspects: fresh fruit consumption, processed to make wine, vinegar or juice, or dried, to produce raisins [26].

2.2.1 *Berry Development*

Grape berries are composed of three distinct types of tissue: skin, flesh and seeds; and during their growth they suffer modifications mainly in size, content, texture, flavour and pathogen susceptibility. Berry development exhibits a double sigmoid growth pattern separated by a lag phase. The main indicators of berry growth start with cell division and then cell enlargement [27]. Literature splits berry growth into three stages [27] [28]:

- **Stage I:** *The first rapid growth phase*. Usually occurs between three to four weeks and immediately after flowering. In this phase the berry growth results both due to cell division as well as cell expansion, and the berry assumes a firm texture and green colour due to presence of chlorophyll. The sugar content is low, however organic acids accumulate, which contributes in some extent to berry expansion. The most prevalent organic acids that are present in this phase are *tartaric* and *malic* acids. *Tartaric* acid concentration is highest at the periphery of the developing berry and *malic* acid accumulates in the flesh cells. *Hydroxycinnamic* acids are also present in the first growth period, being a vital piece in many reactions and also due to their role as precursors of volatile phenols. Additionally *tannins*, also increase during this period and accumulate in the skin and seed tissues and other compounds, such as minerals, aminoacids, micronutrients, and aroma compounds also accumulate during this phase being responsible for the quality of the berry.
- **Stage II:** *The lag phase*. The duration depends on the type of cultivar and its end coincides with the end of the herbaceous phase of fruit (more or less two to three weeks). In this period berry growth slows down and the concentration of organic acids reaches its maximum level. The berry texture persists in its firm state but starts losing chlorophyll.
- **Stage III:** *The second rapid growth phase and fruit ripening*. After the lag phase another rapid growth phase takes place, which corresponds to the beginning of berry ripening. The duration of this phase is up to six to eight weeks and berry growth is restricted to cell enlargement, where berries can double size. This period is also known by the french word *veraison*, meaning *berry softening*, which is used to describe the initial stages of colour development, that symbolizes the start of ripening. In this period many dramatic changes occur in grape composition, including

the softening of the berry texture, the lose of chlorophyll (if the grape is a colored variety, red pigments start appearing and accumulating in the skin), the decrease in acidic content and increase in sugar concentration that is key for the accumulation of aroma and flavour compounds. During this phase most of the solutes remain, however due to the increase in berry volume their concentration is reduced, not only by dilution but to produce other compounds. It is the case of *malic* acid that is metabolized and used as energy in this phase. *Tannins* also decline as well as some aromatic compounds produced in the first rapid growth phase.

Figure 3 demonstrates the distinct berry growth phases as well as the different compounds present at the time.

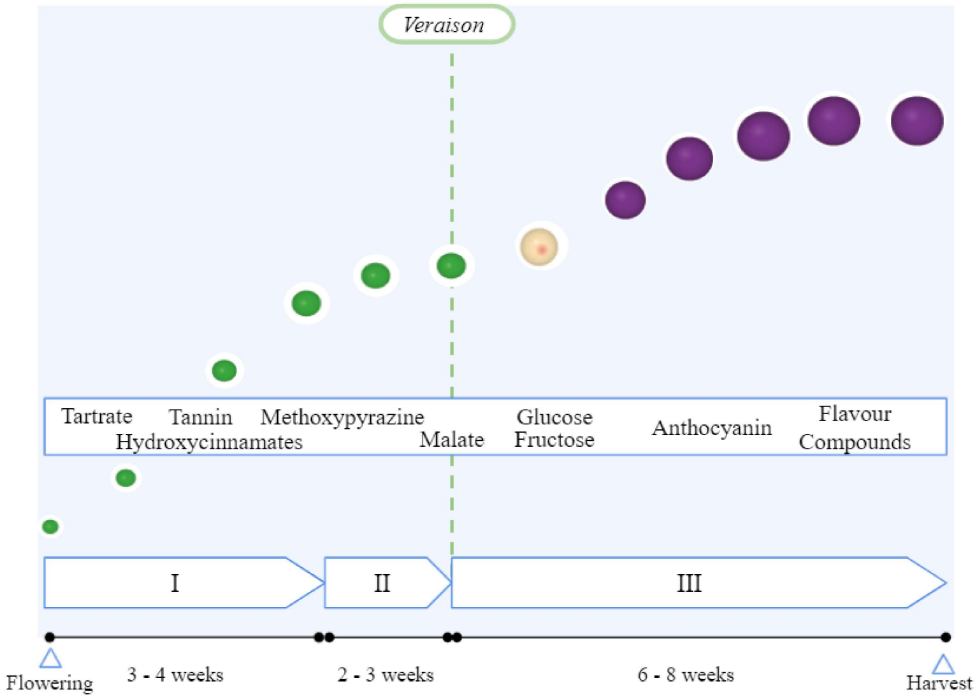


Figure 3: *Illustration of berry growth and different compounds present at each stage*. From flowering to harvest, the berry takes different sizes and colours. At stage I (first rapid berry growth phase), with a duration of 3 to 4 weeks, the berry gets bigger due to cell division and cell expansion, and it starts with a firm texture and green colour due to chlorophyll. At this point the main compounds present in the berry are organic acids. In stage II (the lag phase), spanning between 2 and 3 weeks, the berry growth slows down and the concentration of organic acids reaches its pick. Finally, in stage III (second rapid growth phase and fruit ripening) , a period of 6 to 8 weeks, it reaches veraison, the ripening phase, where the berry texture gets softer, and berry growth is restricted to cell enlargement. The berry also starts losing chlorophyll, and red pigments may appear if the grape is of a colored variety. Additionally, some organic acids are reduced and the sugar content increases, which contributes for the aroma and flavour of the grape.

2.3 SOURCES OF PLANT METABOLIC DATA

Plant metabolic data information can be obtained through generic databases, like the Kyoto Encyclopedia of Genes and Genomes (KEGG) [29]. These databases help understand the biological system's high-level functions and utilities, containing descriptions of metabolic pathways, genes, enzymes, reactions, and metabolites.

On the other hand, perception of plant metabolism can also be inferred by omics data. Omics technologies are defined as high-throughput biochemical assays, including, for example, transcriptomics, proteomics, epigenomic, and metabolomics data. These data are crucial to understand metabolism as they allow the detection and analysis of differential expression patterns in several environmental conditions, thus explaining the metabolic variations that occur in different phenotypes.

2.3.1 *Databases*

Large amounts of data relating to metabolic reactions are available in several databases that can be divided as general or species-specific. Table 1 shows the most pertinent databases for plant metabolism studies. MetaCyc [30] and KEGG [29] are the most generic to obtain information for metabolic pathways from the generic databases. Other generic databases used to extract detailed information on genomes, proteins, transporters, enzymes include the National Center for Biotechnology Information (NCBI) [31], the Universal Protein Resource (UniProt) [32], the BRaunschweig Enzyme Database (BRENDA) [33], the Transporter Classification Database (TCDB) [34] and PubChem [35].

On the other hand, for species-specific cases, the PlantCyc [36], Plant Reactome [37] and MetaCrop [38] are the ones who provide data for most plant species. In turn, SolCyc [39] only covers information for the Solonaceae family and The Arabidopsis Information Resource (TAIR) [40] only provides information for *Arabidopsis thaliana*. More species are available at Plant Metabolic Network (PMN) [36], which contains manually curated and predicted data from 125 plant species.

Lastly, software can be downloaded containing information and profiling data visualisation for plant species, like MAPMAN [41].

Table 1: Description of the most pertinent databases for plant metabolism studies.

Database	Description	Ref
MetaCyc	An all-inclusive database with the largest curated collection of metabolic pathways, containing data about chemical compounds, reactions, enzymes and metabolic pathways from all domains of life.	[30]
BioCyc	Pathway/Genome Databases (PGDB) collection that describes the genome and metabolic pathways of a single species. The database includes metabolites, enzyme activators, inhibitors, and cofactors, and it comprises transport systems and pathway fillers, as well as various tools for visualisation and comparative analysis.	[42]

Continued on next page

Table 1: Description of the most pertinent databases for plant metabolism studies.

Database	Description	Ref
NCBI	Open source repository of several databases that contain information regarding genomics and biomedical sciences, and tools that provide data retrieval systems and computational resources to analyse the structure and function of biologically important molecules.	[31]
KEGG	Database resource for knowledge of functions and applicability's of biological systems, dedicated especially for large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.	[29]
UniProt	Collection of protein sequences and detailed annotations for several organisms.It combines reviewed UniProtKB/Swiss-Prot entries, with unreviewed UniProtKB/TrEMBL entries.	[32]
BRENDA	Website that combines manually curated enzyme data with proteomic and genomic information. It provides an understanding overview on enzymes and combines versatile tools, for analysis, visualisation, and data retrieval, to access enzyme information. The data collection is based on the *Enzyme Commission (EC)* classification system.	[33]
TCDB	Open source curated database for transport protein research, which provides structural, functional, mechanistic, evolutionary and disease/medical information about transporters from several organisms. It is based on the *Transporter Classification (TC)* system.	[34]
PubChem	Largest database that serves the biomedical research communities in many areas like cheminformatics, chemical biology, medicinal chemistry and drug discovery, including information on molecular structure, physical properties and biological activities of compounds.	[35]
PlantCyc and PMN	PlantCyc is a repository for manually curated and reviewed information on metabolic pathways. PMN also provides access to curated and predicted information about enzymes, pathways, and more for several plant species.	[36]
Plant Reactome	Open-source, manually curated and comparative plant pathway database of the Gramene project. It uses *O. sativa* as a reference species for manual curation of metabolic and regulatory pathways and extends to another 82 plant species, also providing a suite of tools for analysis of large-scale omics datasets.	[37]
MetaCrop	Database that outlines information about metabolic pathways in crop plants and grants easy export of information for creation of accurate metabolic models.	[38]
SolCyc	Collection of PGDB for *Solanaceae* species generated using Pathway Tools software from SRI International. Databases generated from the respective genome annotations of tomato, potato, and pepper are available, some of them are curated.	[39]
TAIR	Database of genetic and molecular biology of *A. thaliana*. Data available includes complete genome sequence along with gene structure, gene product information, expression datasets, also including tools for visualisation and analysis of data.	[40]

Continued on next page

Table 1: Description of the most pertinent databases for plant metabolism studies.

Database	Description	Ref
MAPMAN	User-driven tool that displays large datasets onto diagrams of metabolic pathways or other processes providing visualisation of profiling datasets in the context of existing knowledge.	[41]

2.3.2 *Omics Data*

The word *omics* in cellular and molecular biology designates all constituents considered collectively [43]. Given the development of high-throughput technologies, it was possible to generate larger omics datasets. Therefore, omics technologies, referred to as high-throughput biochemical assays, measure thoroughly and simultaneously all molecules of the same type from a sample.

These data help investigate the central dogma of molecular biology through the detection and quantification of DNA, RNA, protein, and metabolites. However, such massive data generation requires bioinformatics to analyse and integrate them, presenting many hurdles, including differences in data cleaning, normalisation, biomolecule identification, data dimensionality reduction, biological contextualisation, statistical validation, data storage, handling, sharing and archiving. Nevertheless, researchers are surpassing this challenge, increasing our understanding of fundamental biological questions ultimately leading to the identifiedation of systems and synthetic biology [13, 44, 45].

The interpretation of molecular intricacy and variations at several levels, including genome, epigenome, transcriptome, proteome, and metabolome, has enabled the use of omics data in many applications, such as the developing a comprehensive understanding of human health and diseases, biological processes and different metabolic phenotypes. Furthermore, it helped transform medicine and biology, creating pathways for integrated system-level approaches [46]. Although many omics exist, genomics, transcriptomics, proteomics, and metabolomics are the main identifiedations of the systems biology field due to their connection with the central dogma.

Genomics is the study of the entire genome sequence and its information, including identification of single nucleotide polymorphisms, copy number, and loss of heterozygosity variants. The primary technique to obtain this data is whole genome sequencing (DNA-seq), used to confirm the oligonucleotide sequence of the DNA template and the resulting protein's chemical composition. The online repositories for genomic datasets are *Sequence Read Archive (SRA)* [47], which stores raw sequence data, *Gene Expression Omnibus (GEO)* [48], which deposits processed genomics data and *NCBI's The Database of Genotypes and Phenotypes (dbGaP)* [49], a public repository with controlled access for genotype and phenotype sequence data [50].

Transcriptomics provides information about the relative abundance of RNA transcripts, demonstrating the active compounds within the cell. The most used high-throughput techniques for transcriptional profiling are RNA-Seq, Microarrays and Serial Analysis of Gene Expression (SAGE). Public databases

for this type of datasets are *GEO*, *SRA* and *ArrayExpress* [51]. Other databases for genomics and transcriptomics datasets are depicted in table 2.

Proteomics identifies and quantifies all proteins expressed by a cell. The more frequently used strategies for proteomic profiling are two-dimensional gel-eletrophoresis and mass spectrometry (MS). MS allows the quantification of proteins and post-translational modifications, as well as identifying novel proteins. However, not all proteins can be detected by this method. Well-known public repositories for proteomics profiling are *PRIDE* [52], *ProteomeXChange* [53], *ProteomicsDB* [54], *PeptideAtlas* [55], *GPMDB* [56], *PAXDB* [57] and *JPOST repository* [58].

Metabolomics identifies all the metabolites present in the cell, resulting in the interaction of the transcriptome, proteome, and more, which gives information about not only the metabolite compounds produced but also the state of the cell. The most used strategies to acquire metabolic profiling are MS, NMR spectroscopy, and vibrational spectroscopy. MS output is used for the quantification of metabolites and new metabolite discovery. Public databases available for metabolic profiling are *Metabolights* [59], *MetabolomeExpress* [60], *GNPS* [61] and *Metabolome Workbench* [62].

Concerning plant databases for genomics, the most relevant is *Plant Omics Data Center (PODC)* [63], which includes core gene expression information regarding gene networks and knowledge-based functional annotations for plants and crops. *PlantExpress* [64] is a public web repository used for gene expression network analysis and microarray data on plants for transcriptomics. Relative to proteomics, the most well-known plant database is *Plant Proteomics Database (PPDB)* [65], storing curated data from mass spectrometry and proteome about protein functions, properties, and subcellular localisation. Lastly, plant metabolomics also has a public database, namely *Plant Metabolome Database (PMDB)* [66], which is a database of secondary metabolites of plants in the three-dimensional structures available in the biological data banks and databases.

As demonstrated, omics data are relevant for the study of systems biology. Nevertheless, the different types of omics are processed and analysed in different manners as they differ in terms of scales and structure. Therefore, there is no standard workflow pipeline able to analyze and process all these contrasting omics data, rather only specific workflows to process and analyse a particular omics data type [13]. In that way, many studies, like [12], [14] and [46], have emphasised the importance of integrating different omics data to have a more comprehensive view of a biological system.

Table 2: Description of the most pertinent databases for plant omics data.

Database	Description	Ref.
SRA	International public archival resource for next-generation sequence data.	[47]
GEO	International public repository that archives and freely distributes raw and processed genomics data, including metadata.	[48]
dbGaP	Public repository for individual-level phenotype, exposure, genotype, and sequence data, and the associations between them.	[49]
ArrayExpress	Public repository for functional genomics datasets and corresponding metadata.	[51]

Continued on next page

Table 2: Description of the most pertinent databases for plant omics data.

Database	Description	Ref.
GenBank	Public repository for collections of annotated nucleotide sequences and their protein translations.	[67]
RefSeq	Open-access database of annotated, curated and publicly available nucleotide sequences (DNA, RNA) and their protein products.	[68]
DDBJ	Biological database that collects DNA sequences at National Institute of Genetics.	[69]
ENA	Repository that provides free and unrestricted access to annotated DNA and RNA sequences and respective metadata.	[70]
Expression Atlas	Public archive that provides curated information on gene expression patterns from RNA-Seq and Microarray studies, and protein expression from Proteomics studies, and respective metadata.	[71]
EVA	Open-access database of all types of genetic variation data from all species.	[72]
MassIVE	Community resource to promote the global, free exchange of raw MS datasets.	[73]
NODE	Resource platform that supports flexible genomics, proteomics, metabolomics and fluorescence imaging data management and effective data release.	[74]
PRIDE	Public data repository of MS based proteomics data, maintained by the European Bioinformatics Institute.	[52]
ProteomeX Change	Consortium established to provide globally coordinated standard data submission and dissemination pipelines involving the main proteomics repositories.	[53]
Proteomics DB	Database for quantitative MS-based proteomics data, RNASeq expression datasets, drug-target interactions and protein turnover data.	[54]
PeptideAtlas	Publicly accessible compendium of peptides identified in MS proteomics experiments, providing tools for processing and analysing raw data.	[55]
GPMDB	Large Proteomics database that helps validate peptide MS/MS spectra as well as protein coverage patterns.	[56]
PAXDB	Public repository that contains genomic and proteomic information from various organisms and tissues. The datasets are scored and ranked by importing the protein network information.	[57]
JPOST repository	Database of integrated proteome datasets, where raw MS data is re-processed and automatically generates high-quality databases for data comparison and integration.	[58]
Metabolights	Repository for metabolic studies that provides research data and metadata as well as metabolite structures, their reference spectra, biological role, location, concentration, and experimental data.	[59]
Metabolome Express	Online server for processing, interpreting, and storing MS metabolomics data.	[60]

Continued on next page

Table 2: Description of the most pertinent databases for plant omics data.

Database	Description	Ref.
GNPS	Public database of raw, processed, and annotated fragmentation of MS data, assisting in the identification and discovery.	[61]
Metabolome Workbench	Repository for metabolomics data and metadata, including tools for analyses access to information on protocols , standard metabolites, tutorials, and more.	[62]
PODC	Public database providing gene expression networks, functional annotations, and additional comprehensive omics resources.	[63]
PlantExpress	Database for GEN analysis, providing functionalities specialised for OryzaExpress and ArthaExpress for *Oryza sativa* and *Arabidopsis thaliana*.	[64]
PPDB	Database for integration of MS-based proteomics data for the species *Z. mays* and *A. thaliana*.	[65]
PMDB	Public repository that collects three-dimensional protein models obtained by structure prediction methods.	[66]

2.4 MACHINE LEARNING

2.4.1 *Overview*

The development of high-throughput technologies led to the generation of vast large-scale omics datasets. Omics data are essential because they provide insights on genetic and molecular profiles, providing a more holistic perception of the organism's metabolism and the fundamental mechanisms that lead to different phenotypes. Nowadays, omics have been applied in numerous areas, such as developing and comprehensive understanding of human health and diseases, biological processes, and characterisation of complex biochemical systems. However, the different types of omics differ in terms of scale and structure; they are very complex and heterogeneous, so to process and interpret this data we need suitable tools. ML algorithms are the most used for this task since they can learn structures and associations, select relevant features from large datasets and make deductions by using example data or past experience without biological assumptions [75, 14].

2.4.2 *Concepts in Machine Learning*

ML is a subset of artificial intelligence and it focuses on the study of algorithms, where a computer uses experimental data to learn and make future predictions without being explicitly programmed. Hence, it can learn and readjust as from experience. For a better comprehension of ML algorithms, a few concepts should be reviewed [76].

An **algorithm** is a procedure that runs on the input data to create the best ML model, based on a given representation structure, that will be further explained in this section [77].The **model** results

from an algorithm that predicts the output values from input variables and generalises from current data.

Regarding the data structure, the processed data is usually represented in a matrix form, also named **dataset**, that is, a table schema, where the rows correspond to **instances** and the columns to **attributes**.

Instances, or objects, are a set of observations that we are interested in, from which the model will learn, or how a model will be used, for example, for predictions.

On the other hand, **attributes**, or features, describe an instance. The attribute type can be *categorical* or *continuous*. A *categorical* attribute is a finite number of discrete values that can be divided into two types: *nominal*, where there is no ordering between the values (e.g. names and colours), and *ordinal*, where there is an ordering (e.g. the attribute takes on the values low, medium, or high). Regarding *continuous* or quantitative attributes, they are described as a subset of real numbers, where there is a measurable difference between the values. They can be, for example, weight or temperature [76].

In some methods, **datasets** are divided into two sets to evaluate rationally the performance of algorithms: the *training set*, a subset to train and build the best model, and the *test set*, a subset to test the trained model and evaluate its effectiveness. Therefore, by using similar data for training and testing, the effects of data discrepancies can be reduced, and we can have a better understanding of the model's characteristics. A third set, named validation set, smaller than the training set, is often used to evaluate models' performance with different hyperparameter values and detect overfitting during the training stages [78].

2.4.3 *Types of Learning*

ML algorithms can be divided into three major categories: supervised, unsupervised and reinforcement learning.

Supervised learning uses previously labelled data, a training dataset, to classify and make predictions on new data. The labels allow the algorithm to correlate the features. The tasks of this type of learning can be *classification* or *regression* problems. A classification problem is a process where the predicted output is a discrete variable, e.g. a label. The quality of the model is usually measured using sensitivity and specificity as accuracy measure.

On the other hand, a regression problem is when the predicted output is a continuous variable, like a quantity or size. For this type of problem, the quality of the model is commonly measured by root mean squared error. In contrast, in unsupervised learning, no external indication is provided, and learning is carried out by finding regularities in the input data. Common tasks for this type of problem are clustering and PCA [79, 80].

Lastly, in reinforcement learning, the machine is not told which actions to take. Instead, it discovers which actions yield the most reward by trying them. In some cases, actions may affect not only the immediate reward but the following. In sum, it learns from the consequences of its past actions [81].

From the three categories, supervised learning is the most used and, in every application, there are

specific steps to develop supervised ML models, as shown in figure 4. The first step is *data collection*, where the quantity and quality of the data will determine our model's accuracy. The second step is *pre-processing and feature selection*, a critical step that will also help with our model's accuracy. The next step is *splitting data into train and test datasets*, following by *selection and optimisation of learning models*. The final step is the *evaluation of the model performance*.

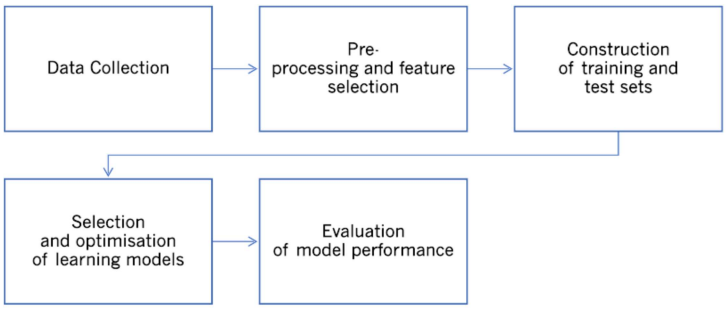


Figure 4: Necessary steps to develop a supervised ML model.

2.4.4 *Model Evaluation*

Evaluating the quality of a model for a given task involves calculating error measures on the test set. As mentioned before, these measures depend on the type of problem: classification or regression.

The confusion matrix that maps the values predicted by a model to real values is usually calculated for a classification problem. In a confusion matrix, the correct and incorrect predictions are compiled with count values arranged by each class. In other words, it shows where the classifier is making wrong predictions. There are two possible classes in a 2x2 matrix: Positive and Negative. The rows represent the real values, while the columns represent the predicted values, as illustrated in figure 5. There are four outcomes. If both the real and predicted values are positive, we have a **True Positive (TP)**; however, if the real value is positive but the predicted is negative then we have a False Negative (FN). On the other hand, if the real value is negative but the predicted value is positive, we have a False Positive (FP), but if both the predicted and real values are negative, we have a **True Negative (TN)**. For problems with more than two classes, a confusion matrix is calculated for each class, in which positive values represent one class and negative values represent the others.

Using the confusion matrix, we can calculate the accuracy of the model, also known as Percentage of Examples Correctly Classified (PECC), by dividing the sum of TN and TP values by the sum of all the other results, as shown in equation (a) of figure 6. Moreover, we can also calculate other error measures. Recall, for instance, is the division of TP values by the sum of all real positive values (TP+FN); as seen in equation (b), it measures the proportion of positive cases correctly identified. Specificity, represented in formula (c), on the contrary, is the proportion of negative cases that are correctly identified, calculated by the division of TN values by the sum of all real negative values (TN+FP). The Precision or Positive Predictive Value (PPV) (equation d) is calculated by dividing

Desired Values \ Predicted Values	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Figure 5: Example of a confusion matrix.

the TP values by the sum of all positive predicted values (TP+FP) and represents the proportion of positive predictive values. Likewise, Negative Predictive Value (NPV) (equation e) represents the negative predictive values and is therefore calculated by the dividing TN values by all the negative predicted values (TN+FN). Finally, the F-score measures the model's accuracy considering both sensitivity and precision and is calculated as shown in equation f).

a) $$PECC = \frac{(TN + TP)}{(TN + TP + FP + FN)}$$

b) $$Recall = \frac{TP}{(TP + FN)}$$

c) $$Specificity = \frac{TN}{(TN + FP)}$$

d) $$PPV(\,Precision) = \frac{(TP)}{(TP + FP)}$$

e) $$NPV = \frac{(TN)}{(TN + FN)}$$

f) $$F1 = 2 * \frac{(Precision\ * Recall)}{(Precision\ + Recall)}$$

Figure 6: Error measures used in classification problems. a) Accuracy. b) recall. c) Specificity. d) Precision or Positive Predictive Value. e) Negative Predictive Value. f) F-score.

For regression problems, measures are calculated based on the error made for each example. This is the difference between the predicted value ($\hat{\mathbf{y}}$) and the real value (y). There are several error metrics, depicted in figure 7, such as Sum of the Square Errors (SSE) that is calculated by summing the square differences of the predicted values ($\hat{\mathbf{y}}$) and the real values (y), as shown in equation g) and Root Mean Square Error (RMSE), that is measured by the square root of the SSE divided by the number of instances (N) (equation h). Moreover, there is also Mean of Absolute Deviation (MAD) that is obtained by the equation i).

As mentioned before, the original dataset should be divided into two datasets, *training set* and *test set*, to evaluate the model performance correctly. Since the model's training requires more examples, the training set usually is larger (80%) and the test set is smaller (20%).

A method generally used for model validation is Cross-Validation (CV) that allows the use of all data observations. In k-fold CV, the entire data is split into k-folds, the model is trained using the k-1 folds and, in each iteration, is tested with the remaining k*th* fold until all the k-folds serve as test set

g) $$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

h) $$\text{RMSE} = \sqrt{\frac{SSE}{N}}$$

i) $$\text{MAD} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}$$

Figure 7: Error measures used in regression problems. g) Sum of the Square Errors (SSE). h) Root Mean Square Error (RMSE). i) Mean of Absolute Deviation (MAD).

[82]. The error is measured by the mean of the errors in each iteration. With this method, we can estimate the variability and reliability of every model that uses that data [83].

However, it may be necessary to maintain the relative proportions of each class in the different sets, so a sample stratification process should guarantee the best possible effort for this result. A particular case of stratified cross-validation is the Leave-one-out method, where k is equal to the number of instances and in each iteration, the model is created using all instances except one, which will be used to test the model. Despite requiring more computational time, this method presents more viable results [84].

Additionally, the bootstrap method can be used to divide the data. This method is inspired by sampling with replacement [85]. Given a set of N original instances, N training instances are selected by resampling. The test examples, on the other hand, will be all those not selected for the training. This process is repeated often to have statistical significance [86], using appropriate statistical tests, such as t-test or ANOVA.

Evaluating the model's performance is a good practice as overfitting problems are often encountered. Overfitting is when the model learns examples too well and loses the ability to generalise, mainly due to the high variance of the model, caused by noise and peculiarities of the memorised training data. This can be addressed by removing redundant data features, increasing the number of data, since small datasets are more prone to overfitting than large datasets, using penalty methods or early stopping methods, and avoiding too complex models compared to the available data.

In practice, the problem is challenging to solve because having simpler models can lead to underfitting. This type of problem is the opposite of overfitting, as the model has low complexity in terms of features or the type of model; therefore, it is incapable of capturing the variability of the data [87].

2.4.5 *Model Selection*

The model selection (or selection of the best hyperparameters) is required for the optimisation process. Which aims to minimise the error over a set of validation instances (not used for the training process), and the error estimation techniques often used are resampling methods, such as cross-validation,

among others. The objective is to look for models that minimise overfitting and underfitting, that is, return models with the right balance between error and complexity.

There are several advantages in reducing the number of input attributes of a model since the model's complexity increases with the number of input attributes, which can cause overfitting. Moreover, data and models can be easily analysed and understood, and the elimination of redundant or contradictory attributes can improve the learning process by reducing noise. A simple technique is to extract the features from the dataset that are not relevant, getting only the ones that will grant our model the best results.

However, feature selection, defined as an optimisation problem that can become complex, given the wide search space can also accomplish this task. Algorithms for feature selection are divided into three groups: *filter*, *wrapper* and *embedded* methods. *Filter* algorithms select the features before the learning process, regardless of the model, and its evaluation is performed with statistical measures. *Wrapper* methods perform attribute selection in parallel with the model construction, they create different subsets of attributes, and the error is estimated by training a model and using the previously mentioned error measures. Forward and backward selection are examples of this type of wrapper heuristic algorithms. Forward selection starts with few attributes and adds attributes until it reaches a satisfactory behaviour. On the other hand, backward selection starts with a large number of attributes and removes one in each iteration. Lastly, *embedded* algorithms perform feature selection, model creation and evaluation at the same time [88].

Additionally, *ensemble* methods can improve learning performance, provide more accurate solutions, reduce overfitting and improve robustness over a single estimator. The result is calculated by a combination function, that given the output of all models, returns a single output value. In classification problems, the combination models return the output that gathers more "votes" and presents greater confidence associated with the class it proposes (winner-takes-all function). Contrarily, in regression problems, the returned output can be the mean of the individual results, a mean function, or a weighted mean, based on prior error estimation processes.

Nevertheless, diverse individual models are necessary to achieve accurate results. There are two approaches to build *ensemble* methods. The first approach manipulates training examples in different ways to create different models and introduce randomness. Three popular algorithms are:

- **Bagging** (Bootstrap aggregation), is based on bootstrap, in which each sample is created using a different bootstrap process.
- **Cross-Validation**
- **Boosting**, also based on bootstrap, takes into consideration the probability of each example being elected, and after creating each training set, the probabilities are updated by decreasing the probabilities of the correctly classified examples and increasing the rest.

The second approach consists of varying the initial parameters of the model to inject randomness into the algorithm. Random Forests are an example of this kind of approach as they are an ensemble of decision trees in which a random subset of attributes is selected to be tested, injecting this way

even more randomness. Additionally, this method also uses bagging to choose the training dataset, and as a way to avoid overfitting, the out-of-bag error is used in the instances not selected [89].

A simple ensemble learning technique is called majority voting and allows the combination of multiple classification models' predictions, turning it into a stronger meta-classifier. This technique balances out the weaknesses of the individual classifiers on a particular dataset as well as gives confident results thanks to the the associated individual weights [90]. The two main approaches in order to incorporate multiple predictions with voting are hard voting and soft voting. When hard voting is taken based on equal weights the predicted label becomes the *mode* of all the predictions. Otherwise, if the voting is based on different weights, the weights are applied to the prediction and the final label is computed accordingly. On the other hand, the soft voting approach, classifies the input data based on the probabilities of all the predictions from the different classifiers weights [90].

2.4.6 *Machine Learning Algorithms*

ML methods are often grouped by algorithm similarity, being the most recognised groups the succeeding ones [79, 91]:

- In each iteration, **regression algorithms** improve the model of the functional relationship between the numerical input features (independent variables) and the numerical output feature (dependent variable) using the error calculated in the model predictions. Three of the most popular algorithms are ordinary least square regression, linear regression, and logistic regression. These algorithms can be further modified to prevent overfitting problems, using regularisation methods like *Ridge Regression*, *Lasso regression* and *Elastic nets*, that maintain all attributes but reduce the magnitude of parameter values by penalisation.
- **Instance-Based Algorithms** do not explicitly create a model that generalises training data, but the data is stored. Stored data is only used when sorting a new example; hence it is called lazy learning. This type of algorithms, such as *K-Nearest Neighbour*, compare new examples with the training data to make predictions.
- **SVM** gained popularity in biology as tools for classification and regression. They use only the important examples (support vectors) and a nonlinear transformation of the inputs to a space of linear characteristics, via a kernel function. SVM ensures that an optimal class separation hyperplane is always achieved, which is also responsible for maximising the distance between both classes' data points.
- **Decision Tree Algorithms** are a favourite in ML due to their speed and accuracy. They build a model similar to a tree, where each node represents a given input attribute, each branch that exits this node corresponds to a possible value for this attribute, and the tree leaves designate a solution, that is, a value for the output attribute. The pathway from root to a leaf corresponds to classification rules. Examples of this type of algorithms are *ID3*, *C4.5*, *C5.0* and *J48* and *Classification And Regression Tree (CART)* algorithms.

- **Bayesian Algorithms** apply Bayes theorem to calculate probabilities of classification and regression, associated with an example belonging to each of the possible classes and use the class values' co-occurrence frequencies and the input attribute's values. The most popular Bayesian algorithms are *Naive Bayes* and *Bayesian Networks*.
- **ANN and Deep Learning algorithms** build neural networks, which simplify the human brain's models. ANN, is trained using *Backpropagation* and *Stochastic Gradient Descent* algorithms, and result in a parallel processor composed of simple processing units called neurons, that receive a set of inputs (data or connections) with a weight associated with each connection. The neurons then do a weighted sum of all these inputs and calculate, based on the activation function that filters the inputs, the signal that will be passed to the output.

  Thanks to technology evolution, more powerful computers were created, enabling more complex and much larger ANN, capable of higher accuracy, and dealing with very large datasets of analogue labelled data, such as image, text, audio, and video. Examples of deep learning algorithms are *Stacked Auto-Encoders*, *Convolutional and Recurrent Neural Networks*.
- **Ensemble algorithms** are composed of several weaker models trained independently to provide a better prediction achieved by the combining all the other predictions. Popular examples of this type of algorithms are *boosting*, *bagging* and *random forests*.
- **Clustering Algorithms** aim to separate groups with similar traits and assign them into clusters. Therefore, same groups' data points are more similar to other data points in the same group than those in different groups. Popular examples of clustering algorithms are *K-means*, *K-medians* and *Hierarchical Clustering*.
- **Dimensionality Reduction Algorithms** seek and exploit the data's inherent structure, selecting relevant features to summarise data and facilitate its interpretation. *PCA* is a popular example of this type of algorithms.

Thus, ML is of extreme importance in many biology areas due to its capability to store and manage information efficiently and extract useful information from large and heterogeneous datasets. The use of methods to transform heterogeneous data into biological knowledge and the underlying mechanisms is a well-known characteristic of ML, and it allows the creation of predictive models. There are several applications for ML tools, such as identification of regulatory elements (transcription factors, promoters), non-coding RNA genes, metabolites from MS metabolomics datasets, and expression patterns of genetic networks; prediction of the location, structure, and function of genes, RNA and proteins; and enable classification, modelling, and induction of genetic networks.

Finally, it also has an important role in evolution for the reconstructing phylogenetic trees through comparisons between different genomes [92].

2.4.7 *Machine Learning in Plants*

Concerning plant studies, ML has been applied in the interpretation of data acquired from high-throughput techniques in all levels of studies: genomics, transcriptomics, proteomics, and metabolomics [93]. Some of those applications are classification of functional proteins, especially ribosomal proteins of plants [94], functional protein classification in a virus family infecting plants [95], image processing to assess salt stress tolerance in wheat [96], classification of grapevine varieties [97] and detection of bacterial infection in melon plants [98], as demonstrated in table 3.

The use of ML algorithms, such as ANN, has been proved to surpass the traditional statistics methods used frequently for the study of omics in plants [93]. In genomics and transcriptomics, ML focuses mostly on the annotating a large number of sequence elements, identifying different gene expressions, and identifying resistance genes and pathogen effector genes. Sequence elements play a crucial role in gene expression, which is the case of micro RNA (miRNA), promoters, and transcription factors targets that ML tools have been successfully identified through genomics and transcriptomics data.

ML tools have been developed regarding miRNA focusing on plant system immunity and are efficient in discovering miRNA. Some examples include *PlantMiRNAPred* [99], which uses SVM algorithm to perform classification and predict plant pre-miRNA, *miRPara* [100], *MaturePred* [101], *MiRduplexSVM* [102] and *miTarget* [103], which also use the SVM algorithm to train the classification models and *mirLocator* [104] that implements the random forest algorithm.

Plant promoters, on the other hand, help develop disease-resistant or abiotic stress-tolerant plant varieties and are predicted using ML tools, such as *TSSP-TCM* [105], *PromMachine* [106] and *PromoBot* [107] that apply SVMs and also *TSSPlant* [108] that uses ANNs.

At last, transcription factor target genes are essential for the mechanisms that regulate the global gene expression, and ML approaches have also been applied to help identify these elements, for instance, [109], [110], [111] and [112] use SVM algorithms and [113] resorts to a Hidden Markov model strategy.

Concerning the global analysis of gene expression, ML methods were first used in [114], which developed a decision tree model. Several others have followed, like studying the stress response in *Loblolly pine* using Inductive Logic Programming (ILP) in express microarrays [115], identification of transcription networks regulated by glucose and ABA in *Arabidopsis* through relevance vector machine model [116] and to predict the class of plant varieties using gene expression profile to elucidate whether they can be distinguished by expression profiles of close-related plant genotypes, using SVM algorithms [117]. Some ML-based tools were also developed, for example, *MLDNA* to predict candidate stress-related genes using random forests [118] and *Beacon GNR* inference tool to predict gene regulatory networks in *Arabidopsis* seed development using SVM algorithm [119].

Furthermore, regarding plant immunity, two tools were developed for the prediction of plant disease resistance proteins against the pathogens of plants using SVM algorithms, *NBSPred* [120] and *Disease resistance plant protein predictor (DRPPP)* [121]. Additionally, ML tools were created to identify

effectors secreted by plant-pathogens, it is the case of *Localizer* [122], which uses the SVM algorithm, and *ApoplastP* [123], related to the prediction of localisation in the apoplast, using random forest.

ML techniques have recently, been applied in metabolomics, for instance, [124] combined network analysis and ML, using Random Forest, AdaBoost, SVM, and Naive Bayes algorithm to predict metabolic pathways from tomato metabolomics data. Finally, deep learning is, as well, being applied in plant molecular studies for phenotyping, disease identification, and genomics [125, 126]. Despite the wide range of ML applications in plant molecular biology, the lack of plant scientists with the required programming skills does not allow the evolution of these types of utilisation. Consequently, the development of easy-to-use programs for analysing plant omics data is still much necessary to thoroughly examine the possibilities of ML in plant omics data and processing information.

Table 3: Literature of several applications of ML in plants.

First Author (publication date)	Approach	Method	Type of Omics data	Ref
Xuan (2011)	PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNA	SVM	sequence-based	[99]
Wu (2011)	MiRPara: prediction of most probable microRNA coding regions in plants	SVM	sequence-based	[100]
Xuan (2011)	MaturePred: identification of microRNAs within novel plant pre-miRNA	SVM	sequence-based	[101]
Karathanasis (2015)	MiRduplexSVM: prediction and evaluation of miRNA-duplex	SVM	sequence-based	[102]
Kim (2006)	miTarget: microRNA target gene prediction	SVM	sequence-based	[103]
Cui (2015)	miRLocator: prediction of plant microRNAs	Random Forest	sequence-based	[104]
Shahmuradov (2005)	TSSP-TCM: Plant promoter prediction	SVM	sequence-based	[105]
Anwar (2008)	PromMachine: Pol II promoter prediction	SVM	sequence based	[106]
Azad (2011)	PromoBot: Prediction of plant promoters	SVM	sequence-based	[107]
Shahmuradov (2017)	TSSPlant: prediction of plant Pol II promoters	ANN	sequence-based	[108]
Holloway (2005)	Prediction of transcription factor binding of plants	SVM	gene-expression	[109]
Jiang (2007)	OSCAR: accurate recognition of cis-elements of plants	SVM	sequence-based	[110]
Dai (2007)	Predict target genes of transcription factors in *A.thaliana*	SVM	gene co-expression	[111]
Cui (2014)	Predict plant's transcription factor target genes	SVM	sequence-based	[112]
Dai (2017)	Sequence2vec: Modeling of transcription factor binding affinity landscape in plants	Markov Models	gene co-expression	[113]
Kell (2001)	Explanatory analysis of plant expression profiling data	Decision Tree	metabolomics	[114]
Heath (2002)	Studying stress response in loblolly pine	ILP	gene-expression	[115]
Li (2006)	Discover glucose-and ABA-regulated transcription networks in *A. thaliana*	Relevance Vector Machine	gene-expression	[116]
Ancillo (2007)	Class prediction of closely related plant varieties using gene expression profiling	SVM	gene-expression	[117]
Ma (2014)	MLDNA: A study of stress-responsive transcriptomes in *A. thaliana*	Random Forest	gene co-expression	[118]
Ni (2016)	BIT: Predict gene regulatory networks in seed development in *Arabidopsis*	SVM	gene co-expression	[119]
Kushwaha (2016)	NBSPred: Plant resistance protein NBSLRR prediction	SVM	sequence-based	[120]
Pal (2016)	DRPPP: Prediction of disease resistance proteins in plants	SVM	sequence-based	[121]
Sperschneider (2017)	LOCALIZER: subcellular localisation prediction of both plant and effector proteins in the plant cell	SVM	sequence-based	[122]
Sperschneider (2018)	ApoplastP: prediction of effectors and plant proteins in the apoplast	Random Forest	sequence-based	[123]
Toubiana (2019)	Prediction of metabolic pathways from tomato metabolomics data	Random Forest, AdaBoost, SVM and Naive Bates	gene co-expression	[124]
Ubbens (2018)	The use of plant models in deep learning: an application to leaf counting in rosette plants	Convolutional neural networks	images of synthetic plants	[125]
Ferentinos (2018)	Deep learning models for plant disease detection and diagnosis	Convolutional neural networks	leaves images of healthy and diseased plants	[126]

2.5 INTEGRATION OF MULTIOMICS DATA

The evolution of high-throughput technologies enabled the generation of extensive amounts of biological data, also known as *omics data*, concerning different cellular behaviour conditions. However, the analysis of a single type of omics data restricts the system's knowledge extraction. Therefore, integrating multiple omics data is fundamental to have a more holistic understanding of the biological system. This way, we will identify complex biological interactions and have a better perception of the genotype-phenotype relationship until then concealed. Nonetheless, this procedure has many challenges related to the different sizes, formats, and proportions of the data being integrated and the complexity, noisiness, contents and agreement between datasets [12].

ML has been the main focus of these integration methods, as it can integrate and manage information from large heterogeneous datasets and is capable of several further analysis, such as prediction, clustering, dimension reduction and association. ML, when trained with heterogeneous data (data from multiple sources) is designated as *"multi-view"* or *"multimodal"* learning [127].

Data integration is divided into two main approaches: *Multi-staged* and *Meta-dimensional* analysis. *Multi-staged*, a step-wise or hierarchical analysis based on a linear hypothesis that models the relationship between two given omics data, aiming to uncover cause-effect links, e.g cis relationships [128]. On the other hand, *Meta-dimensional* analysis integrates multiple different data types, combining them in a simultaneous step, and builds a multivariate model associated with a given outcome [128]. The latter method is the most promising for massive data integration and can be categorised into concatenation- (or early), transformation- (or intermediate), and model- (or late) based integration, as depicted in figure 8, and defined below [129, 12]:

- **Concatenation-based integration** joins the datasets into a single dataset before constructing a model. Since it requires a transformation of the datasets into a common representation, it may result in information loss, and it may be hard to identify the approach that best combines these data - due to the noise and different scales -, which may need a normalisation step. Moreover, data reduction may be required as this process may inflate high-dimensionality (the number of samples being smaller than the number of measurements for each sample). However, if the approach determined is successful, it will be relatively easy to use statistical methods.
- **Transformation-based integration** creates intermediate forms for each dataset individually, transforming each data type into an intermediate representation (kernel or graph) that will then be merged into a more elaborate model. The main disadvantage of this method is that, since this process transforms the data independently, the interaction effects is difficult to detect. Nonetheless, it is a good way to integrate several types of data with unifying features, preserve data-type-specific properties and it is a robust method for different measurement scales.
- **Model-based integration** generates multiple models using different data types as training sets and develops a final model from the multiple models created throughout the training stage. Therefore, this strategy is of great use when dealing with extremely heterogeneous data and

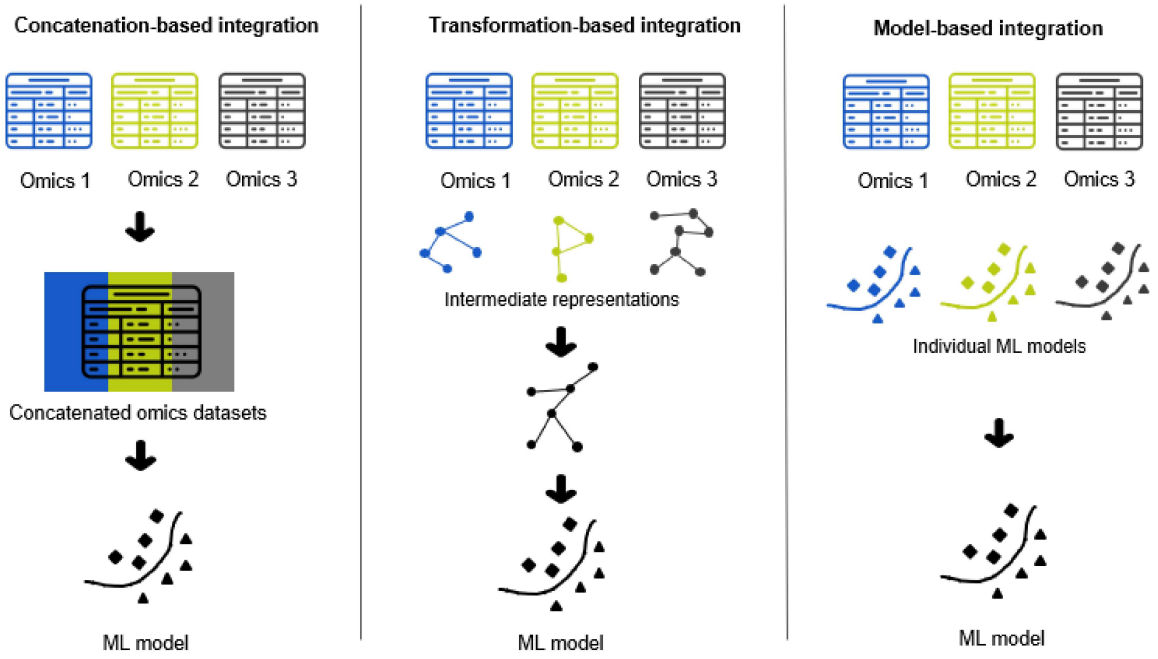


Figure 8: *Methods for multiomics data integration using ML techniques.* In the concatenation-based (early-stage) integration all the datasets are joined into a single dataset before constructing the model. Transformation-based (intermediate-stage) integration develops intermediate forms for all the datasets individually and transforms them into a intermediate representation to merge them into a more complex model. Model-based (late-stage) integration produces individual ML models for each of the datasets, that are then combined into a final ML model.

offers more flexibility. However, by transforming the data individually, it disregards the collective relationships, resulting in reduced in the final model performance.

The increase in multiomics studies led to more strategies and approaches for multiomics data analysis and integration that have been reviewed for instance in [12] and [127]. Although there is not a clear classification, they can be divided in the following groups: *Dimension reduction*, *Network-based*, *Bayesian* and *Multiple Kernel Methods*.

2.5.1 *Dimension Reduction Approaches*

Dimension reduction is a multivariate statistical approach that aims to transform the higher dimension datasets into a smaller dimension, guaranteeing that it provides similar information. The features, variables or columns in a dataset are known as dimensionality. A higher dimensionality hinders data

visualisation and predictions. Therefore, this approach helps obtaining a better predictive model while solving classification and regression problems [130].

PCA [131] is an unsupervised method that converts the original variables (typically correlated), using covariance analysis, into a set of non-correlated variables (linearly) that are called principal components (PC). Each PC is generated to explain the maximum variability of the part not yet explained, needing to be orthogonal to the previous PC. These PCs are ordered by the decreasing amount of variability that they explain. This way, it can increase interpretability while minimising information loss. However, this technique is sensitive to the data scaling, so previous normalisation is recommended.

Canonical Correlation Analysis (CCA) is a multidimensional exploratory statistical method in which the primary purpose is to explore sample correlations between two sets of quantitative variables observed in the same experimental units [132]. When this method decomposes each set, it finds the loading factors (linear combinations of variables), which maximise the correlation between sets while explaining the variability. Since traditional CCA cannot analyse omics datasets due to their high dimensionality, penalisation and regularisation terms were added cooperatively to develop more stable and sparse solutions. It is the case of L1-penalized sparse CCA (sCCA) [133] and elastic net CCA [134] created to make the results more easily interpretable in biological terms. Furthermore, structure-constrained CCA (ssCCA) [135] and CCA-sparse group [136] were created [137] to consider the group effects of features as structures embedded within the datasets.

Partial Least Squares (PLS) is a multivariate method used to identify latent structures of both predictors and responses by maximizing the covariance between them [138], and it is well-known for integrated omics studies. This technique can avoid the sensitivity of outliers and finds the fundamental relationships between two sets of data. Nonetheless, it has some difficulties in high-dimensionality data, so it is preferable to obtain sparse solutions for better interpretations. Therefore, extensions of this method were created, for instance, sparse PLS (sPLS) [139] that uses LASSO penalization for integrated omics, and extended Orthogonal signal correction PLS (OPLS) [140] that filters the "structured noise" (removes the systematic variation of predictors not correlated to the response) and enables a better interpretation of the data. Additionally, Sparse Multi-Block PLS (sMBPLS) [141] was developed to overcome the two datasets limitation and therefore integrate multiple omics data types, decomposing the datasets into sets of features that are strongly associated with the response. Furthermore, Sparse Network regularized PLS (SNPLS) [142] specializes in the identification of gene expression and drug-response relationship through assimilation of interaction network structures.

MixOmics is a R package with a whole range of multivariate methods capable of reducing the dimension of the data by using components, defined as a combination of all variables able to produce useful graphical outputs that enable better understanding of the relationships and correlation structure between the different datasets that are integrated [143].

A method used by this package is **DIABLO**, based on PLS and extends regularized and sparse generalized canonical correlation analysis (sGCCA) to a classification or supervised problem. sGCCA is a multivariate dimension reduction technique, and this method uses singular value decomposition and selects co-expressed variables from multiple omics datasets. It maximizes the covariance between linear

combinations of variables (latent component scores) and projects the data into smaller dimensional dimensional subspace spanned by the components. The data is maximally correlated using a design matrix that specifies the correlation between datasets. The identification of a multiomics panel is obtained through l_1 penalties in the model that shrink the variables coefficients defining the components to zero [144].

SMSPL proposed by Yang and Yan-Qiong, contrary to other dimension reduction techniques, uses a self-paced learning as training loss method, which instead of prioritising samples with higher training loss values chooses samples with smaller training loss values as easy samples, since they are more likely to be high confidence samples. This technique is a more suitable option for heavy noise scenarios and gives a desirable generalisation capacity.

An important aspect of SMSPL is that it takes into account the interaction between different modalities to recommend high-confidence samples for training the classifiers, using the advantage of common knowledge in sharing sample confidence between the several modalities. Additionally, when updating the training pool besides using the high-confidence samples justified by other modalities, it can also feed the pool with high confidence samples provided by very small loss values calculated on the current modality, which makes the proposed method utilize more-reliable high confidence knowledge from the prediction knowledge of the current classifier. Another aspect is that contrary to DIABLO and other methods that use majority voting it predicts samples by solving, making this method more accurate in terms of discriminating equivocal samples. Therefore, this method can simultaneously identify potentially significant multiomics signatures as well as predict subtypes during the integration process [145].

MFA is an unsupervised learning algorithm used for the analyse of different groups of variables related to the same observations. This sets of variables can be of different nature (qualitative or quantitative) between the groups but of same nature within the group. MFA follows the common structures present in all or most of these sets. Hence, a important step is to make these groups of variables comparable. Therefore, MFA starts by performing a PCA on each dataset and normalizes the data by dividing all its elements by the square root of the first eigenvalue obtained from the PCA. Then, the normalized datasets are concatenated and combined into an exclusive matrix to be submitted to a final PCA [146].

2.5.2 *Network-based Approaches*

Networks are widely used in biological representations, depicting the relations between entities, such as gene regulation and metabolism. A node represents a network entity, like genes or proteins, and the link between the nodes describes their relations. They are well-known for their capability to infer missing relations, and there are several biological networks with different features. In the context of multiomics data analyses, the multiple layer networks allow to uncover model relations between biological entities (genotype-phenotype) on a multi-scale. Each layer can represent a type of omics data, and inter-layer links can represent correlations between omics types. Multiple layers can be

displayed as networks, and they can be discriminated into multiplex networks and Heterogeneous Multi-Layered Networks (HMLN) [12, 147].

HMLN considers different kinds of nodes, each type corresponding to a different layer of biological information. In this type of network, intra-layer and inter-layer connections are treated in the same way, even if they have different weights. On the other hand, each layer of the *Multiplex networks* represents a different characterization of the same nodes. This difference is depicted in figure 9 [148, 147]. Network diffusion and other ML algorithms applied in multiomics studies have been reviewed in [147, 149].

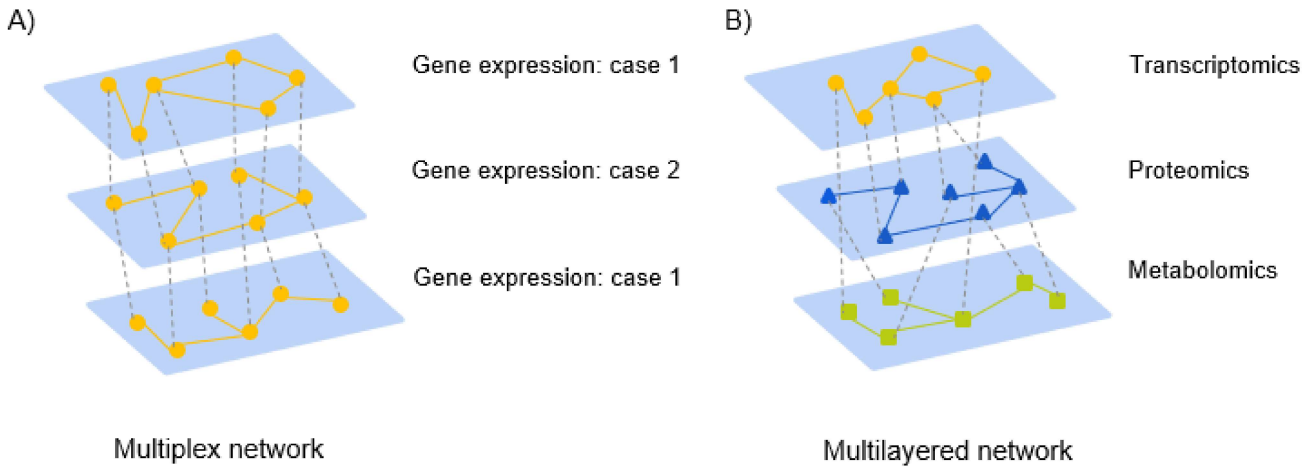


Figure 9: *Illustration of Multiplex and Heterogeneous Multilayered Networks.* **A)** In a multiplex network each layer represents a different characterisation of the same nodes, for example, genes. **B)** In a Heterogeneous Multilayered Network each layer represent a different group of nodes, for instance, genes, proteins and metabolomes.

Similarity Network Fusion (SNF) [150] is a method capable of computing and fusing patient similarity networks obtained from each omics separately in order to find disease subtypes and predict phenotypes [148]. SNF creates an individual network for each data type and fuses them into a single similarity network using a nonlinear fusion approach. It uses a local K -Nearest Neighbour (KNN) approach combined with ohter layers' global similarity matrices. With each iteration, the networks get more related to each other. This methods main advantage is that the weak connections, like noise, disappear with iterations, whereas the strong connections are propagated till convergence [46].This approach has been proposeed to analyse gene expression, mRNA expression, DNA methylation and miRNA of cohort cancer patients for tumor subtyping and survival prediction [127].

Graph-Composite Association Network (CAN) is a graph-based algorithm which primary task is to classify the unlabeled nodes using the network structure related to these nodes. Its objective function makes use of the *Laplacian* matrix of the network, that if is not sparse it's computing time can be very time-consuming and memory intensive. However, it is possible for it to be very sparse therefore, allowing the graph-based algorithm to be applied in large scaled networks. The CAN approach resolves

this limitation by using linear regression to obtain the weights of different data sources. Thanks to only having to solve one linear regression problem it usually performs better than other graph-based networks in terms of accuracy, F1 score and AUC. Additionally, it assigns weights for each data source by minimizing the least square error between the target network and composite weight matrix, then predicting via the combined weight matrix, making its training process less complicated than other graph-based networks, and therefore being an excellent choice to integrate different data sources when considering a graph-based approach [151].

NEighborhood based Multi-Omics clustering (NEMO) is an unsupervised multi-omics clustering method. Its unique characteristic lies in its simplicity and capability of supporting partial data. It is based and build on prior similarity-based methods, such as SNF and rMKL-LPP (see section 2.5.4). NEMO performs three tasks. First, it builds an inter-patient similarity matrix for each omics, then the matrices of the different omics are combined into one matrix and finally the network is clustered. Distinctively to other approaches, this method does not require iterative optimization and is faster [152].

2.5.3 *Bayesian Approach*

The Bayesian approach uses previous knowledge about the data probability distribution to incorporate this information in predictive or exploratory models through the computation of the updated posterior probability knowledge, using the Bayes' approach, depending on the dataset measurements [127, 148]. This approach's major advantage in multiomics integration is that it can make assumptions regarding various types of datasets with different distributions and the correlations between the datasets [137].

Patient-Specific Data Fusion (PSDF) [153] uses a Bayesian non-parametric model for clustering - Dirichlet Process model-, that integrates Copy Number Variations (CNVs) and gene expression data checking its concordance, to categorize them into sub-groups. The concordant samples will fuse, indicating patient-specific fusion models. This method reduces noise from data, selects only the important features, and estimates the number of disease subtypes from the datasets [137, 46].

iCluster [154] focuses on generating a single cluster assignment using a joint latent variable for integrative clustering based on simultaneous inference from multiomics data. This unsupervised method uses an expectation-maximization algorithm to infer. The associations between contrasting data types and the variance-covariance structures are incorporated flexibly into a framework while simultaneously reducing the datasets' dimensionality and then achieving data integration. This way, *iCluster's* goal is to identify a set of driving factors that define biologically and clinically relevant subtypes of disease, for instance, a type of cancer. However, it does not handle both categorical and numerical data; therefore an enhanced method was created *iClusterPlus*, which uses generalized linear regression to create of a joint model with categorical and numeric variables. It handles different types of omics data, like genomics, epigenomics, and transcriptomics and uses k driving factors to predict the most relevant variables and capture biological variation. Additionally, it uses the LASSO regression approach to indicate the subset of features that contribute to the biological variation between the subtypes [148, 138].

Bayesian Networks (BN) are part of probabilistic graphical models and model multi-view data with mixed distributions for classification and feature-interaction identification purposes. A BN is a directed acyclic graph, where nodes represent random variables and the edges represent Conditional Probabilities Distributions (CPDs). These CPDs can model conditional dependencies in continuous data, numerical data or a combination, granting this method the ability to capture noisy conditional dependence between multiomics data. This method has been applied system's biology including protein signalling pathways and gene function prediction. The search for an optimal BN structure is a NP-complete problem; thus, a heuristic method to solve this problem is required. An example of a BN method is the *Naive Bayes* classifier [127, 12].

PARADIGM *(Pathway Recognition Algorithm using Data Integration on Genomic Models)* [155] is a BN method that aims to infer the activity of patient-specific biological pathways from different types of omics data. It produces a matrix with integrated Pathway Activities which grant a set of essential profiles that contribute to delineate the subtypes regarding the survival outcomes [148].

Bayesian Consensus Clustering (BCC) is an unsupervised learning algorithm, that uses integrative statistics to allow a separate clustering of the objects for each data source. This separating clusters comply roughly to an overall consensus clustering, thus not being independent. BCC is a more flexible and robust approach than joint clustering of all data sources and it is more powerful than clustering each data individually. Its advantageous come from the capacity of modelling uncertainty and the ability to borrow information across sources. Although it is used for biomedical data, its applications are potentially unlimited [156].

2.5.4 *Multiple Kernel Learning Approach*

Kernel-based algorithms are a class of statistical ML methods often applied to pattern analysis, such as clustering, classification, regression, correlation and feature selection. They map, often on a kernel matrix, the original data to a high-dimensional space, denominated *feature space*, in which the pattern analysis is achieved. A popular kernel-based algorithm often used in biologic predictive problems is the SVM. This method's main advantages are that their optimizations are independent of the number of features, which is known as dimension-free, and it can integrate multiple data types and only requires defining the kernel function [127].

For multiomics integration studies, the *Multiple kernel learning (MKL) approach* is used, in which, instead of constructing a single kernel, multiple kernels for a single dataset are constructed, using different measures of similarity. These kernels are then linearly integrated into one kernel for further analysis, noticing that all kernel matrices representing different datasets should be constructed in the same feature space to be correctly combined. Generally, this approach provides a better performance. As this process follows a transformation-based approach, where data is transformed or projected into the same feature space, it leads to information loss, and the detection of interactions between different omics might be difficult, which represents a downside of this approach [127, 12].

Regularised MKL-Locality Preserving Projection (rMKL-LPP) [157] is used to perform multiple omics data integration dealing with gene expression, DNA methylation and microRNA expression

profiles and to perform subtype identification. The data is projected into a lower-dimensional and integrative subspace for clustering. This method automatically assigns higher weights to high information content and avoids overfitting using a regularization term, allowing different kernel types. The LPP is applied to conserve the sum of distances for each samples' K- Nearest Neighbours. The final cluster is performed applying a k-means on the distance summation. This approach is more flexible as it provides different options of dimension reduction methods and a variety of kernels per data type [137, 46].

Kernel-Relevance Vector Machine (RVM) is a ML method with a similar function as SVM, but applies a Bayesian inference to obtain probabilistic results. Although the performance can be similar to the SVM, the RVM is more competitive in several aspects, such as the results being sparser than SVM, and the computation time broadly decreased. Furthermore, it can grant probabilistic predictions for classification problems by returning the class probabilities. Additionally, it does not require any specification for loss parameters and the kernel is more flexible [151].

However since RVM is computationally intensive, **Ada-BoostRVM** could lower the computational cost. Ada-Boost RVM is an algorithm capable of combining different types of learners to improve the final performance, being the final classifier the weighted sum of several weak learners. Since its concept is to sample small training sets from the original training set and then train each model with the smaller training set, the computational cost is reduced. Although is difficult to distinguish these two methods the accuracy differences is very small between the two [151].

2.5.5 *Integration of multiomics data in Plants*

Plants are subjected to a wide range of environment al stresses that can be abiotic (e.g. temperature, salinity, drought) and biotic (e.g. attacks by pathogens), and because of these stresses, they activate complex interactions of multiple pathways in their metabolism as a coping mechanism [158]. Therefore, studies integrating different omics are fundamental to contribute to a more holistic view of the plant's metabolism and adaptation to the surroundings [159]. Recent plant multiomics studies are using parallel data integration, identifying the most useful features for each dataset individually and combining them into a final dataset that will train the model to predict the outcome more accurately. Additionally, simple correlation techniques, like Pearson and Spearman correlation coefficients, are being used to integrate and compare different omics datasets [159].

Regarding multiomics studies in *Vitis vinifera*, some studies used Pearson correlation. For instance, Ghan et al. [160] evaluated biochemical differences in biological systems by integrating transcriptomics, proteomics, and metabolomics data, where data dimension was reduced using PCA. A linear regression model was fitted to the transcript-protein pairs and computed using Pearson's correlation to investigate the linear relationship between the relative transcript abundance and therelative protein abundance. Zaini et al. [161] aimed to reduce the *Xylella fastidiosa* infection in *Vitis vinifera* and increase its health,using transcriptomics, proteomics and metabolomics responses to the disease compared with healthy plants. In this study, Pearson's correlation was measured between all pairs of samples of all datasets. As expected they inferred that the correlation among the experimental methods

(transcriptome, proteome, or metabolome) was a stronger determinant of higher correlation than the experimental groups (infected or non-infected). Furthermore, Savoi et al. [162] investigated the impact of water deficit on the secondary metabolism of white grapes using transcriptomics and metabolomics profiling data in a season of prolonged drought. They performed PCA on all metabolite profiles and transcriptome datasets, and used the Pearson correlation coefficient as similarity index between any two variables of the dataset (i.e metabolites).

More elaborate correlation methods and regression models, for example, CCA, PLS and O2PLS, and multi-omics networks are being applied in ither plant studies. O2PLS is similar to OPLS, but OPLS only returns one predictive component while O2PLS returns two [163].

Rajasundaram et al. [164] tried to establish the relationship between cotton fibre properties and non-cellulosic cell wall polysaccharides using data from glycomics and phenomics. CCA was used to obtain a global view of the association between the system levels, and sPLS was applied to predict cell wall polysaccharides linked with fibre characteristics. They proved the importance of these types of studies to obtain and develop high-quality fibre. Another example is the Bylesjo et al. [165] study, where they used transcriptomic and metabolomic data to investigate different light conditions on populations of wild *Populus tremula* x *Populus tremoloides*. They used O2PLS to identify the key transcripts and metabolites that defined most of the systematic variation across the two datasets. Subsequently, the two datasets' predictive features were used to infer the class and to show that the related structures captured the implicit class information through OPLS-discriminant analyses (DA). Zamboni et al. [166], focused on grapevine berry development and withering by integrating transcriptomics, proteomics, and metabolomics, also used O2PLS-DA to analyze each data set regarding different developmental stages and withering intervals to identify the key information (well-correlated transcripts, proteins, and metabolite variables) contained in the data, using PCA was able to recognize the three distinct classes.

A few studies extended O2PLS to handle multiple omics datasets. For example, Srivastava et al. [167] investigated system's responses to oxidative stress in *Populus*, integrating transcriptomics, proteomics, and metabolomics using a more sophisticated method of orthogonal projection to latent structures, named OnPLS. This more recent technique did not depend on the analysis order when more than two blocks were analyzed. Therefore the authors proved that pathways related to redox regulation, carbon metabolism, and protein degradation were significantly influenced in transgenic plants, providing information on the ROS metabolism and responses to oxidative stress, which indicated that some initial responses to oxidative stress shared common pathways.

Anesi et al. [168] aimed to understand the *terroir* effect in *Vitis vinifera* cultivar in Corvina in seven different sites over three years. They characterized the metabolome and transcriptome berries and used PCA, PLS-DA, O2PLS-DA, and orthogonal constrained PLS discriminant analyses as correlation analyses so that they could identify a *terroir* pattern in the berry metabolome composition for each site.

Other studies used different techniques, like Acharjee et al. [169], to find genetic and metabolic pathways related to phenotypic traits of interest of a population of potatoes using transcriptomics, metabolomics and proteomics data. They applied Random Forest regression to predict the four quality

traits, and constructed a partial correlation network for each trait, with genes, metabolites, proteins, and traits as nodes and correlation values as edges, and obtained relatively small sets of interrelated omics variables that could predict, with higher accuracy, a quality trait of interest. Another study by Wong and Matus [170] used Network analysis to investigate fruit composition regulation in grapevine using genomics, transcriptomics, proteomics, and metabolomics. In this particular study, they identified new non-identified transcription factors and various microRNAs that were responsible for regulating different steps of the phenylpropanoid pathway. The integrated network included genes, transcription factors, and RNA types as nodes and the interactions and correlation values as edges.

Jiang et al. [171] also used Network methods to integrate genomics, transcriptomics, proteomics, and phenomics data from *Z. mays* to build a multi-omics network to obtain information on its development. Weighted networks were created for each data type and combined into a final weighted network. The merged networks were analyzed, and the transcription factors with a key role in maize development were obtained from the orphan nodes.

A recent study by Jiang et al. [172] investigated the contribution of KLU (cytochrome P450 gene) to leaf longevity and drought tolerance using transcriptomic and metabolomic data from *Arabidopsis thaliana*. They used hierarchical clustering based on the heatmap of sample distances to identify the differential expression of the transcriptomics data, and PCA revealed that the harvesting dates grouped the samples. This study demonstrated that KLU overexpression activates cytokinin signalling by coordinately repressing cytokinin catabolism genes and the negative cytokinin response regulatory genes, and consequently, KLU-overexpression plants showed delayed leaf senescence.

Finally, a study by Nguyen et al. [173] used a method named *ManiNetCluster*, alongside the time-series gene expression dataset of *Chlamydomonas reinhardtii* microalgae, in different conditions (light and dark) to identify functional links between conditions. This study integrated transcriptomics data. The program takes two different datasets as input, and for each condition, a dataset is taken into account and builds a co-expression network for each of the datasets. The two networks are aligned into a single one, and this final network is clustered to allow the identification of genes linking functions from the different datasets. The studies described are organized in Table 4.

2.5.6 *Combination of experimental omics and predicted fluxomics*

Recently, a few studies emerged using ML to analyse experimental omics in combination with fluxomics data that were predicted by metabolic models (figure 10). The combination of omics with fluxomics data can improve our holistic view of plants' metabolism and discover genotype-phenotype associations not captured by high-throughput data alone.

An example is the work of Plaimas et al. [174], who presented an ML strategy to study and validate essential enzymes of a metabolic network in *E.coli*. First, for each reaction, features were defined describing local topology in the network, genomics and transcriptomics data and biomass rate predicted by Flux Balance Analysis (FBA). A table of all the reaction profiles was created and used for training the SVM classifier to differentiate between essential and non-essential reactions. The study showed that the approach improved FBA predictions results.

Table 4: Examples of multiomics studies in plants.

First Author (publication date)	Approach	Method	Type of Omics	Ref.
Ghan (2015)	Multiomics study to differentiate biochemical characteristics of grapevine cultivars.	PCA, Pearson correlation	transcriptomics, proteomics and metabolomics	[160]
Zaini (2018)	Discovering the *Vitis vinifera* metabolic response to the infection by *Xylella fastidiosa*.	Pearson correlation	transcriptomics, proteomics and metabolomics	[161]
Savoi (2016)	Analyse of the phenylpropanoid and terpenoid pathway of *Vitis vinifera* when subjected to prolonged drought.	PCA, Pearson correlation	transcriptomics and metabolomics	[162]
Rajasundaram (2014)	Study about the relationship between cotton fiber properties and non-cellulosic cell wall polysaccharides.	CCA, sPLS	glycomics and phenomics	[164]
Bylesjo (2007)	Analyse of short-day inducing effects on a population of *Populus tremula x Populus tremoloides*.	O2PLS, PCA	transcriptomics and metabolomics	[165]
Zamboni (2010)	Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks.	O2PLS, PCA	transcriptomics, proteomics and metabolomics	[166]
Srivastava (2013)	Study of oxidative stress responses in a population of *Populus tremula x Populus tremoloides* plants.	OnPLS	transcriptomics, proteomics and metabolomics	[167]
Anesi (2015)	Study of terroir conditions and their influence in the metabolome of the *Vitis vinifera*' s berry.	PCA, PLS, O2PLS	transcriptomics and metabolomics	[168]
Acharjee (2016)	Multiomics integration for the prediction of phenotypic traits of *S. tuberosum*.	Random Forests, Network analysis	transcriptomics, proteomics and metabolomics	[169]
Wong (2017)	Identification of new secondary metabolic pathway regulators for fruit composition in *Vitis vinifera*.	Network analysis	genomics, transcriptomics, proteomics and metabolomics	[170]
Jiang (2019)	Study of *Z. mays* development.	Network analysis	genomics, transcriptomics, proteomics and phenomics	[171]
Jiang (2020)	Contribution of KLU to leaf longevity and drought tolerance in *Arabidopsis thaliana*.	Hierachical Clustering, PCA	transcriptomics and metabolomics	[172]
Nguyen (2019)	Analyse the different light conditions effect on the metabolism of *Chlamydomonas reinhardtii* regulation of fruit composition in grapevine.	Network analysis	transcriptomics	[173]

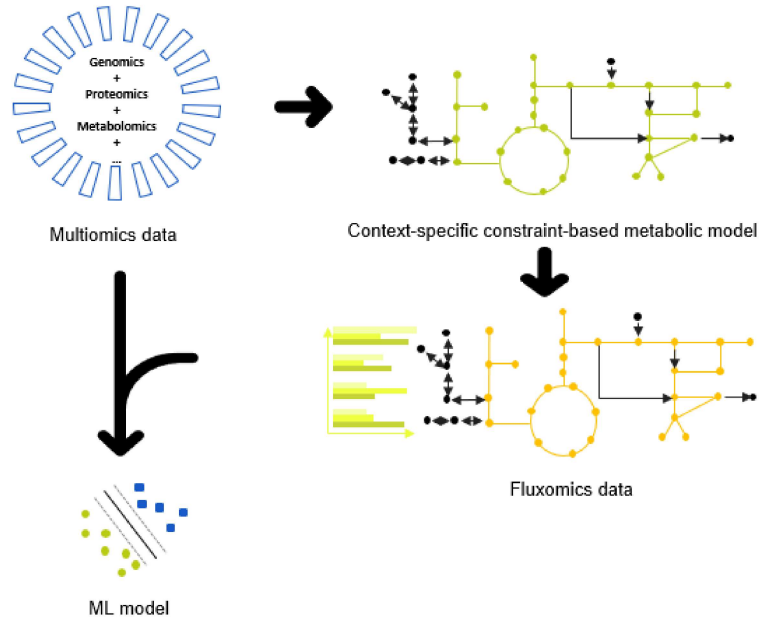


Figure 10: *Multiomics analysis*. Fluxomics data predicted by metabolic models can be analysed by ML in combination with omics data from high-throughput technologies.

An identical technique was applied by Szappanos et al. [175] to predict genetic interactions in *S. cerevisiae*. Using FBA, the authors computed *in silico* interaction degrees and single-mutant fitness. Gene-pair characteristics such as gene-expression, paralogy, network topology and protein annotations

were also defined. Then they used these features to train random forests and logistic regression and classify genetic interactions. The study demonstrated that incorporating FBA-derived fitness and genetic interaction scores into statistical methods boosted the predictions' precision, indicating that biochemical modelling provides unique information that is not captured by purely statistical data integration.

Another example of combining experimental omics data with predicted fluxomics is the work of Kim et al. [176] that created a database for *Escherichia coli* with well-annotated and normalized multiomics data, named ECOmics. They used transcriptomics, proteomics, metabolomics, phenomics (growth rate) and integrated these multiomics with fluxomics data obtained by condition-specific models. This compendium provides excellent data for predictive analysis, resulting in an incremental increase in the prediction performance.

3

MATERIALS AND METHODS

3.1 PLANT DATA COLLECTION

The datasets of our choice must have a considerable amount of samples to obtain reliable results. The larger the number of our samples, the more representative is the population, and the less influence the outlier observations have. Additionally, a large sample size provides better results among variables that are significantly different, setting a better picture for the analysis [177]. An extensive search in several databases was made with these requisites.

For the first case study (**Case Study I**), we opted for two datasets of *Vitis vinifera* transcriptomics and metabolomics data, both with two hundred and nineteen (219) samples with biological triplicates as replicates, for a total of seventy-three (73) samples. These datasets were taken from the Fasoli et al. study: the transcriptomics dataset was retrieved using the GEO accession number mentioned in the article (GSE98923) and the metabolomics dataset was available in the additional files, as well as the samples' metadata. In this research, two different cultivars were studied *Cabernet Sauvignon* and *Pinot Noir* and the main goal of this article was to get a better insight into the key molecular events controlling berry formation and ripening.

In the second case study (**Case Study II**), we selected two datasets of *Arabidopsis thaliana*, with only twenty-six (26) samples in total, exploring transcriptomics and fluxomics data. The transcriptomics dataset was identified in the Siriwach et al. article and retrieved from Gene Expression Omnibus under accession number GSE65046. The fluxomics dataset was identified in the supplementary material. The main goal of this article was to understand the mechanisms evolved in the adaptation of plants against drought in order to facilitate the development of drought-tolerant crops for agricultural use.

3.1.1 Case Study I

Sample Design

During three consecutive years (2012, 2013 and 2014), berries were collected, from the fruit set phase until harvest, at 10-day intervals during 2012 and weekly in 2013 and 2014. The authors defined *veraison* as clusters with at least 50% coloured berries and collected the samples at the same time of the day (8 am) in random blocks for each cultivar. In total, eight vine blocks for *Pinot Noir* and

six-vine blocks for *Cabernet Sauvignon* were placed and replicated along three rows for each cultivar in order to obtain biological triplicates and a total of 219 samples: 120 for *Cabernet Sauvignon* (39, 42, and 39 during vintages 2012, 2013, and 2014, respectively) and 99 for *Pinot Noir* (30, 33, and 36 during vintages 2012, 2013, and 2014, respectively). Each sample replicate comprised 26 clusters of berries from each vine block [178]. Since the 219 total samples were biological triplicates and therefore replicates, we calculated the mean for both the transcriptomics and metabolomics data, ending with 73 samples in total.

Metadata

Metadata contains information about the samples that are collected. In this case, the authors recorded with a specific identifier, for each harvested grape, the harvesting date (*Sample Date*), the *Variety* of each sample (*Cabernet Sauvignon* or *Pinot Noir*), the *Vintage* in which they were collected (2012, 2013 or 2014), and the *Time Point*, which corresponds to the reference time point of their reaping (the first ones to be harvest are noted as 0 and so on, until the last harvesting at the time point 14, for each *vintage*). Additionally, the metadata has information regarding the days after veraison (being the landmark for *veraison* the 0 days and ranging from -37 to 80 days), the content of malic acid (mg/L), reducing sugars (RS) (g /100 ml $\pm$ 5%) and finally the berry weight (g/berry).

Transcriptomics Dataset

The authors followed the steps mentioned in the article to obtain the transcriptomics data and were able to align the reads to the grapevine 12x genome PN40024 [180], in which an average of 86,7% of reads were mapped for each sample, allowing the reconstruction of the transcripts and the reference genome annotation V1 (http://genomes.cribi.unipd.it/DATA) [178]. This procedure resulted in the number of features present in table 5. The dataset downloaded from GEO with accession number GSE98923 contained the expression measurements of 29971 transcripts (rows) for 219 samples (columns). After pre-processing, this dataset was transposed so that the samples are represented by the rows.

Metabolomics Dataset

The authors collected sixty berries, from six confined clusters selected randomly from the vine blocks, and the samples were analysed by Gas chromatography–mass spectrometry (GC-MS) and Ultra-high performance liquid chromatography-quadrupole time-of-flight mass spectrometry (UHPLCQTOF-MS) to obtain the metabolome information. For the GC-MS analysis, the authors annotated 140 metabolites resorting to a reference library search, and 72 metabolites using UHPLCQTOF-MS analyse. The resulting number of metabolites is shown in table 5.

3.1.2 Case Study II

Sample Design

The work of Siriwach et al. provided one hundred and eight (108) distinct samples, that were part of four biological replicates (13 time points in control and drought conditions, plus a zero set). Additionally, the authors evaluated the changes in gene expression between time points, by comparing samples from adjacent sampling times, and assessed treatments, biological replicates and sampling times.

Metadata

The information regarding the samples from the *Arabidopsis thaliana* datasets is available in the metadata table. The authors recorded the day the samples were harvested, corresponding to the different time points (0 to 13) (*"time (day):ch1"*), the biological replicate (A,B,C,D) (*"biological replicate:ch1"*) and the treatment they were under (control or drought) (*"treatment:ch1"*).

Transcriptomics Dataset

The authors retrieved microarrays data from the Gene Expression Omnibus under the accession number GSE65046. The data was previously normalised providing a total of 26 samples and 32501 features, as depicted in table 5.

Fluxomics Dataset

The fluxomics dataset was retrieved from the supplementary files of the work by Siriwach et al.. The authors reconstructed context-specific Genome-scale Metabolic (GSM) models using Gene Inactivity Moderated by Metabolism and Expression (GIMME) to integrate the transcriptomics dataset within the GSM models. Then, the authors use the models to determine fluxes through FBA. As shown in table 5, the dataset has 26 samples and 1602 reactions.

Table 5: Dimension of the two *Vitis vinifera* datasets (total of samples and features) before pre-processing. The samples are the same in both datasets.

Omics Dataset	**Samples**	**Features**
Case Study I		
Transcriptomics	73	29971
Metabolomics	73	212
Case Study II		
Transcriptomics	26	32501

Fluxomics	26	1602

3.2 PRE-PROCESSING

Pre-processing data is one of the most fundamental steps of ML, to improve the data quality. These steps transform the data to be used by ML models. Four well-known data pre-processing tasks are missing value imputation, data normalization, data discretization and filtering/feature selection [181].

Pre-processing starts with data exploration, to easily view possible problems, which are usually associated with the incorrect reading of files, and to visually uncover data trends and points of interest. This can be accomplished by verifying the sample and feature names, the dimensions of the dataset as well as the type and class of the data. The R software [182] allows users to use the *names()* function to return or set the names of a data object, the *dim()* function to check the dimensions and the *class()* and the *typeof()* function to return the class and type of the data object.

3.2.1 *Missing Values*

To check whether the transcriptomics or proteomics dataset had any missing values, the *is.na()* function of R was used. In the present work, the rows/columns containing the missing values were deleted. However, depending on the analysis to be performed, other approaches can be used, for instance: substituting the values for another value, like the mean of the row/column, using the KNN approach to use information from neighbour lines to opt for a better value, or a valid approach in transcriptomics analysis is to insert the value zero for missing values. Regarding the fluxomics dataset, several of the reactions (features) contained only zeros, therefore we removed the rows with only zero values.

3.2.2 *Data Standardization*

Data Standardization is required to compare variables. Both transcriptomics and metabolomics data, from **Case Study I**, were previously normalized by the authors, as well as the transcriptomics data from the **Case Study II**. For the **Case Study I**, the normalized expression of each transcript was calculated as Reads Per Kilobase Million (RPKM) for each sample. Regarding metabolomics, the authors used median scaling to normalize the metabolite measurements across vintages: for each annotated metabolite, the abundance level in a given year was divided by the ratio of the median of that specific year to the median for that metabolite in all samples of all years [178].

On the other hand, for the **Case Study II** the transcriptomics dataset was normalized within the array by Lowess normalization and the variation due to arrays and dyes was also removed by a random-effects model, and averaged, in log space, using a modified version of R/MAANOVA (MicroArrayANalysis Of VAriance).

A possible alternative is scaling the data using the R function *scale()*, which normalizes the values

by subtracting them by the mean and dividing them by the standard deviation. This function was used in the metabolomics dataset from the **Case Study I** and fluxomics from **Case Study II** to further scale the data and better plot visualization.

3.2.3 *Data Discretization*

Data discretization is a useful variable transformation technique that involves dividing the range of possible values into sub-ranges and considering each one of them as a category. This procedure was used in the **Case Study I** sample's metadata to transform the numeric variable *Time Point*, ranging from values 0 to 13, to a variable named **Berry_development** with two factors, referring to the pre-Veraison and post-Veraison berry development stages (wherein, the post-Veraison stage includes the Veraison stage) and later three factors, pre-Veraison, Veraison and post-Veraison separated. The *Time Point* variable (x) was categorized in the following categories:

- **Berry_development** - two factors: *PreV*, if $x \in [0, 3]$ and *PosV* , if $x \in [4, Inf[$.
- **Berry_development** - three factors: *PreV*, if $x \in [0, 4[$, *Veraison*, if $x \in [4, 5[$, and *PosV* , if $x \in [5, Inf[$.

3.3 FEATURE SELECTION

As omics datasets are very large, a decisive step to improve the interpretability and performance of the model, as well as speed up the learning process, is feature selection. This technique extracts a subset of relevant features, consequently decreasing the size of the original dataset. The three general classes of methods for feature selection often referred in the literature are *Filter methods*, *Wrapper methods* and *Embedded methods* as explained in chapter 2 [181].

In this work, we used mainly *filter methods*, which apply an external statistical measure to compute a rank for each feature depending on the assigned score [183]. The features are then selected to be kept or removed from the dataset. This method was used in **Case Study I** and **Case Study II** transcriptomics dataset, where three filters were executed. First, we filtered the transcripts where the mean for that transcript was less than one, assuring that at least one transcript per cell existed. Then, we performed a median filter where it calculated a median expression level for the dataset as a cutoff value and checked whether each gene was "expressed" in each sample, selecting the genes where the expression occurred in more than 2 samples. Finally, we executed a flat patterns filter, which assumed that genes with very consistent values do not contribute to relevant information and therefore filtered genes whose maximum ratio value over the minimum value of expression was greater than 2. The number of features was, therefore, decreased as shown in table 6, except for the metabolomics dataset that was not filtered, and the fluxomics dataset that underwent feature selection but the filters did not remove any features.

Table 6: Dimension of the two *Vitis vinifera* datasets (total of samples and features) before filtering and after filtering. The samples are the same in both datasets.

Omics Dataset	Samples	Features[a]	Features[b]
Case Study I			
Transcriptomics	73	29971	32501
Metabolomics	73	212	212
Case Study II			
Transcriptomics	26	32501	725
Fluxomics	26	1602 (407) [c]	407

a Original feature size
b Feature size after filtering
c Feature size after missing values operation

3.4 MODELS

For the individual omics analyses, three models in R were created for each dataset, using different algorithms, namely:

1. Support Vector Machines (SVM) using method *"svmLinear"* from package *caret*;
2. Random Forests (RF) using method *"rf"* from package *caret*;
3. Artificial Neural Networks (ANN) using method *"nnet"* from package *caret*.

These models were specifically selected for the individual omics analysis due to their higher performance in comparison to other family classifiers. As explicit in the article by Fernández-Delgado et al. the best family of classifiers is RF, followed by SVM and ANN.

For multiomics integration, several models were used, differing in the programming language, ML algorithm function and the type of learning, unsupervised or supervised, as depicted in table 7. The two programming languages used in this project were R [185] and Python [186], as R is a recognized and successful language for ML used to discover models for multiomics integration, and Python is a flexible and simple language with a large number of libraries and frameworks, commonly used to streamline large complex data sets.

For the Concatenation-Based Integration we used the two programming languages, R and Python. The algorithms used on R were:

- DIABLO from the package *mixOmics* [144], selected due to being one of the only first multivariate integrative classification methods that builds a predictive model to predict new samples, and also being able to identify correlated (or co-expressed) variables measured on

heterogeneous data sets, which also explain the categorical outcome of interest (supervised analysis) [144];

- SMSPL by Yang and Yan-Qiong, that besides predicting subtypes can also identify potentially significant multiomics signatures, and deal with the high noise of multiomics data, which is the main cause of overfitting and poor performance;
- Stacked generalisation using the package *h2o* [187] selected to evaluate how well the performance of several base models of the concatenated multiomics data produced into one optimal predictive model would be;
- Lasso Regression from the package *glmnet* [188], as considered in the article [189] a good concatenation-based integration approach.

We also used Python to execute SVM with the *SVC()* function, ANN with the function *MLPClassifier()* and RF with the function *RandomForestClassifier()* from the *scikit-learn* [190] library. These algorithms were advised in the [189] article and as mentioned in Fernández-Delgado et al. are the best classifiers for each family and for this reason, were selected for the individual omics analysis. For unsupervised learning, we used the MFA method from the R package *FactoMineR*.

Regarding Transformation-Based Integration, only models designed in R were identified, more specifically, SNFtool from the package *SNFtool* [150] mentioned in various articles [148, 137] and Graph-CAN and Kernel-Integrated RVM, both from package *MDIntegration* [151], based in kernels and graphs that were also advised in [189]. The unsupervised learning model used for transformation-based integration was NEMO, using the R package installed with *devtools* from the GitHub repository of *Shamir-Lab* https://GitHub.com/Shamir-Lab/NEMO/tree/master/NEMO), mentioned in articles like [191].

Lastly, Model-based integration focused on the use of ensemble classifiers with hard and soft voting implemented, created through various functions identified on the library *scikit-learn* [190], as suggested by [189]. The unsupervised learning model used in this integration was BCC using as source file the R BCC function taken from the GitHub repository https://GitHub.com/ttriche/bayesCC, an interesting model considered in [191].

The majority of the multiomics models selected for this work were retrieved from different publications [191, 127, 189, 12, 148, 137, 46, 191]. However, most of these were all implemented and evaluated using human tissue samples, for instance, the well-known TCGA (human cancer) multiomics dataset.

3.5 MODEL EVALUATION

Different methods were selected depending on the algorithm to evaluate our models' performance. For the individual omics analysis models, we selected the repeated CV (*method="repeatedcv"*) in the *"trainControl()"* function of the package *caret*, that served as input for the *"train()"* function. Hence, we assure that all data is used for training and testing by dividing our data into training and test sets, which improves the performance and results in a finer model evaluation. The CV was performed using 10 folds and 3 repetitions.

Table 7: Description of the several models used for the different integration approaches (Concatenation, Transformation, and Model-Based Integration), as well as the package or function executed in the corresponding programming language.

Model	Package/ Function	Type of Integration	Programming Language	Ref
DIABLO	mixOmics	Concatenation-Based	R	[144]
SMSPL	GitHub: ZiYi Yang	Concatenation-Based	R	[145]
Stacked Generalisation	h2o	Concatenation-Based	R	[187]
Lasso Regression	glmnet	Concatenation-Based	R	[188]
SVM	SVC (scikit-learn)	Concatenation-Based	Python	[192]
ANN	MLPClassifier (scikit-learn)	Concatenation-Based	Python	[190]
RF	RandomForestClassifier (scikit-learn)	Concatenation-Based	Python	[190]
MFA[1]	FactoMineR	Concatenation-Based	R	[193]
SNFtool	SNFtool	Transformation-Based	R	[150]
Graph-CAN	MDIntegration	Transformation-Based	R	[151]
Kernel-Integrated RVM and Boosted-RVM model	MDIntegration	Transformation-Based	R	[151]
NEMO[1]	GitHub:Shamir-Lab/NEMO	Transformation-Based	R	[152]
Ensemble Classifier with different ML algorithms (Hard and Soft Voting)	Various functions from scikit-learn	Model-Based	Python	[190]
BCC[1]	GitHub:ttriche/bayesCC	Model-Based	R	[156]

[1] Unsupervised Learning

In addition, the metrics selected to evaluate our models were PECC, recall and precision, which could be automatically extracted from the confusion matrices created for each model, using the function *confusionMatrix()* in the *caret* package, and also the ROC curve and AUC value, obtained using the package *MLeval*. Furthermore, for all the models in individual omics analyses, we were able to extract the top 20 most relevant features using the *caret varImp()* function.

Regarding multiomics integration algorithms, as shown in table 8, most algorithms used 5-fold CV, except for DIABLO who used repeated 10-fold CV with 10 repeats, SMSPL that opts for a 10-fold CV and Stack Generalisation that uses a 15-fold CV. For this implementation, DIABLO used the parameter ($validation = 'Mfold', M = 10, nrepeat = 10$) in its algorithm. Cross-validation was implemented in SMSPL, as well as in Lasso Regression, using the default value 5 in the *cv.glmnet()* function. Regarding Stack Generalisation, the CV was implemented in the *h2o.glm()* function changing the following parameters *nfolds* and *fold_assignment* to 15 and "Modulo", respectively. On the other hand, scikit-learn models SVM, ANN and RF used *cross_val_score* and *cross_val_predict* functions with argument *cv=5* for their implementation. SNFtool implements an internal (M-fold) CV analogous to that of the DIABLO algorithm. Graph-CAN and Kernel-RVM also implement an internal CV method and both use 5-fold CV. The Ensemble Classifier does not use any cross-validation method.

In addition, the evaluation metrics used to evaluate the multi-view models' performance were the same described for single-view models, using the function *confusionMatrix()* in the *caret* package, except for the DIABLO method, which used a specific function *evaluate.DIABLO.performance()*, and Lasso regression, that used evaluation metrics specific to regression problems: RMSE and R-square, obtained using *assess.glmnet()* that produces a list of vectors of measures.

In respect to the Python models SVM, ANN and RF, three functions in the *scikit-learn* library,

Table 8: Description of the several metrics and the validation method used to evaluate the performance of the different models used for the different integration approaches (Concatenation, Transformation, and Model-Based Integration). PECC (accuracy), Precision and Recall for the classification algorithms, and RMSE and R square for the regression algorithms. Additionally, ROC curves and AUC values were also determined.

Model	Validation Method	Metrics						
		Classification Metrics					Regression metrics	
		PECC	Precision	Recall	ROC curve	AUC Value	RMSE	R Square
DIABLO	Repeated 10-fold CV with 10 repeats	X	X	X	Yes	Yes		
SMSPL	10-fold CV	X	X	X	Yes	Yes		
Stack Generalisation (Ensemble)	15-fold CV	X	X	X	No	Yes		
Lasso Regression	5-fold CV	X	X	X	Yes	Yes	X	X
SVM	5-fold CV	X	X	X	Yes	Yes		
ANN	5-fold CV	X	X	X	Yes	Yes		
RF	5-fold CV	X	X	X	Yes	Yes		
SNFtool	None	X	X	X	Yes	Yes		
Graph - Composite Association Network (CAN)	5- fold CV	X	X	X	Yes	Yes		
Kernel - Integrated RVM and Boosted-RVM model	5- fold CV	X	X	X	Yes	Yes		
Ensemble Classifier with different ML algorithms (Hard and Soft Voting)	None	X	X	X	Yes	Yes		

were used to extract the metrics, namely, *metrics.accuracy_score()*, *metrics.recall_score()* and *metrics.precision_score()* as well as the *classification_report()*. The *classification_report()* was also used in the Ensemble Classifier model to extract the metrics.

Finally, the Receiver Operating Characteristics (ROC) curves and Area Under Curve (AUC) values were also obtained using different functions depending on the algorithm. DIABLO used the *auroc* function from the *mixOmics* package for the ROC curve and the function *evaluation.DIABLO.performance()* for the AUC value. SMSPL, Stack Generalisation, Lasso Regression, Graph- CAN and Kernel-RVM models used the *roc()* function from the *pROC* library that automatically accessed the AUC value. The Python models used the *roc_curve()* function and the *auc()* function both from the *scikit-learn* package to plot the ROC curve and obtain the AUC value.

3.6 MODEL OPTIMIZATION

ML models are parameterized to best adapt to different problems. A favourable set of parameters can improve the models' performance, opposed to a unfavorable set of parameters that decreases the models' performance.

For R algorithms, ANN, RF and SVM, in individual omics analysis, the function *expand.grid()* from the *caret* package was used to perform grid search with the manual grid to look for the best model. Additionally, automatic grid search was implemented using the parameter *tuneLength* in the *train()* function.

Regarding the multiomics integration models, ANN, RF and SVM in Python, the best models were obtained using *sklearn* functions *GridSearchCV()* and *RandomizedSearchCV()*. Table 9 displays all the values used for both grid search and automatic/random search.

On the other hand, most of multiomics integration algorithms tune models in different ways. DIABLO fits a model without variable selection to assess the performance and selects the best number of components for the final DIABLO model, that indicate the number of sufficient components to discriminate all phenotype groups. Lasso Regression uses CV to find the best lambda, as well as MDIntegration. SMSPL and SNFtool tune the model as part of the algorithm in an automatic way.

Table 9: Description of the several hyperparameters and different values used to perform Manual Grid Search and Automatic/Randomized Grid Search to find the best performance, in both R and Python SVM, RF and ANN models.

	Model	Parameters	Manual Grid Search	Automatic Grid/ Randomized Search
Individual Analysis[1]	SVM	C (cost)	seq(0, 2, length= 20)	tuneLength=15
	RF	mtry	c(2,3,6,7,10) floor(sqrt(ncol(x)))	tuneLength=20
		ntree	c(500, 1000, 1500, 2000, 3000) c(20,40,60,80,100)	
	ANN	size	seq(1, 10, 1) c(3, 5, 10, 20)	
		decay	seq(1, 100, 10) seq(0.1,0.5,0.1) c(0.5, 0.1, 1e-2, 1e-3,1e-4, 1e-5, 1e-6, 1e-7)	tuneLength=20
Multiomics Integration[2]	SVM	C	[1, 3, 10, 100]	[1,3,10,100]
		gamma	[0.01, 0.001]	[0.01, 0.001]
		kernel	('linear'; 'rbf')	('linear', 'rbf')
	RF	max_depth	[2,3,None]	[2,3,None]
		max_features	[2,4,6]	[2,4,6]
		min_samples_split	[2,4,6]	[2,4,6]
		min_samples_leaf	[2,4,6]	[2,4,6]
		bootstrap	[True,False]	[True, False]
		criterion	['gini',entropy']	['gini',entropy']
	ANN	hidden_layer_sizes	(8,8,8)	((15,),(25,),(50,),(75,),(100,))
		activation	'relu'	('identity, 'logistic', 'tanh', 'relu')
		solver	'adam'	('lbfgs', 'sgd', 'adam')
		max_iter	1000	
		alpha		(0.0001,0.001,0.01)

[1] R programming language.

[2] Python programming language.

3.7 COMPUTATIONAL FRAMEWORK

The developed tool will be available in an open-source computation framework, that can be identified in https://insilicoplants.pt/, and is currently available on the GitHub page https://GitHub.com/InesFaria-UM/Master_Thesis.git.

4

DEVELOPMENT

The pipeline for this work was created using Python 3.8.11 in the integrated development environment (IDE) Spyder, and connects both Python, a multifaceted programming language, with R, a well-known language preferred by many data analysts, statistics and graphics, using the *rpy2* package, which creates an interface between both languages, allowing access to the libraries of one language while working on the desired language. This pipeline follows 5 steps, as depicted in figure 11.

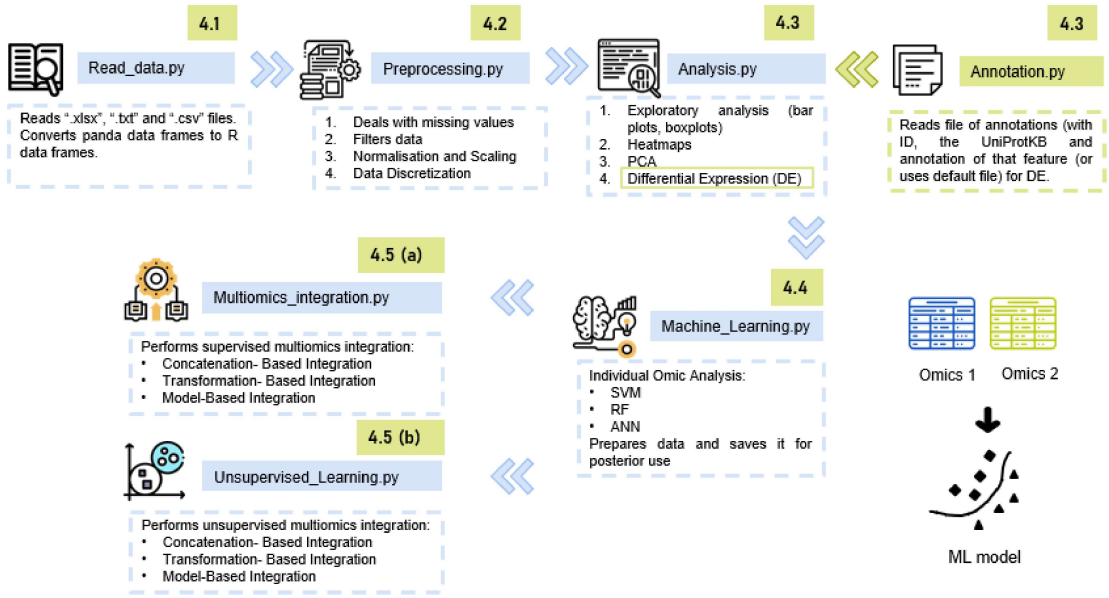


Figure 11: *Pipeline Schema*. Order and steps of the different development stages.

4.1 PLANT DATA UPLOADING

The omics files and metadata selected by the user are uploaded to the open-source framework. The data files are then converted into easier formats for analysis. At this stage, the computer runs the script **Read_data.py**.

– **Read_data.py**

Input:

1. **omics_files:** A dictionary, where the keys describe the type of omics (for example, "Transcriptomics"), and the values corresponds to the file path that leads to the dataset.
2. **skip_row:** A dictionary, where the keys describe the type of omics and the values the number of lines the function should skip to start reading the file.
3. **feature_name_col_index:** Another dictionary, that consists of the name of the omics as key and the index of the column corresponding to the features names as value.
4. **header:** Takes as default the value 0 indicating that all the datasets have an header.

The selected data files should have the features in the rows and the samples in the columns, except for the metadata. Only datasets with ".txt", ".xlsx" and ".csv" format can be introduced, for now. This script takes the file paths and the type of omics, converts them into a dictionary (omics_files) and, using the rest of the input parameters, it runs the function *"read_data()"* that reads the data and creates a pandas dataframe, in which the index corresponds to the feature names (rows), based on the *feature_name_col_index* parameter, and the *skip_row* parameter allows an efficient reading of the file, so that it can then be converted into an R dataframe. The R dataframe grants an easier base format for later analysis.

Output:

1. **data** - A dictionary that contains as keys the type of omics and as value the R dataframes.

4.2 PRE-PROCESSING

After data assimilation, an important and necessary step to prepare the data and assess its quality for posterior analyse is data pre-processing to make data more tractable for ML models. However, depending on the type of omics, the pre-processing method can differ. Nevertheless, four main steps in pre-processing are common: dealing with missing values, feature selection, data standardization and data discretization. The machine runs the script **Preprocessing.py**.

– **Preprocessing.py**

Input:

1. **data:** The dictionary obtained from the **Read_data.py** script.

Firstly, the *"na_delete()"* function searches for *NAs* in the dataframe, if true it deletes the rows that contain those *NAs*.

Secondly, the *"filtering()"* function performs feature selection based on filter methods, as mentioned in section Feature Selection in chapter 3.

Thirdly, using the filtered dataframes, and receiving as input the *scaled* parameter (a dictionary where the keys correspond to the omics type and the values are Booleans indicating whether the data is normalised or not) the function *"normalization()"* performs the scaling of the data that is not normalised.

Finally, the *"data_discretization()"* function allows the user to choose a numeric vector or string name (parameter *col_index*) to be converted to a factor by cutting. In order to transform that variable into a categorical vector, the following parameters should be supplied:

- *cut:* List of two or more unique cut points or a single number (greater than 2) giving the number of intervals into which the data is to be cut.
- *labels:* List of labels for the levels of the resulting category.
- *name_variable:* The new name for the resulting variable.

Output:

1. **data_norm_final:** A dictionary with the data types as keys and the resulting pre-processed data as values.

4.3 EXPLORATORY ANALYSIS

The exploratory analysis stage helps the user to have a better comprehension of the data at hand. It resorts to boxplots and barplots to better visualize a desired column, as well as a summary of the data. Additionally, heatmaps can be created for a specific omics or both omics, and PCA can be performed to help the user seek and understand the inherent structure, select relevant features to summarise the data and facilitate its interpretation. Another step involves differential expression of both or a selected omics, as an important step for feature selection. The machine runs the script **Analysis.py**, until the last function also executing **Annotation.py**.

- **Analysis.py**

  Input:

  1. **data_norm_final:** Dictionary obtained from the **Preprocessing.py** script.
  2. **items:** Dictionary with the type of exploratory analysis the person wants to do (barplot or boxplot) as key and the selected column name or index as value: if the value is a list of columns then each one is perceived as a different output; if the value is a tuple then the output will be a measure of the relation between the first and second element.

  First, using the function *"exploratory()"* the script uses the data obtained from the pre-processing stage and the dictionary *"items"* to perform the barplots and boxplots selected by the user.

  Next, the user has the option to request a brief summary of the data. The function *"summary()"* calculates the statistics of the omics data selected. It informs about the minimum value, the 1^{st} quartile (25^{th} percentile), the median, mean, 3^{rd} quartile (75^{th} percentile), and maximum value.

  After that, using the function *"pheatmap()"*, a heatmap can also be created, which gives a graphical representation of the data with colors, indicating whether there is a high degree of correlation or not between the different features. The function needs two parameters: *which_data*, the name of the omics we want to select or "ALL", for all the omics; and *columns* a list or a single value of the column we want to include in the heatmap for a better visualization.

Furthermore, the function *"pca()"* executes the PCA algorithm, a dimensionality reduction algorithm that grants the user the ability to see the inherent structure of each omics, select relevant features to summarise the data and facilitate its observation. This function also needs the following parameters to be executed:

- *which_data:* string of the name of the omics wanted or string "ALL" to execute for all the omics;
- *variable:* a single value - the wanted column - or a dictionary of different variables and how its represented in the plot, example {"shape":"Variety", "color":"Vintage"}.

Finally, the function *"differential_expression()"* grants the user the ability to execute differential analysis to the omics type desired, using the parameter *which_data*, and the outcome we want to focus on (parameter *y_pred* indicating the name of the column in metadata we want to analyse). Furthermore, *filter_results* allow the user to select the number of features to keep for further analyse (feature selection). However, to have the biological function of the features studied, the dataframe of the annotation file is necessary (parameter *annotation*). If working with transcriptomics of *Vitis vinifera*, a default file can be used running the **Annotation.py** script.

Output:

1. **top:** The top most differential expressed features. The length of the top parameter depends on the number inserted in the parameter *filter_results*.

– **Annotation.py**

Input:

1. **default:** A Boolean value that indicates whether the user wants to use the default annotation file, the PN40024 *Vitis vinifera* reference genome and annotation from http://plants.ensembl.org/Vitis_vinifera/Info/Index.
2. **annotation_file:** The file path of the annotation file, in case the user has a specific file. This file should contain the ID, the UniProtKB and the annotation/function of that feature. The header should contain the following columns as "ID", "UniProtKB" and "Annotation".

This script runs the function *"read_annotation()"*. If *default* is *True*, the function reads the default file from *Vitis vinifera*.

On the other hand, if *default* is *False*, then the function reads the *annotation_file*, considering certain parameters, namely:

- *header*, the default value is 0, indicating the row where the header is;
- *feature_col*, column (0-indexed) to use as the row labels of the dataframe, the default value is 0;
- *skip_rows*, Line numbers to skip (0-indexed) or number of lines to skip (int) at the start of the file. The default is 1.

Output:

1. **annotation:** dataframe with the columns ID, UniProtKB and Annotation to use in differential expression analysis, function *"differential_expression()"* of **Analysis.py** script.

4.4 INDIVIDUAL OMICS ANALYSES

We first need to understand how each omics influences the final result to clearly highlight a more holistic perception of the organism's metabolism and the fundamental mechanisms that lead to different phenotypes with multiomics integration, to can compare the different methods (individual and multiomics integration analysis), comprehend the insights obtained for each analysis, and verify the advantageous of multiomics integration. Therefore, the **Machine_Learning.py** script starts by analysing each omics with different models, preparing and saving them for multiomics analysis, allowing to compare the insights obtained by individual and multiomics analysis and understand the advantages obtained by multiomics integration.

- **Machine_Learning.py**

  Input:

  1. **top:** the output of the **Analysis.py** script.
  2. **data_norm_final:** The dictionary obtained from the **Preprocessing.py** script.
  3. **y_pred:** The name or index of the metadata column we want to analyse/predict.

  The first function *"train_models()"*, depending on *which_data* the user chooses, uses the *data_norm_final* parameter to produce the final dataset, where the rows correspond to the samples, and the columns to the features for that specific omics type, including the *y_pred* the user wants to analyse. This method returns a dictionary *train_test_datasets* as output, that for each key (the omics type) includes, as value, a list containing the dataset, the trainData and testData.

  Next, the *"SVM()"*, *"RF()"* and *"ANN()"* functions use the output from *"train_models()"* to execute the individual omics analyses. The output is written to a file in the user's directory, containing the dimension of the dataset, train dataset and test dataset and also the error metrics, precision, recall, PECC and the confusion matrix.

  Lastly, the *"save_data()"* function creates the dataset concatenated with the two omics: first, it saves the dataset into the file named *dataset_concat.xlsx*, and then organises the data in a list with the train and test dataset for the concatenated dataset, which will be used by the concatenated-based integration models. Furthermore, it builds another list for the other types of integration, dividing the omics and creating a train and test dataset for each omics.

  Output:

  1. **res_concat:** A list containing the train and test dataset of the concatenated dataset.
  2. **res_multi:** A list consisting of the train and test datasets for each omics and the Y_train and Y_test.
  3. **omic_1:** Dataset for the first omics.

4. **omic_2:** Dataset for the second omics.
5. **Y:** The column the user wants to analyse.

4.5 MULTIOMICS INTEGRATION

When the individual omics analyses is done, the machine executes different methods for each type of multi-omics integration (concatenation, transformation and model-based). In the first sub-stage (a), the machine executes supervised learning multiomics algorithms running the **Multiomics_integration.py** script, and in the second sub-stage (b) the machine performs unsupervised learning multiomics algorithms by running the **Unsupervised_Learning.py** script.

(a) **Multiomics_integration.py**

Input:

1. **res_concat:** A list containing the train and test dataset of the concatenated dataset, obtained from the *"save_data()"* function from the **Machine_Learning.py** script.
2. **res_multi:** A list consisting of the train and test datasets for each omics and the Y_train and Y_test, obtained from the *"save_data()"* function from the **Machine_Learning.py** script.
3. **omics_1:** Dataset for the first omics, given as output from the **Machine_Learning.py** script.
4. **omics_2:** Dataset for the second omics, given as output from the **Machine_Learning.py** script.
5. **Y:** Output feature the user wants to analyse, given as output from the **Machine_Learning.py** script.
6. **y_pred:** The name or index of the column to analyse/predict.
7. **train_test_datasets:** Dictionary obtained executing the function *"train_models()"* from the **Machine_Learning.py** script.
8. **omics:** A tuple with the name of the two omics selected for multiomics analyse.

The user has the option to choose which model wants the machine to run, however the **Multiomics_integration.py** script divides itself into three steps. The first step is concatenation-based integration. There are seven models to chose from:

- *DIABLO():* Performs DIABLO from the *mixOmics* package.
- *SMSPL():* Executes SMSPL model referred in the article from Yang and Yan-Qiong;
- *Stack_Generalisation():* Runs stacked generalisation using the package *h2o*;
- *Lasso_Regression():* Executes lasso regression from the package *glmnet*;
- *SVM()*, *RF()*, and *ANN()* functions that perform SVM, RF and ANN from the package *caret*.

Then, for the transformation-based the user has the following options:

- *SNFtool():* Executes model from the package *SNFtool*;
- *CAN_TBI():* Performs Graph-CAN from package *MDIntegration*;
- *RVM_Ada_TBI():* Runs Kernel-Integrated RVM also from the package *MDIntegration*.

Lastly, for the model-based integration, the machines focuses on ensemble classifiers with hard and soft voting, using several function from the Python library *scikit-learn*.

- *Ensemble_classifier():* Ensemble Classifier with different ML algorithms.

  For this function, the user can input the option and the voting strategy they prefer in the parameter *option* and *voting* respectively. The five options for the parameter *option* are:

  - **Option 1** Ensemble Classifier with two SVM models;
  - **Option 2** Ensemble Classifier with different models (SVM, ANN, Decision Trees and Guassian Naive Bayes (NB));
  - **Option 3** Ensemble Classifier with many ANNs;
  - **Option 4** Ensemble Classifier with NB (combination of models using voting classifier with NB, recommended for a model-based integration approach, according to Lin and Lane).
  - **ALL**: executes all options.

Regarding the parameter *voting* the input values can be *"Soft"*, *"Hard"* and *"ALL"*.

Output:

1. **Text files** (.txt) and **pictures** (.png) with the error metrics for all the models generated.

(b) **Unsupervised_Learning.py**

Input:

1. **train_test_datasets:** Dictionary obtained executing the function *"train_models()"* from the **Machine_Learning.py** script.
2. **omics:** A tuple with the name of the two omics selected for multiomics analyse.
3. **y_pred:** The name or index of the output column to analyse/predict.

The **Unsupervised_Learning.py** script provides the user three algorithms to perform unsupervised learning in multiomics data. The three algorithms are executed using the following functions:

- *MFA():* A concatenation-based integration algorithm, that can be identified in the *FactoMineR* package for the R language. The user need to insert one parameter to execute this function:
  - *type_:* Represents the type of variables in each group (the groups are separated into omics1, omics2 and Y). By default, the first two groups are numeric and the Y group is categorical, ["c","c","n"]. Allowed values include "c" or "s" for quantitative variables. If "s", the variables are scaled to unit variance; "n" for categorical variables;"f" for frequencies (from a contingency tables).

- *NEMO:* A transformation-based integration method designed in R and developed by Rappoport and Shamir. To run this function the following parameters should be given:
  - *num_clusters:* Number of clusters, if not given NEMO decides the values itself.
  - *num_neighbours:* Number of nearest neighbours, if not given NEMO decides the values itself.
  - *K:* The number of neighbors to use for each omics. It can either be a number, a list of numbers or NA. If it is a number, this is the number of neighbors used for all omics. If this is a list, the number of neighbors are taken for each omics from that list. If it is NA, NEMO chooses its value.
- *BCC:* A model-based integration method develop by Lock and Dunson. The following parameters should be given, to execute this function:
  - *K:* The (maximum) number of clusters. Default value is 10;
  - *a and b:* a and b are hyperparameters for the Beta(a,b) prior distribution on alpha. Default values are 1 for both.
  - *IndivAlpha:* Indicates whether the alpha should be separate for each data source ("TRUE" or "FALSE"). Default value is "False";
  - *Concentration:* Dirichlet concentration parameter for the overall cluster sized. Default value is 1.
  - *NumDraws:* Number of MCMC draws (NumDraws/2 is used as the "burn-in"). Default value is 1000.

Output:

1. **Text files** (.txt) and **pictures** (.png) with the error metrics for all the models generated.

5

RESULTS AND DISCUSSION

The first step and the most important in our project was plant data collection. Despite many searches in omics databases, the ideal plant dataset was almost impossible to find. Although there were several *Vitis vinifera* omics datasets, only a few of them had a sufficient number of samples or the same samples analysed for different types of omics, which is necessary for this project to achieve good results. This limitation was identified for other plants too.

Nonetheless, as mentioned in chapter 3, two omics datasets of transcriptomics and metabolomics from *Vitis vinifera* were selected as **Case Study I** and two omics datasets of transcriptomics and fluxomics from *Arabidopsis thaliana* were selected for the **Case Study II**. These omics were then pre-processed, individually analysed and integrated to provide a more holistic interpretation, improve our knowledge on plant's metabolic phenotypes and their underlying products, when facing environmental stresses and diseases.

5.1 CASE STUDY I: *Vitis vinifera*

5.1.1 *Pre-processing*

The two datasets of *Vitis vinifera*, as depicted in table 5, originated from the same 73 biological samples and a different number of features.

Missing Values

In the first step of pre-processing, we could see that the datasets did not contain any missing values; hence, no row was deleted. However, the number of rows (corresponding to the features) was still very high in the transcriptomics dataset, thus feature selection was performed.

Feature Selection

Three filter methods were applied to transcriptomics data. The first filter removed rows where the mean for such row was less than one, assuring that at least one transcript per cell existed, resulting in a transcriptomics dataset with 17720 features (12251 features removed). The second filter, the median

filter, calculated a median expression level for the dataset and selected which genes were expressed in each sample, to filter only genes present in at least 2 samples. The resulting dataset encompassed 13011 (4209 features removed). The final filter, the flat patterns filter, assumed that genes with very consistent values did not contribute to relevant information, filtering genes whose maximum ratio value over the minimum value of expression was greater than 2, resulting in a transcriptomics dataset containing 3447 features (deleting 9564 features), as shown in table 6. On the other hand, the metabolomics dataset encompassed 212 features; thus, it was considered small in comparison to the transcriptomics dataset and feature selection was not performed.

Normalisation and Scaling

Next, even though the metabolomics dataset was previously normalised by the authors, it was scaled not only to provide better visualisation and comprehension of the plots but also to be suitable for ML algorithms.

Data Discretization

Finally, as the variable that we wanted to predict, *Time Point*, for this case study was quantitative, we performed data discretization to a variable named **Berry_development**, as specified in the section 3.2.3. We opted for the two-factor variable, as most multiomics integration algorithms could not predict variables with more than two levels, for instance, *Stack Generalisation*, *Graph-CAN* and *Kernel-Integrated RVM and Boosted RVM* model.

5.1.2 *Exploratory Analysis*

Barplots and Boxplots

Barplots and Boxplots provide a better comprehension of the data, allowing to visualise data distribution and to comprehend the range of values. Information regarding the boxplots and barplots is available in the Case Study I section, in the supplementary figures A.

Therefore, we opted to analyse the *"Time Point"* as our predictive variable, as it allows studying the different development stages and having a better look into the metabolism of the berry during growth.

Heatmaps

Next, we provide a summary of the data confirmed that the metabolomics dataset was scaled. Two different heatmap analyses were performed. The first analysis identifies similarities within the same vintage year or within samples from different varieties in both datasets. As shown in supplementary figure S4, the transcriptomics heatmap allows visualization of samples regarding *Variety*, while the

heatmap analysis for the metabolomics dataset allows differentiating the vintage year from which the samples were collected.

On the other hand, the heatmap from the second analysis allowed identifying similarities regarding the different varieties, but also for the outcome we wanted to predict, which indicated the development stage of the berries (pre-Veraison or post-Veraison, including Veraison). As depicted in supplementary figure S4, both analysis can explain berry development.

PCA

Additionally, PCAs were carried out, as shown in Figure 12, which illustrates a PCA regarding the variety and development stage of berries. Data discretization of the factor *Time Point* into *Berry_Development*, allows demonstrating that the first two principal components (PC1: 71.41% and PC2: 18.58%) in the PC1-PC2 plot of the transcriptomics distribution, divided the samples into 3 main groups: pre-veraison berries of both varieties (Group 1), post-veraison *Pinot Noir* berries (Group 2) and post-veraison *Cabernet Sauvignon* berries (Group 3). Additionally, with the metabolomics dataset (figure 12B), we could see that PC1 (55.666%) can also differentiate metabolite content in two groups: berries in Early development (pre-veraison) and berries in Veraison and Late Ripening stage (post-veraison).

Thus, these results indicate that the cultivars in post-Veraison are different regarding gene expression, compared to the pre-Veraison phase, that shows a similar gene expression for each variety. Contrarily, in the metabolite expression, the pre-Veraison and post-veraison phases are very different.

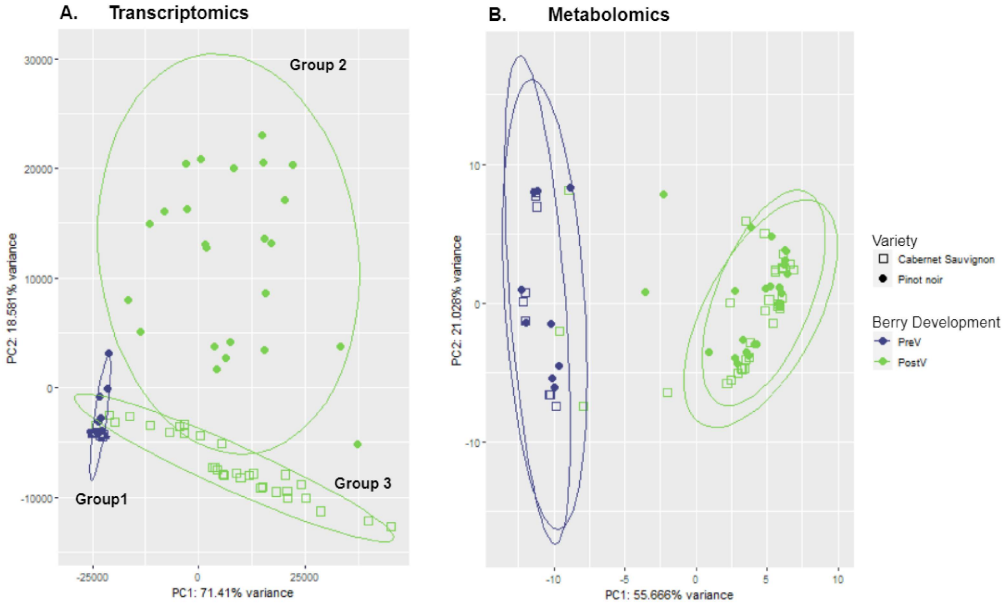


Figure 12: *Exploratory Analysis*. PCA of (A) transcriptomics and (B) metabolomics regarding Variety and Berry_Development. In (A) we see that the transcriptomics analysis divided the samples into 3 main groups: pre-Veraison berries of both varieties (Group 1), post-Veraison berries *Pinot Noir* berries (Group 2) and post-Veraison *Cabernet Sauvignon* berries (Group 3). In (B) the analysis differentiates the metabolite content in two groups: Early-development (pre-Veraison) berries and berries in Veraison and Late Ripening stage (post-Veraison).

In supplementary figure S4A, we can observe that the PC1-PC2 principal component analysis evaluates changes in gene expression regarding different varieties, showing that the second principal component (18.581%) distributed samples according to cultivar (*Pinot Noir* and *Cabernet Sauvignon*). Whereas metabolomics data does not provide a clear division between varieties.

On the other hand, in supplementary figure S5A, the principal components in the PC1-PC2 plot of transcriptomics regarding variety and vintage did not divide vintage years, but in PC1-PC2 of metabolomics dataset, we can see that the second principal component (PC2: 21.026%) distributed samples according to vintage years (2012, 2013 and 2014).

Differential Expression

The features of the transcriptomics dataset were narrowed further to match the number of features in the metabolomics dataset and facilitate the analysis by multiomics algorithms. We opted for performing differential expression of only the transcriptomics dataset. In table 10, we have the top 20 significantly expressed transcripts regarding the contrast between pre-Veraison and post-Veraison stages, and their respective annotation obtained with the PN40024 *Vitis vinifera* reference genome and description from http://plants.ensembl.org/Vitisvinifera/Info/Index.

Table 10: Top 20 up-regulated and down-regulated differential expressed genes from transcriptomics dataset.

Up-Regulated			Down-Regulated		
ID	**UniProtKB**	**Annotation**	**ID**	**UniProtKB**	**Annotation**
VIT_02s0154g00070	D7TN10	Abnormal floral organs	VIT_15s0048g01920	D7U7J4	F-box family protein
VIT_14s0083g01030	D7SMN6	MADS-box APETALA 1	VIT_18s0001g11930	F6H040	Thaumatin SCUTL2
VIT_17s0000g06880	F6GT30	Heparanase protein 2 precursor	VIT_03s0091g00260	F6H699	Zinc finger protein 4
VIT_09s0002g00960	D7TZQ4	Inter-alpha-trypsin inhibitor heavy chain	VIT_04s0044g01110	A5C0I8	Alcohol dehydrogenase 6
VIT_06s0004g07880	D7SJK7	Allergen	VIT_12s0059g01080	D7TEB6	Acid phosphatase/vanadium-dependent haloperoxidase
VIT_08s0007g07550	F6HLA6	GATA transcription factor 11	VIT_14s0006g02020	F6HSN0	NADH dehydrogenase (ubiquinone) Fe-S protein 1
VIT_09s0002g04260	D7U0H5	Unknown protein	VIT_03s0063g00730	F6HQC9	CXE carboxylesterase CXE10
VIT_13s0019g03040	F6HNN8	Indole-3-acetate beta-glucosyltransferase	VIT_13s0019g04410	D7TM83	Oligosaccharyltransferase complex subunit alpha (ribophorin I)
VIT_01s0026g01780	F6HPE9	Leucine-rich repeat transmembrane	VIT_06s0061g00550	D7SNC1	Xyloglucan endotransglucosylase/hydrolase 32
VIT_04s0044g01870	F6I0B3	Auxin efflux carrier	VIT_00s0207g00280	F6I234	WAK receptor protein kinase
VIT_13s0064g00890	F6HB61	Cellulose synthase CESA3	VIT_03s0091g00520	D7SXT3	Prolyl 4-hydroxylase
VIT_06s0004g05340	NA	Tropinone reductase	VIT_09s0002g01800	D7TZW9	Dihydrolipoamide S-acetyltransferase
VIT_14s0066g01650	F6HV18	Nodulin MtN21 family	VIT_16s0050g00910	F6H6E7	MATE efflux family protein
VIT_18s0001g10550	E0CP66	LHCA5 (Photosystem I light harvesting complex gene 5)	VIT_08s0007g03830	A5B118	fructose-bisphosphate aldolase cytoplasmic isozyme
VIT_14s0083g00850	D7SML9	Lipase GDSL 7	VIT_04s0023g03020	F6GWP9	Serine carboxypeptidase S28
VIT_13s0064g01030	D7T2Z6	Zinc finger (C3HC4-type ring finger)BIG BROTHER	VIT_08s0040g00430	D7TQ50	COP9 signalosome complex subunit 2
VIT_06s0004g03020	D7SKW9	Beta-galactosidase	VIT_06s0004g05890	F6GUV6	Ceramidase
VIT_05s0020g04880	D7T7A2	Seed specific protein Bn15D14A	VIT_07s0005g01250	A5AS62	Unknown protein
VIT_05s0020g02690	F6HDL7	Copper-binding family protein	VIT_14s0066g01600	F6HV14	NHL repeat-containing protein
VIT_09s0002g04080	F6HYD4	IAA9	VIT_07s0031g00460	F6H4J6	Nicotinate phosphoribosyltransferase

As in the work by Fasoli et al., we identified some insightful differential expressed transcripts, contrasting the pre-Veraison stage with the post-Veraison stage, as depicted in table 10. Looking at the up-regulated genes in pre-Veraison, we identified genes involved in fruit development and developmental processes, such as, cell differentiation and proliferation, enabling growth, which is the case of D7TN10 (Abnormal floral organs). Additionally, we identified F6GT30 (Heparanase protein 2 precursor) that enables beta-glucuronidase activity, notably prominent in young regions of developing organs and associated with cell elongation [194]. These genes are associated with the first growth phase characterized by pericarp enlargement, caused by cell division and elongation [178]. Moreover, this phase also accumulates organic acids but little sugar, explaining the up-regulation of genes involved in carbohydrate metabolism, TCA, carbon fixation and photosynthesis.

Organic acids are frequently formed during carbohydrate metabolism, especially during the TCA cycle or carbon fixation that occurs during the dark phase of photosynthesis. The differential expression analysis identified genes like D7SKW9 (Beta galactosidase) involved in the carbohydrate metabolic process and E0CP66 (LHCA5 (Photosystem I light-harvesting complex gene 5)) in photosynthesis.

Furthermore, this stage also accumulates tannins, hydroxycinnamates, and phenolic precursors, in which several of these compounds were up-regulated. Tannins describe a group of phenols present in *Vitis vinifera* that promote berry and seed colour, from which flavonoids, such as flavonols and anthocyanins, are present. Consequently, we identified several genes, e.g., F6I4E7 (UDP-glucose:flavonoid 7-O-glucosyltransferase) and D7TJX3 (flavonol synthase) related to flavonols, and F6H6Q6 (UDP-glucose: anthocyanidin 5,3-O-glucosyltransferase) related to anthocyanins formation. In their most basic form, these anthocyanins are called anthocyanidins, but when they bind with glucose, anthocyanins are formed [195]. In addition, hydroxycinnamates were also present, like D7TJI5 (Anthranilate N-hydroxycinnamoyl/benzoyltransferase) involved in the coumarin biosynthetic process. This type of hydroxycinnamates possess antimicrobial and antioxidant capacities [196]. Phenolic precursors like D7T5P6 (Polyphenol oxidase II, chloroplast precursor) were also identified.

Several auxin signal transduction components were also identified in up-regulated genes, like F6I0B3 (Auxin efflux carrier) and F6HNN8 (Indole-3-acetate beta-glucosyltransferase), which is in concordance with the article and the fact that auxin indole-3-acetic acid inhibits ripening.

This early development phase also requires genes involved in oxidative phosphorylation, like F6HDL7 (Copper-binding family protein) that enables protein kinase activity and ATP Binding. Several other genes were related to genetic information processing, for instance, F6HYD4 (IAA9), enabling ATP binding, and D7SMN6 (MADS-box APETALA 1), which regulates transcription, enabling DNA-binding.

Regarding the down-regulated genes in pre-veraison, we identified genes highly expressed related to the post-veraison grapes, and also the included Veraison phase. This second growth phase is defined by changes that make the fruit edible and alluring, including changes in the skin colour, accumulation of sugars, loss of organic acids and tannins. Therefore, since grape ripening takes place, is normal to find proteins related to ripening, for instance, F6I7I4 (MATE efflux family protein ripening responsive) and F6H0Y9 (Ripening regulated protein DDTFR18), putative ripening-related proteins.

Furthermore, we also identified differential expressed genes involved in iron ion binding, F6HSN0 (NADH dehydrogenase (ubiquinone) Fe-Sprotein 1)) and D7SXT3 (Prolyl 4-hydroxylase). Iron (Fe) is an essential element for the growth and reproduction of plants. The increase in Fe content improves the production of Reducing Sugar (RS), and a good RS/TCA ratio, that improves wine grape quality, if this ratio is decreased the ripening phase will be delayed. Plus, Fe promotes anthocyanin accumulation [197].

During ripening the berry also decreases its water content occurring berry softening, which is in concordance with differential expressed genes related to cell wall organization and biogenesis, e.g., D7SNC1 (Xyloglucan endotransglucosylase/hydrolase 32).

The accumulation of sugars, loss of organic acids and tannins, and synthesis of volatile aromas can be explained by the differential expression of A5BB118 (fructose-bisphosphate aldolase cytoplasmic) and other genes that participate in the glycolytic process and other occurrences of the carbohydrate degradation metabolism.

In this phase, we also identified several genes involved in transport, like F6H6E7 (MATE efflux family protein) that enables transmembrane transporter activity and antiporter activity, and also genes involved in detoxification, like D7TEB6 (Acid phosphatase/vanadium-dependent haloperoxidase).

Lastly, hormones also play an essential role in ripening and maturation, as ethylene and abscisic acid (ABA) induce ripening. Ethylene induces alterations in colour, aroma, texture and flavour besides other biochemical and physiological parameters in berry. A differential expressed gene identified in this analysis involved in the transcription factor of ethylene was F6I2P2 (Ethylene-responsive transcription factor ERF105), among others. Ethylene production occurs just before veraison so it is normal to find genes related to ethylene in the PreV phase, although they can also be identified differentially expressed in the PostV phase, like D7TFI7 (Ethylene-responsive transcription factor ERF114).

On the other hand, ABA helps in the accumulation of anthocyanins and sugars and the up-regulation of genes involved in ripening and is also differentially expressed in this analysis as we can see by the gene A5BHW6 (ABA-responsive protein (HVA22a)) in the up-regulated genes and F6HW11 (GRAM domain-containing protein / ABA-responsive) in the down-regulated genes [178].

5.1.3 *Classic Machine Learning Methods*

Classic ML models were used, in this project, both for the individual omics analysis and for the multiomics integration analysis, such as the SVM, RF and ANN models.

In the individual omics analysis stage, we used the *"train()"* function of the package *caret* in R language to split both our data individually into train and test datasets (70% train and 30% test). Table 11 shows the dimensions of the original dataset and the resultant train and test datasets. Then, using only the train dataset, we trained the different models and evaluated them with 10 fold and 3 repetitions CV, and used the test dataset to predict the output. On the other hand, for the multiomics integration analysis, the three models were created using the library *scikit-learn* for Python and 5-fold CV was implemented.

This way, we could determine which classical ML algorithm better suited our data, having an idea of how well the individual datasets predict the outcome compared to the multiomics integration analysis, and determine the features for each dataset that best explain the outcome. Table 12 in the supplementary tables contains the different error metric values (Accuracy, Recall and precision) obtained for both datasets, for the individual omics analysis, whereas table 18 shows the results for all the error metrics in the multiomics integration analysis, in the three models.

Table 11: *Individual Omics Analysis*. Dimensions (samples, features) of the original transcriptomics and metabolomics datasets and their respective train and test datasets.

	Transcriptomics	**Metabolomics**
Original	(73,213)	(73,213)
Train	(50,213)	(50,213)
Test	(23,213)	(23,213)

Table 12: Values of the different error metrics (Accuracy, Recall and Precision) for each model (SVM, RF and ANN) for both the transcriptomics and metabolomics datasets.

Model	**Metrics**	**Transcriptomics**	**Metabolomics**
SVM	Accuracy	0.95	0.9
	Recall	0.8571	1
	Precision	1	0.75
RF	Accuracy	1	0.8
	Recall	1	1
	Precision	1	0.6
ANN	Accuracy	0.8	0.9
	Recall	0.4286	1
	Precision	1	0.75

Support Vector Machine

- *Individual Omics Analysis*

The SVM model has very good performance, the results are shown in table 12. It as an accuracy, recall and precision of 0.95, 0.8571 and 1 in the transcriptomics dataset, and 0.9, 1, 0.75 in the metabolomics dataset, respectively. Figure 13 shows the ROC plot and AUC value for both datasets.

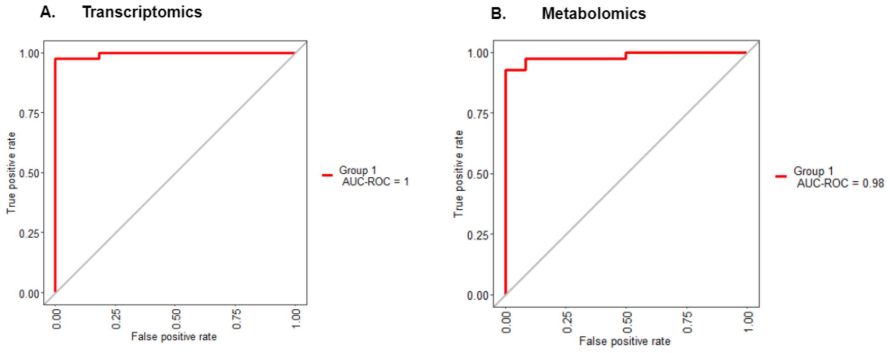


Figure 13: *Individual Omics Analysis.* ROC curve of the SVM model for the (A) transcriptomics dataset, with an AUC value of 1 and (B) metabolomics dataset, with an AUC value of 0.98.

- *Multiomics Integration Analysis*

Grid search and random search were executed to obtain the best estimator for our data in the SVM model. The best estimator was obtained with grid search and the hyperparameters were: C=1, gamma=0.01, kernel='linear'. The model's accuracy, precision and recall for the grid search were 0.933, 0.666 and 1.0, respectively. Furthermore, we calculated the ROC curve, depicted in figure 14, for grid search and the AUC value was 0.96. Lastly, we obtained the top 10 most important features that explain the outcome of the model. Table 13 shows the top 10 features for the grid search model, which has better accuracy.

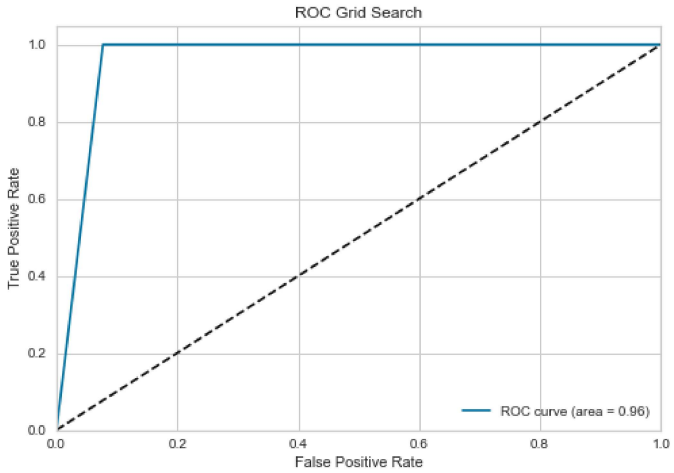


Figure 14: *Multiomics Analysis Integration.* ROC curve for the SVM model in the concatenation-based integration. Grid Search ROC curve with an AUC value of 0.96.

Table 13: Most important features regarding the concatenation dataset for the SVM model.

Most important features			
Variable	**UniProtKB**	**Annotation**	**Score**
VIT_11s0016g05840	F6HHB3	Protease inhibitor/seed storage/lipid transfer protein (LTP)	0.00268498
VIT_19s0090g01570	F6HEM1	Ribosomal protein S8 (RPS8A) 40S	0.00201202
VIT_05s0020g02690	F6HDL7	Copper-binding family protein	0.00192066
VIT_18s0001g01490	A5ASV7	Oxidoreductase N-terminal domain-containing	0.00159169
VIT_13s0064g00890	F6HB61	Cellulose synthase CESA3	0.00131851
VIT_12s0028g01080	A5B1D3	Photosystem II oxygen-evolving complex precursor, 32kda PSBP	0.00117213
VIT_06s0004g03240	A5AFS1	Elongation factor 1-alpha 1	0.00100799
VIT_19s0014g03850	A5BX41	Cytochrome B6-F complex iron-sulfur subunit, PETC	0.000736657
VIT_11s0016g01230	D7TCG0	12-oxophytodienoate reductase 3	0.000499583
VIT_02s0025g03540	F6HUC8	Tubulin beta-6 chain	0.000497651

Random Forest

- *Individual Omics Analysis*

For the RF model, the transcriptomics dataset had an accuracy, recall and precision of 1, which indicates that it correctly classifies all samples but it might be overfitted, even though cross-validation was performed, although it is not possible to prove unless we test with new data. On the other hand, the metabolomics dataset returned accuracy, recall and precision of 0.8, 1 and 0.6, respectively. Figure 15 depicts the ROC plot and AUC value for both datasets.

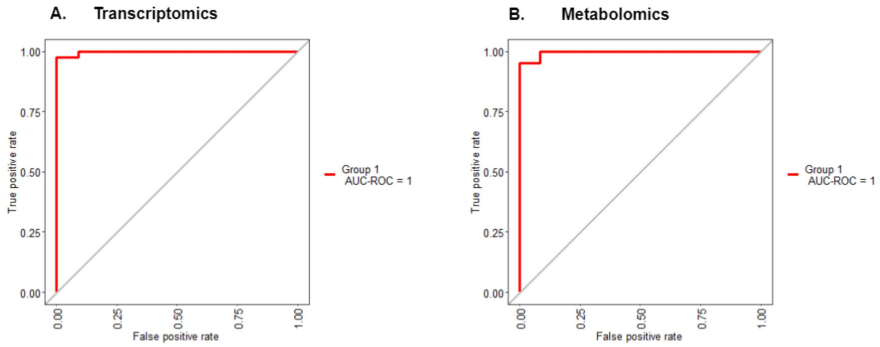


Figure 15: *Individual Omics Analysis.* ROC curve of the RF model for the (A) transcriptomics dataset, with an AUC value of 1 and (B) metabolomics dataset, with an AUC value of 1.

- *Multiomics Integration Analysis*

For the concatenation-based integration, grid and random search were executed to obtain the best estimator for our data in the RF model. The best estimator had the following hyperparameters: bootstrap=False, max_depth=2, max_features=4, min_samples_leaf=2. Additionally, we identified the features with greater importance. The model's accuracy, precision and recall were 0.87, 0, 0, respectively. Figure 16 shows the ROC curve for the RF model with an AUC value of 0.5. Table 14 depicts the most important features for this model.

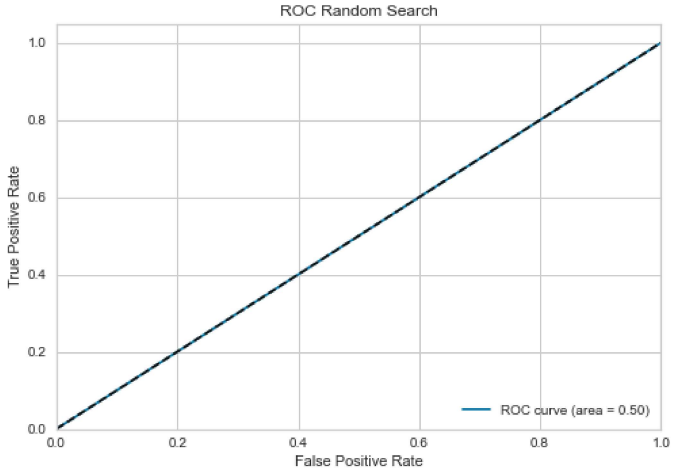


Figure 16: *Multiomics Analysis Integration.* ROC curve for the RF model in the concatenation-based integration. The AUC value corresponds to 0.5.

Table 14: Most important features regarding the concatenation dataset for the RF model.

Most important features			
Variable	**UniProtKB**	**Annotation**	**Score**
VIT_11s0016g05840	F6HHB3	Protease inhibitor/seed storage/lipid transfer protein (LTP)	0.0306963
VIT_04s0044g01870	F6I0B3	Auxin efflux carrier	0.0303425
VIT_06s0004g07880	D7SJK7	Allergen	0.0290973
VIT_01s0026g01780	F6HPE9	Leucine-rich repeat transmembrane	0.0289725
VIT_18s0001g09390	E0CP50	Protein phosphatase 2C	0.0205442
VIT_01s0026g00330	D7TNP7	NHL repeat-containing protein	0.02
VIT_06s0004g07310	F6GUL1	Indole-3-acetate beta-glucosyltransferase	0.0199708
VIT_18s0001g08240	F6GZM2	Leucine-rich repeat transmembrane protein kinase	0.0196217
VIT_18s0001g09070	F6H199	ZCW7 protein	0.0196005
VIT_16s0013g01510	D7U7B0	WD-repeat protein 8	0.019315
fructose	–	–	0.0192617

Artificial Neural Network

- *Individual Omics Analysis*

Lastly, the ANN model for the transcriptomics dataset exhibited accuracy and precision of 0.8 and 1, respectively, and a lower value of recall (0.4286). This means that the model is identifying few pre-Veraison cases. In turn, the metabolomics dataset presented an accuracy, recall and precision of 0.9, 1 and 0.75. Figure 17 illustrates the ROC plot and AUC value for both datasets. This model obtained better results for the metabolomics dataset, while the others performed best with transcriptomics data.

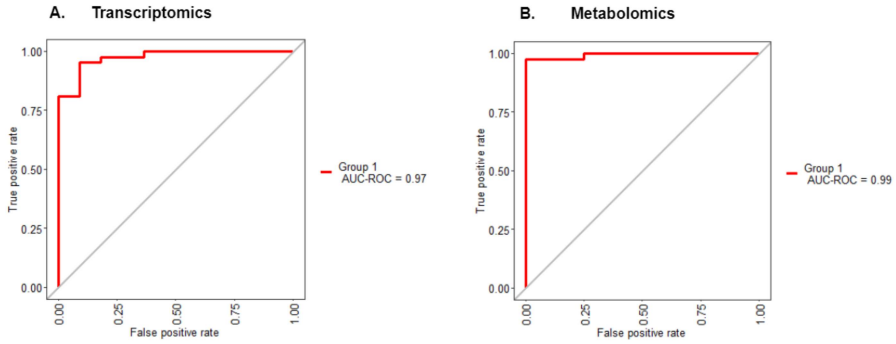


Figure 17: *Individual Omics Analysis*. ROC curve of the ANN model for the (A) transcriptomics dataset, with an AUC value of 0.97 and (B) metabolomics dataset, with an AUC value of 0.99.

- *Multiomics Integration Analysis*

Lastly, we implemented an ANN model. Using the random search and grid search we selected the best estimator, with the following hyperparameters: activation='tanh', alpha=0.001, max_iter=900,

solver='lbfgs'. As grid search took too long to run, we opted to use only the random search results. The model's accuracy, recall and precision for the best estimator were 0.733, 0.5 and 0.25, respectively. Then, we used the Python package *LIME* to calculate feature relevance. However, only the important features for each specific sample at each time are returned. Still, we can see which features are more important to consider for each class. Figure 18 displays the ROC curve, which the AUC value corresponds to 0.58. In table 15, we find the features with greater importance.

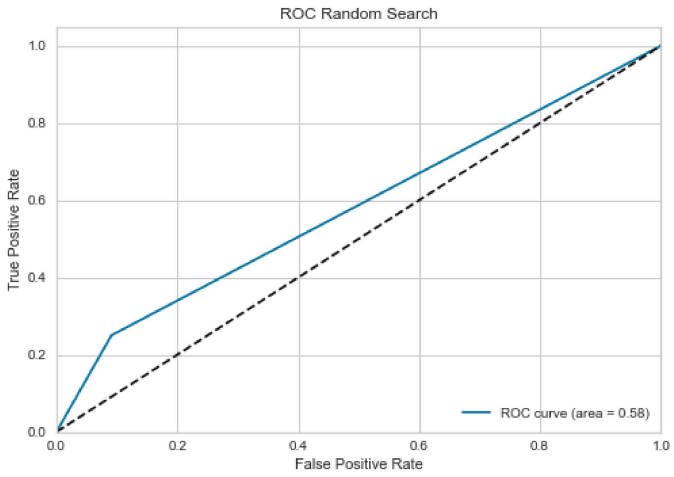


Figure 18: *Multiomics Analysis Integration.* ROC curve for the ANN model in the concatenation-based integration. The AUC value corresponds to 0.833.

Table 15: Most Important features regarding the concatenation dataset for the ANN model.

Most important features		
Variable	**UniProtKB**	**Annotation**
VIT_05s0020g02690	F6HDL7	Copper-binding family protein
VIT_19s0090g01570	F6HEM1	Ribosomal protein S8 (RPS8A) 40S
VIT_13s0064g00890	F6HB61	Cellulose synthase CESA3
VIT_16s0022g00670	F6HAU0	Vacuolar invertase 1, GIN1
VIT_14s0083g01110	D7SMP3	Brassinosteroid-6-oxidase
VIT_18s0001g01490	A5ASV7	Oxidoreductase N-terminal domain-containing

Feature Relevance

- *Individual Omics Analysis*

Regarding the individual omics analysis, looking at the most important features in the transcriptomics dataset that could explain the outcome, for each model (SVM,RF and ANN - supplementary tables S1,

table S3 and table S5 respectively) we identified transcripts common to some models. Table 16 depicts the most common features identified in the three individual omics models that were identified in at least two models or all models, their respective annotation, and the role they play in berry development.

Analysing the results, we identified similar biological processes, like plant development, cell population proliferation and cell growth, ion binding and oxidoreductase activity, chlorophyll, plant cell wall organization and biogenesis, regulation of transcription and auxin-related, which indicate some of the most important events in the pre-Veraison stage. For instance, the cell wall of the berry is still hard and green, explaining the cell wall biogenesis and the chlorophyll. Additionally, this stage is also notable for the intense cell proliferation events and the auxin hormone that inhibits ripening.

Regarding the metabolomics dataset (Supplementary tables S2, table S4 and table S6, respectively) the most important metabolites that explain the outcome for all three models were fructose, tartaric acid, malic acid, malvidin-3-o-glucoside, glucose, peonidin-3-glucoside, sucrose, stearic acid and benzenemethanol. There are several organic acids, like tartaric and malic acid that are the most prevalent organic acids in the pre-Veraison stage. On the other hand, glucose, fructose, sucrose and benzenemethanol are compounds identified in the post-Veraison phase that indicate the increase in sugar concentration and accumulation of aroma and flavour compounds.

- *Multiomics Integration Analysis*

Looking at the results obtained by the three classic ML models for the concatenation-based integration approach, we identified 5 transcripts in common with at least two models. Table 17 depicts the most relevant features in the three classical ML models for multiomics integration analysis identified in at least two models, the respective annotation, and the role they play in berry development. The biological processes discovered were related to plant growth and development, ribosomal protein, oxidative phosphorylation, nucleotide-binding and cell wall biogenesis, cellulose biosynthetic process and cell wall organization. With these results, we identified new essential functions like nucleotide-binding and ribosomal protein, that appear in other publications as overexpressed functions in berry development [198, 199]. Furthermore, the other functions were previously mentioned in the prior analyses, such as plant development, oxidoreductase activity and plant cell wall biogenesis, cellulose biosynthetic process and cell wall organization.

On the other hand, for the metabolomics dataset, only one metabolite was considered as a relevant feature to explain the outcome. It was the case of fructose, indicated by the RF model, a sugar identified in the post-veraison phase in grapes.

Thus, comparing the three classic ML models and the results for individual omics analysis and multiomics integration analysis, we determined that the individual omics analyses can indicate a greater number of biological processes and metabolites present in grape development and relevant for predicting the outcome. However, both analyses identified two transcripts in common, F6HHB3 (Protease inhibitor/seed storage/lipid transfer protein (LTP)) and F6HB61 (Cellulose synthase CESA3) referring to plant growth, development, and importance for fruit ripening and cell wall biogenesis, cellulose biosynthetic process and cell wall organization, key roles in berry development.

Table 16: *Individual Omics Analysis*. Feature Relevance. Most common transcripts in the three classical ML models, their annotation, respective function in berry development and which models have in common.

Transcripts (UniProtKB)	Annotation	Function	Models	Ref.
F6HB61	Cellulose synthase CESA3	Involved in plant-type primary cell wall biogenesis and organization and cellulose biosynthetic process	All	[200]
D7T7A2	Seed specific protein Bn15D14A	Maintenance of cell morphology or regulation of cell growth and development	All	[201]
F6I6F6	Lateral organboundaries protein 1	Plant organ and pollen, plant regeneration, photomorphogenesis and pathogen response	SVM and RF	[202]
F6HNN8	Indole-3-acetate beta-glucosyltransferase	Involved in the regulation of auxin levels	SVM and RF	[203]
D7ST07	Porphobilinogen deaminase, chloroplast precursor	Production of chrolophyll, heme, siroheme and phytochromobilin in plants, affecting vegetative and reproductive development when in state of deficiency	SVM and RF	[204]
D7T2Z6	Zincfinger (C3HC4-type ring finger) BIG BROTHER	Organ development, protein ubiquitination and negative regulation of cell population proliferation	SVM and RF	[205]
F6HQG7	CYP82C1p	Involved in defense responses enabling iron ion binding, heme binding and oxidoreductase ativity	SVM and RF	[200]
D7UA38	LNG1 (LONGIFOLIA1)	Involved in cell growth	SVM and RF	[200]
F6GVV2	Alpha-1,4-glucan-proteinsynthase 1	Involved in plant-type cell wall organization or biogenesis, and UDP-L-arabinose metabolic process	SVM and RF	[200]
F6HHB3	Protease inhibitor/seed storage/lipid transfer protein (LTP)	Plant growth and development and fruit ripening	SVM and ANN	[206]
F6H5G9	GPRI1 (GOLDEN21)	Involved in the positive regulation of transcription, DNA templated	SVM and RF	[200]
F6HYD4	IAA9	Auxin/indole-3-acetic acid (Aux/IAA) protein, essential for plant growth and development involved in auxin-activated signaling pathway and regulation of transcription, DNA-templated.	SVM and RF	[200]

Table 17: Multiomics Integration Analysis. Feature Relevance. Most common transcripts in the three classical ML models, their annotation, respective function in berry development and which models have in common.

Transcripts (UniProtKB)	Annotation	Function	Models	Ref.
F6HHB3	Protease inhibitor/seed storage/lipid transfer protein (LTP)	Involved in plant growth and development, and fruit ripening	SVM and RF	[206]
F6HEM1	Ribosomal protein S8 (RPS8A) 40S	Ribosomal protein that may suggest a shift in protein synthesis, that occurs in the transcriptome reprogramming during berry development	SVM and ANN	[198]
F6HDL7	Copper-binding family protein	Involved in oxidative phosphorylation, that enables protein kinase activity and ATP Binding	SVM and ANN	[200]
A5ASV7	Oxidoreductase N-terminal domain-containing	Enables nucleotide binding, a biological process notably identified in berry development	SVM and ANN	[199]
F6HB61	Cellulose synthase CESA3	Involved in plant-type primary cell wall biogenesis, cellulose biosynthetic process and cell wall organization	SVM and ANN	[200]

5.1.4 *Novel models for Multiomics Integration Analysis*

For the multiomics integration stage, we studied three different integration based approaches: concatenation-based, transformation-based and model-based integration. The multiomics integration stage uses both datasets simultaneously, which may result in better and more holistic outcomes, and give interesting variable relations. In this stage, we are exploring different models and integrating approaches and try to complement berry development studies, giving more insightful information. Table 18 shows the different algorithms in each integration approach and the corresponding accuracy, recall, precision and AUC values.

Due to the poor results, in the models suggested in chapter 3, we opted to show results obtained for the multiomics integration models. For the concatenation-based approach, we studied DIABLO, SMSPL, Stack Generalization (Ensemble) and the three classical ML methods, removing the lasso regression model. Regarding the transformation-based models, we analyzed the SNFtool model but left out the graph-CAN and kernel- Integrated RVM and Boosted-RVM models. Lastly, for model-based integration, we selected the first of the four options. However, all models can be run in the developed pipeline. The link can be identified in the section 3.7.

Table 18: *Multiomics Integration Analysis*.Results of the error metrics: PECC (accuracy), Precision, Recall and the AUC values for all models executed.

Model	**Metrics**			
	Classification Metrics			
	PECC	**Precision**	**Recall**	**AUC Value**
DIABLO	0.85	0.66	0.875	0.8375
SMSPL	0.952	1	0.833	0.969
Stack Generalisation (Ensemble)	0.9523	1	1	0.9875
SVM	0.933	0.666	1.0	0.96
RF	0.87	0.87	1	0.5
SNFtool	0.933	0.916	1	0.875
Ensemble Classifier with different ML algorithms (Hard and Soft Voting)	-	-	-	-
Option 1 (Soft Voting)	0.93	1	0.92	0.96
Option 1 (Hard Voting)	1	1	1	-

Concatenation-Based Integration

- **DIABLO**

First, we set our data as a list of data matrices matching the same samples in the rows. The omics datasets were named as blocks, **block omics1** corresponding to transcriptomics and **block omics2** equivalent to metabolomics dataset.

For the matrix design, a symmetrical matrix, that indicates which blocks are connected (ranging from 0 to 1) is required. We first opted to examine the correlation between the different blocks via the modelled components, and therefore, we executed a partial least square regression and a Sparse Partial Least Squares regression to find the best value to input in the design matrix. We opted for a correlation of 0.88.

Regarding the tuning of the parameter "number of components" (the number of sufficient components to discriminate all phenotype groups), the mixOmics package suggests first fitting a DIABLO model without variable selection to assess the global performance and obtain the number of components for the final DIABLO model. The plot in figure 19 shows the performance of the executed DIABLO model.

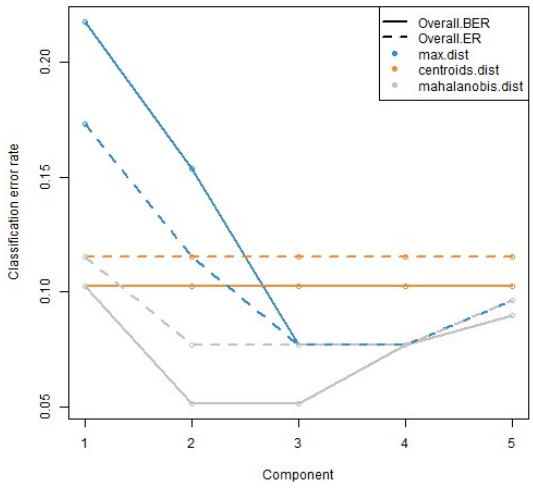


Figure 19: *Multiomics Integration.* Diablo model performance for number of components tuning. Looking at the weighted vote of overall balanced error rate (BER) and the centroids distance (centroids.dist), the model obtained a total of 5 components for the final model.

Looking at the performance plot, we extract the number of components regarding the weighted vote of overall balanced error rate (BER) and the centroids distance (centroids.dist), obtaining a total of five components for the final model.

Then, we executed the final model with the *"block.splsda()"* function using the five components

and obtained the following sample plot that projects the samples into the space extended by the components of each block (figure 20).

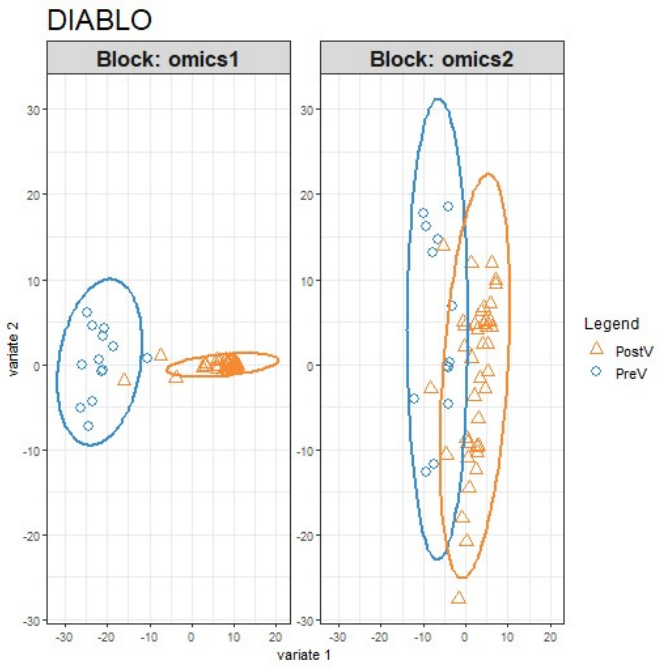


Figure 20: *Multiomics Integration.* Plot resultant of the *"plotIndiv()"* function of mixOmics, that projects each sample into a space extended by the components of each block. Block: omics1 concerning the transcriptomics dataset and block: omics2 regarding the metabolomics dataset.

As shown in figure 30, variate 1 explains with success the outcome of berry development in the transcriptomics dataset. The same can also be shown in the arrow plot in supplementary figure S8.

MixOmics also allows the execution of variable plots to visualize and analyze the associations between the selected variables. The variable plot (figure 21) depicts the variables from all blocks selected on components 1 and 2.

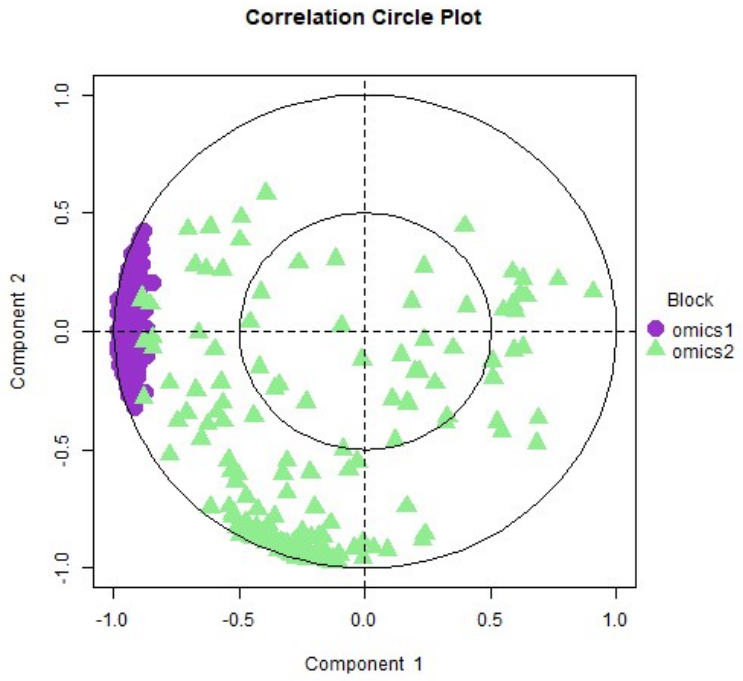


Figure 21: *Multiomics Integration.* Plot resultant of the *"plotVar()"* function of mixOmics, that allows the visualization and analyse of the variations between selected variables.

As clusters of points indicate a strong correlation between variables we can see that the transcriptomics dataset variables are more correlated than the metabolomics dataset variables.

On the other hand, the Circos plot in figure 22 can provide a good insight into the correlations between variables of different types. However, the number of samples makes the interpretation difficult. Nevertheless, we can see that the features from the metabolomics dataset explain both the preV and postV stages, while in the features from the transcriptomics dataset, we cannot see a clear distinction, and if we take a closer look we see that the space between the blue (PreV) line and orange (PostV) line is larger in the transcriptomics dataset, which means this dataset explains better the preV stage.

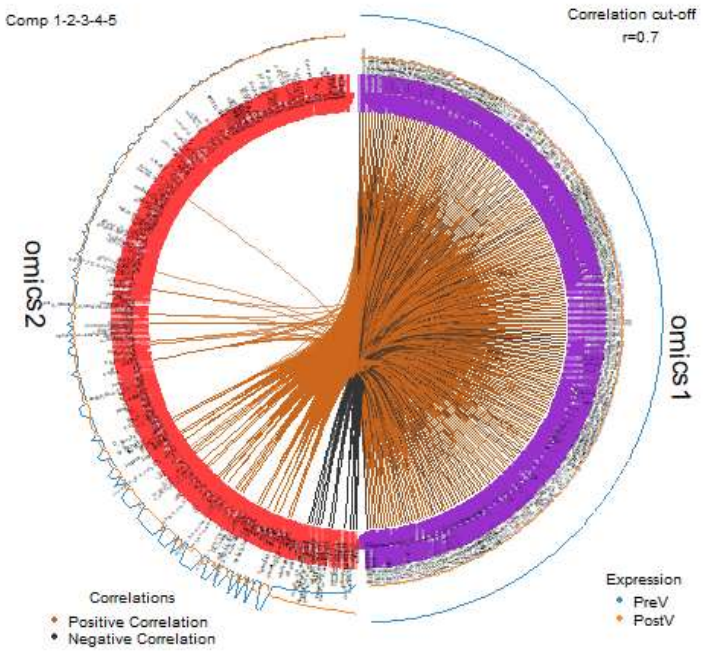


Figure 22: *Multiomics Integration.* Plot resultant of the *"circosPlot()"* function of mixOmics, that allows the visualization and analyse of the variations between selected variables.

In supplementary figure S9, we can see a network plot made using the *"network()"* function from mixOmics, which allows us to visualize the correlation between the different datasets. Using a cutoff value of 0.9 we observe a relation between some transcripts and malvidin-3-O-glucoside and tartaric acid.

The final plot was executed using the *"plotLoadings()"* function of mixOmics, and aids in visualizing the loading weights of each selected variable on each component and each data set, see figure 23 that demonstrates the top 20 features for each omics dataset. As shown in the Circos plot, the transcriptomics dataset explains only the preV stage while the most important features of the metabolomics dataset are related to both preV and postV stages. Thus, the metabolomics dataset has a larger weight in component 1.

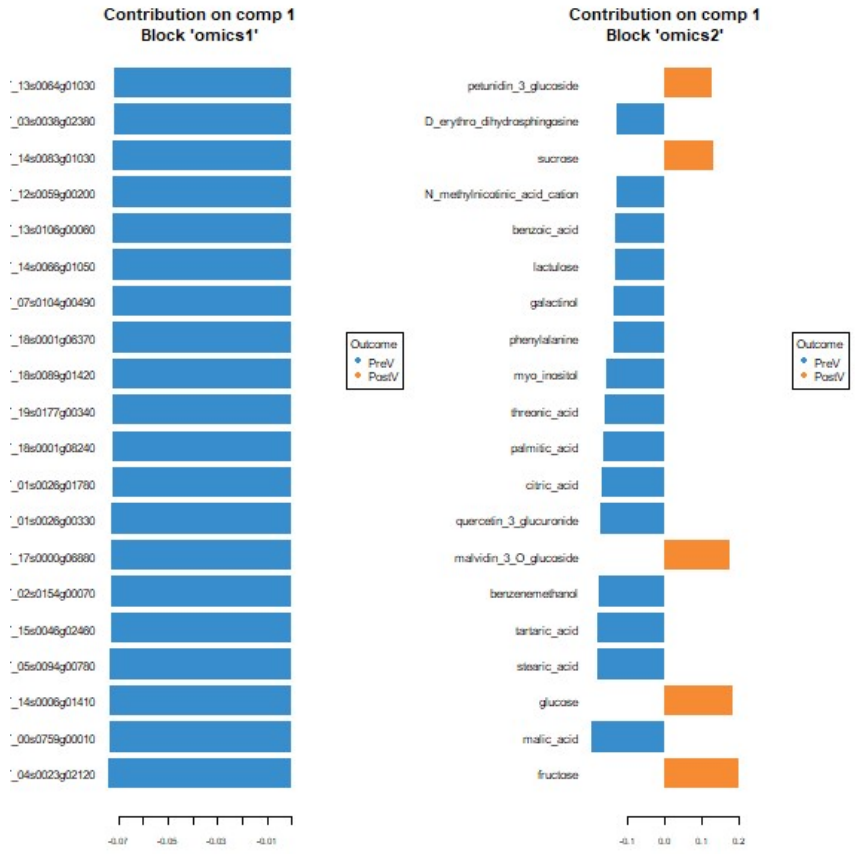


Figure 23: *Multiomics Integration.* Plot resultant of the *"plotLoadings()"* function of mixOmics, that allows the visualization of the loading weights of each selected variables on each component and each data set.

Finally, to evaluate the performance of the DIABLO model we opted for a 10-fold cross-validation repeated 10 times using the *"perf()"* function. This function runs another *"block.splsda()"* with the output of the final model but on cross-validated samples. Then, we obtained the ROC curve for both the transcriptomics and metabolomics blocks and the correspondent AUC value for the model performance. Figure 24 shows the final AUROC curve. As shown in table 18, DIABLO's accuracy, precision and recall were 0.85, 0.66 and 0.875, respectively.

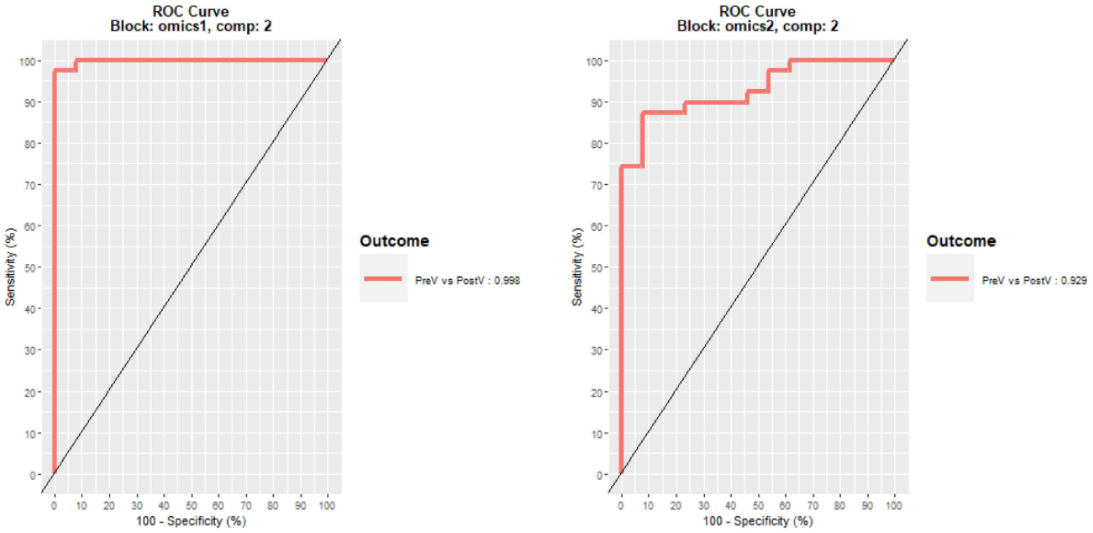


Figure 24: *Multiomics Integration.* DIABLO AUROC curve. AUC value equal to 0.998 for transcriptomics dataset and metabolomics dataset 0.929.

Lastly, we used the *"selectVar()"* function to evaluate the most important features for the DIABLO algorithm and obtained the following 20 most important features of transcriptomics dataset (table 19) and of the metabolomics dataset (table 20).

Table 19: Multiomics Integration. Most relevant features obtained from the DIABLO model for the transcriptomics dataset.

Transcript	UniProtKB	Annotation
VIT_00s0759g00010	D7ST07	Porphobilinogen deaminase chloroplast precursor
VIT_05s0094g00780	D7T2G6	Cysteine synthase
VIT_14s0066g01650	F6HV18	Nodulin MtN21 family
VIT_19s0177g00340	D7T0C7	Unknown protein
VIT_12s0028g02160	F6H4Z7	Ribulose bisphosphate carboxylase
VIT_04s0023g02120	D7SPA6	GTP-binding protein era
VIT_08s0040g03010	D7TQB7	Pigment defective 149
VIT_17s0000g06880	F6GT30	Heparanase protein 2 precursor
VIT_13s0064g01030	D7T2Z6	Zinc finger (C3HC4-type ring finger)BIG BROTHER
VIT_18s0001g06370	E0CRP4	L-ascorbate peroxidase, chloroplast
VIT_15s0046g00290	F6I6F3	Auxin response factor 18
VIT_09s0002g02110	D7TZZ7	Ribonuclease III
VIT_14s0006g01410	D7TSR0	fructokinase-2
VIT_01s0026g00330	D7TNP7	NHL repeat-containing protein
VIT_00s0274g00060	F6GZ22	Glycine-rich protein
VIT_13s0019g05380	D7TM13	Unknown protein
VIT_12s0028g03100	F6H5G9	GPRI1 (GOLDEN2 1)
VIT_06s0004g08450	D7SJF6	Unknown protein
VIT_13s0106g00060	F6HVM2	Ankyrin repeat
VIT_03s0091g00870	F6H656	Adenylylsulfate kinase 1 (AKN1)

Table 20: Multiomics Integration. Most relevant features obtained from the DIABLO model for the metabolomics dataset.

Metabolites	
fructose	threonic acid
malic acid	benzenemethanol
tartaric acid	myo-inositol
glucose	D-erythro-dihydrosphingosine
stearic acid	phenylalanine
citric acid	1,2-anhydro-myo-inositol NIST
malvidin-3-O-glucoside	benzoic acid
palmitic acid	sucrose
quercetin-3-glucuronide	myricetin

- **SMSPL**

For the second concatenation-based approach, we selected the SMSPL algorithm, which can simultaneously predict subtypes and identify potentially significant multiomics signatures. The datasets were divided into train and test datasets (70% train and 30% test) to precisely evaluate the accuracy of the SMSPL. First, we created a list of omics matrices, one for the training datasets and another for the testing datasets, and initialized all the parameters. Then, we proceeded to the initialization of the classifier and started the optimization stage. Lastly, we identified the best validation map and evaluated its performance. In figure 25, we can visualize the ROC curve and AUC value for both the train and test predictions.

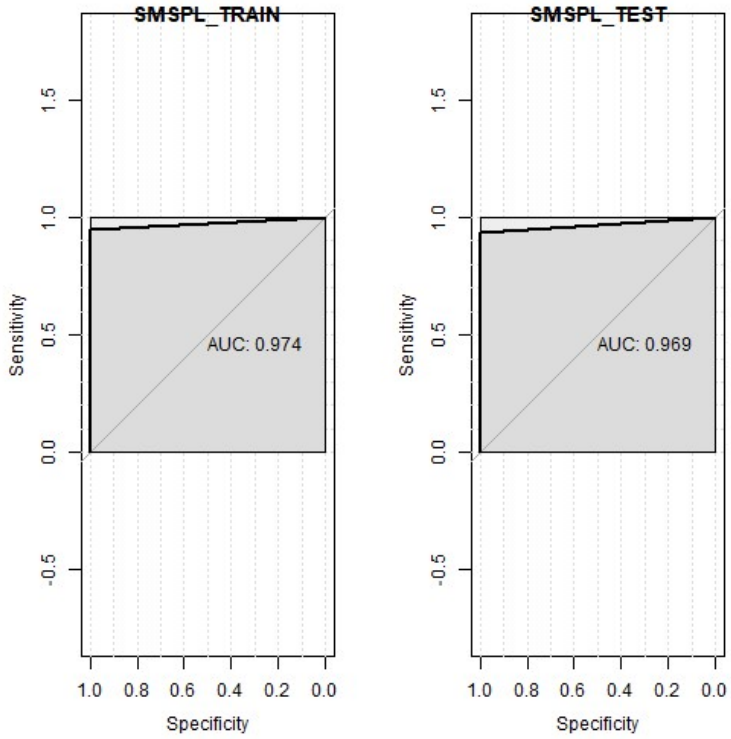


Figure 25: *Multiomics Integration.* SMSPL AUROC curve. AUC value equal to 0.974 for prediction with the train dataset and an AUC value of 0.969 for the performance of the test dataset.

The test dataset prediction demonstrated an accuracy of 0.952, precision of 1.000 and recall of 0.833, hence showing good performance for the SMSPL algorithm in the concatenation-based integration approach. This model can also indicate the most important features. Table 21 and table 22 depict the features for the transcriptomics and metabolomics dataset, respectively.

Table 21: Most Important features regarding the transcriptomics dataset for the SMSPL model. The positive class corresponds to the preV phase, hence positive coefficient values coincide with the preV phase.

Most important features			
Transcriptomics	**UniProtKB**	**Annotation**	**Coef.**
VIT_05s0020g02690	F6HDL7	Copper-binding family protein	0.0041
VIT_09s0002g04080	F6HYD4	IAA9	0.0148
VIT_12s0057g01080	D7ST21	Unknown protein	0.0250
VIT_03s0091g00500	F6H673	Unknown protein	0.0084
VIT_16s0039g02550	F6HEH4	Seed specific protein Bn15D1B	0.0002
VIT_04s0008g07340	F6H3L9	CONSTANS-like protein 4	0.0013
VIT_04s0023g03010	F6GWQ0	fructose-bisphosphate aldolase, chloroplast precursor	0.0001
VIT_13s0106g00060	F6HVM2	Ankyrin repeat	0.4867
VIT_15s0048g00940	D7U7S2	ATP-dependent DNA helicase 2 subunit 2	0.1242

Table 22: Most Important features regarding the metabolomics dataset for the SMSPL model. The positive class corresponds to the preV phase, hence positive coefficient (Coef) values specify PreV features, and negative coefficient values postV features.

Most Important Features	
Metabolites	**Coef**
stearic acid	0.3457
malic acid	0.0423
glucose	-0.0223
fructose	-0.4194
L-aspartic-acid	0.2119
malvidin-3-O-glucoside	-0.1341
N8-acetylspermidine	-0.1239

- **Stack Generalisation**

Next, stack generalisation was performed using the *h2o* library from R, to fit two Generalized Linear Model (glm)s to the training dataset, from the concatenation-based dataset. Then, the two models were stacked and the berry development stage was predicted for the test dataset. The model's performance was, as depicted in table 18, 0.9523 for accuracy and 1.000 for recall and precision. The AUC value was 0.9875. Table 23 shows the 15 most important variables for both models. In this case, since both glm models had the same data and hyperparameters it resulted in the same variable importance results.

Table 23: Most important features of the concatenation dataset for the Stack Generalisation models.

Most important features			
Variable	**UniProtKB**	**Annotation**	**Relative Importance**
VIT_06s0004g07880	D7SJK7	Allergen	0.09238
VIT_15s0046g00290	F6I6F3	Auxin response factor 18	0.07791
VIT_08s0007g07550	F6HLA6	GATA transcription factor 11	0.07595
VIT_12s0057g01080	F6HHQ2	Kelch repeat-containing protein	0.07279
VIT_04s0044g01870	F6I0B3	Auxin efflux carrier	0.06921
VIT_14s0068g01020	D7SVG9	Unknown protein	0.06392
VIT_14s0066g01650	F6HV18	Nodulin MtN21 family	0.06371
VIT_07s0104g01440	D7TP32	Phototropic-responsive NPH3	0.06349
VIT_06s0004g05340	NA	Tropinone reductase	0.06225
VIT_09s0002g00960	D7TZQ4	Inter-alpha-trypsin inhibitor heavy chain	0.06211
VIT_18s0001g15520	E0CQN6	Leaf senescence protein	0.05795
fructose	–	–	0.05769
VIT_09s0002g04260	D7U0H5	Unknown protein	0.05631
VIT_18s0001g10550	E0CP66	LHCA5 (Photosystem I light harvesting complex gene 5)	0.05628
VIT_01s0026g01640	D7TNE5	Band 7 family	0.05624

Transformation-based Integration

- **SNFtool**

For the transformation-based integration, we used the R library *SNFtool* to implement the SNFtool model. First, we calculated the pair-wise distance using the *"dist2()"* function. Then, we constructed the similarity graphs, that have complementary information about the clusters (presented in the supplementary figure S8). Next, the graphs were fused and the overall matrix was computed using the *"SNF()"* function for similarity network fusion. We identified the features with more influence using the *"rankFeaturesByNMI()"* function that returned the list containing the rank based on the normalized mutual information for each feature (table 24). Additionally, we also executed spectral clustering using the function *"spectralClustering()"*, that gave information regarding the final subtype information. The *"displayClusters()"* function can also display the spectral clustering of the fused graphs (shown in supplementary figure S9). Using the output from the spectral clustering in the function *"calNMI()"*, we evaluated the accuracy of the obtained clustering results. In the final step,

we predicted the new labels using label propagation with the library function *"groupPredict()"*. The model's accuracy, precision and recall were 0.933, 0.916 and 1, respectively. The ROC curve is depicted in figure 26, with an AUC value of 0.875.

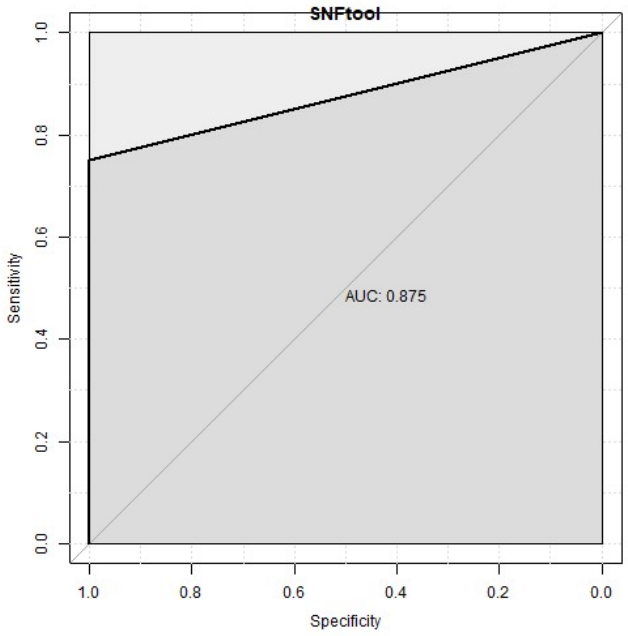


Figure 26: *Multiomics Analysis Integration.* ROC curve for the SNFtool model in the transformation-based integration. The AUC value corresponds to 0.875.

Table 24: Most important features regarding the transformation-based integration for the SNFtool model.

Most important features			
Transcriptomics			**Metabolomics**
Variable	**UniProtKB**	**Annotation**	**Variable**
VIT_14s0083g01030	D7SMN6	MADS-box APETALA 1	fructose
VIT_06s0004g07880	D7SJK7	Allergen	glucose
VIT_14s0083g00850	D7SML9	Lipase GDSL 7	petunidin-3-glucoside
VIT_18s0001g01490	A5ASV7	Oxidoreductase N-terminal domain-containing	malvidin-3-O-glucoside
VIT_07s0104g00490	D7TPB3	Unknown	sucrose
VIT_06s0004g02230	D7SL39	Unknown protein	peonidin-3-glucoside
VIT_18s0041g01350	F6I3U7	Receptor-like protein kinase HAIKU2	delphinidin-3-O-glucoside
VIT_18s0001g15520	E0CQN6	Leaf senescence protein	quercetin-3-glucuronide
VIT_03s0063g02010	D7TPW9	Protease	stearic-acid
VIT_16s0013g00220	D7U738	Metacaspase AtMCP1b	cyanidin-3-glucoside

Model-based Integration

- **Ensemble Classifier with different ML algorithms(Hard and Soft Voting)**

For model-based integration, we used as supervised model an ensemble classifier with different ML algorithms implemented in Python. The model offers 4 different ensemble classifiers, depending on the models selected to perform: the parameter *"option"* enables the user to choose which option to consider (*"opt1"*, *"opt2"*, *"opt3"* or *"opt4"*) or run *"ALL"* the options. Additionally, it enables two different voting approaches (Hard and Soft voting) that can be selected with the *"voting"* parameter, executing *"Hard"*, *"Soft"* or *"ALL"* approaches.

Although four options were tested and analyzed, we presented only the option that gave a better performance, option 1.

- **Option 1**

Option 1 executes an ensemble classifier with 2 SVM models, predicts the outcome of each one and calculates the final prediction using the voting strategy. For this case study, we opted for both voting strategies. The final prediction had an accuracy, precision and recall of 0.93, 1 and 0.92 for soft voting and accuracy, precision and recall of 1, 1, 1 for hard voting. The ROC curve plotted for soft voting is depicted in figure 27. The AUC value is 0.96.

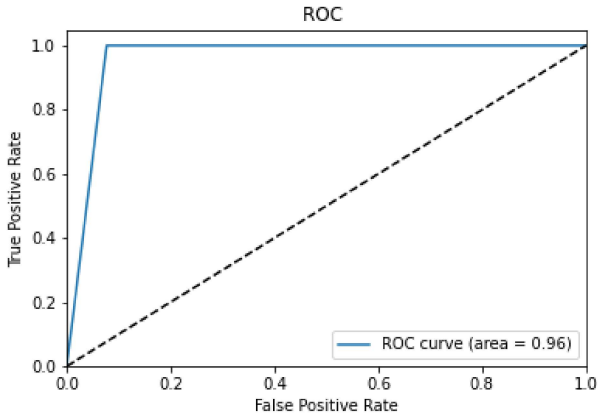


Figure 27: *Multiomics Analysis Integration.* ROC curve for the option 1 of the ensemble classifier model in model-based integration using soft voting. The AUC value corresponds to 0.96.

We also identified the top 4 features with more importance for the transcriptomics and metabolomics dataset, shown in table 25.

Table 25: Most important features according to option 1 of ensemble classifier model in model-based integration.

Most Important Features			
Transcriptomics			**Metabolomics**
Variable	**UniProtKB**	**Annotation**	**Variable**
VIT_06s0004g03240	A5AFS1	Elongation factor 1-alpha 1	malic acid
VIT_08s0007g07680	Q107W9	Aquaporin SIP1;1	leucine
VIT_19s0014g01350	F6H2N3	Ribulose bisphosphate carboxylase, large chain	fructose
VIT_08s0007g00840	D7THJ7	Ribulose bisphosphate carboxylase/oxygenase activase, chloroplast	glucose

Feature Relevance

In conclusion, looking at all the results from the multiomics integration analysis, we identified features relevant for all models, both in transcriptomics and metabolomics datasets. For the transcriptomics dataset, the most frequent transcripts can be identified in table 26 and for the metabolomics dataset, the most relevant features are depicted in table 27.

Table 26: Multiomics Integration Analysis. Feature Relevance. Most common transcripts in all the novel models, their annotation, respective function in berry development and which models have it in common.

Transcripts (UniProtKB)	Annotation	Function	Models	Ref.
F6HV18	Nodulin MtN21 family	Involved in transmembrane transport, dealing with transport of various solutes throughout plant development.	Diablo, Stack Generalisation	[200]
F6I6F3	Auxin response factor 18	Involved in regulation of transcription, DNA-templated and auxin-activated signaling pathway	Diablo, Stack Generalisation	[200]
F6HVM2	Ankyrin repeat	Involved in diverse functions including: transcriptional initiators, cell cycle regulators, cytoskeletal, ion transporters, and signal transducers.	Diablo, SMSPL	[200]
D7SJK7	Allergen	Involved in guard cell differentiation and stomatal complex development. Guard cells are crucial in preV phase for various reasons, like, regulating the accumulation of calcium (Ca) in the berries, hence this gene is up-regulated at this stage.	Stack Generalisation, SNFtool	[207]
E0CQN6	Leaf senescence protein	Enables O-acetyltransferase activity that allows the production of anthocyanin, flavonoid compounds responsible for red/purple colors in the leaves and berries.	Stack Generalisation, SNFtool	[208]

Table 27: Multiomics Integration Analysis. Feature Relevance. Most common metabolites identified in the novel multiomics integration algorithms, the group they represent,and the models they were identified in common with.

Metabolites	Group	Models
frutose	sugar	ALL
malic acid	organic acid	Diablo, SMSPL and Ensemble
glucose	sugar	DIABLO, SMSPL, SNFtool and Ensemble
stearic acid	organic acid	DIABLO, SMSPL and SNFtool
malvidin-3-O-glucoside	anthocyanin	DIABLO, SMSPL and SNFtool
quercetin-3-glucuronide	flavonoid	DIABLO and SNFtool
sucrose	sugar	DIABLO and SNFtool

All features are related to the PreV stage, being in charge of some crucial functions. The common transcripts play similar roles in the PreV stage, such as gene information processing, auxin regulation, signal transduction, allergen, plant development, production of anthocyanins and flavonoids. However, when analysing biological processes we also identify other roles such as, cell wall biogenesis, regulation and structure (D7SML9 (Lipase GDSL 7); F6GZ22 (glycine-rich protein) and F6HB61 (Cellulose synthase CESA3)) and defence responses (F6HPE9 and F6GZM2(Leucine-rich repeat transmembrane)), cell division and elongation (D7SPA6 (GTP-binding protein era)), carbohydrate metabolism, TCA cycle (F6GWQ0 (fructose-bisphosphate aldolase); D7TSR0 (fructokinase-2)), carbon fixation and nucleotide-binding (F6HUC8 (Tubulin beta-6 chain)).

As mentioned before, the PreV stage is characterized by pericarp enlargement, caused by cell division and elongation and overall plant development. Furthermore, it also accumulates organic acids, hence explaining the transcripts involved in carbohydrate metabolism, TCA cycle, carbon fixation and photosynthesis. Organic acids are frequently formed in carbohydrate metabolism, especially in the TCA cycle or carbon fixation reaction that occurs in the dark phase of photosynthesis. Additionally, this stage also accumulates tannins, hydroxycinnamates, and phenolic precursors, so it is reasonable to find transcripts related to the production of anthocyanins and flavonoids. Auxin regulation and signal transduction components are also important since the auxin hormone inhibits ripening. The increase in chlorophyll levels in this phase is also normal. The berries are mainly green and the berries skin is hard, therefore transcripts for cell wall biogenesis, regulation and structure are needed. We also identified nucleotide-binding and defence responses, in some of the transcripts in this stage, just like gene information processing, due to the constant cell division.

On the other hand, in the metabolomics dataset, the most common metabolites are depicted in table 27. In concordance with the information observed in the transcriptomics dataset, we identified that the metabolites are divided into 3 groups. The groups consist of flavonoids, where the anthocyanins belong, organic acids and sugar compounds. The anthocyanins group is composed of malvidin-3-O-glucoside, and although not in common with the models we also identified petunidin-3-glucoside, peonidin-3-glucoside, delphinidin-3-glucoside, quercetin-3-glucoside and cyanidin-3-glucoside. Another flavonoid was identified, corresponding to quercetin-3-glucuronide. The organic acids more frequently observed ,were malic acid and stearic acid, although other organic acids were identified, such as tartaric acid. The most relevant sugars were fructose, glucose and sucrose. Lastly, one other compound group was identified, when looking at all the results. It was the case of aromatic compounds, for instance, benzenemethanol and phenylalanine.

Unsupervised Learning

Finally, besides supervised learning algorithms, we also implemented unsupervised learning for multi-omics integration. We used three models for each type of integration to discover similarities, differences, hidden patterns or data grouping in multiomics unlabeled datasets.

For the different integration based approaches, we chose different multiomics unsupervised models. For the concatenation-based approach we executed MFA, for the transformation-based integration we

performed NEMO, and for the model-based integration we opted for BCC. However, both NEMO and BCC did not offer much information, so they were omitted from the discussion of the results.

- **Concatenation-Based Integration**
  - **MFA**

For the concatenation-based unsupervised learning, we implemented the MFA algorithm in R, using the library *FactoMineR*. First, we selected the concatenated dataset and formed a vector with the number of variables present in each group. We also selected the type of variables that were present in each group. Since we divided the concatenated dataset into 3 groups, transcriptomics dataset, metabolomics dataset and output, our parameter *"type"* corresponded to ["c","c","n"].

Then, we extracted the proportion of variances explained by the different dimensions and obtained the scree plot, which is illustrated in figure S12. As shown in the scree plot, the first dimension explains the most variance of our dataset.

The MFA model also offers several different graphs of variables, figure S13 shows the plotted groups of variables. The same conclusions can be observed in figure S14, which demonstrates the contribution of the groups to dimension 1 and dimension 2.

We can use the *"fviz_contrib()"* function (figure 28) to visualize in more detail the contribution of the quantitative variables (in %) to the definition of the dimensions. The transcriptomics dataset contributes more to explain dimension 1 and the metabolomics dataset is better to explain dimension 2.

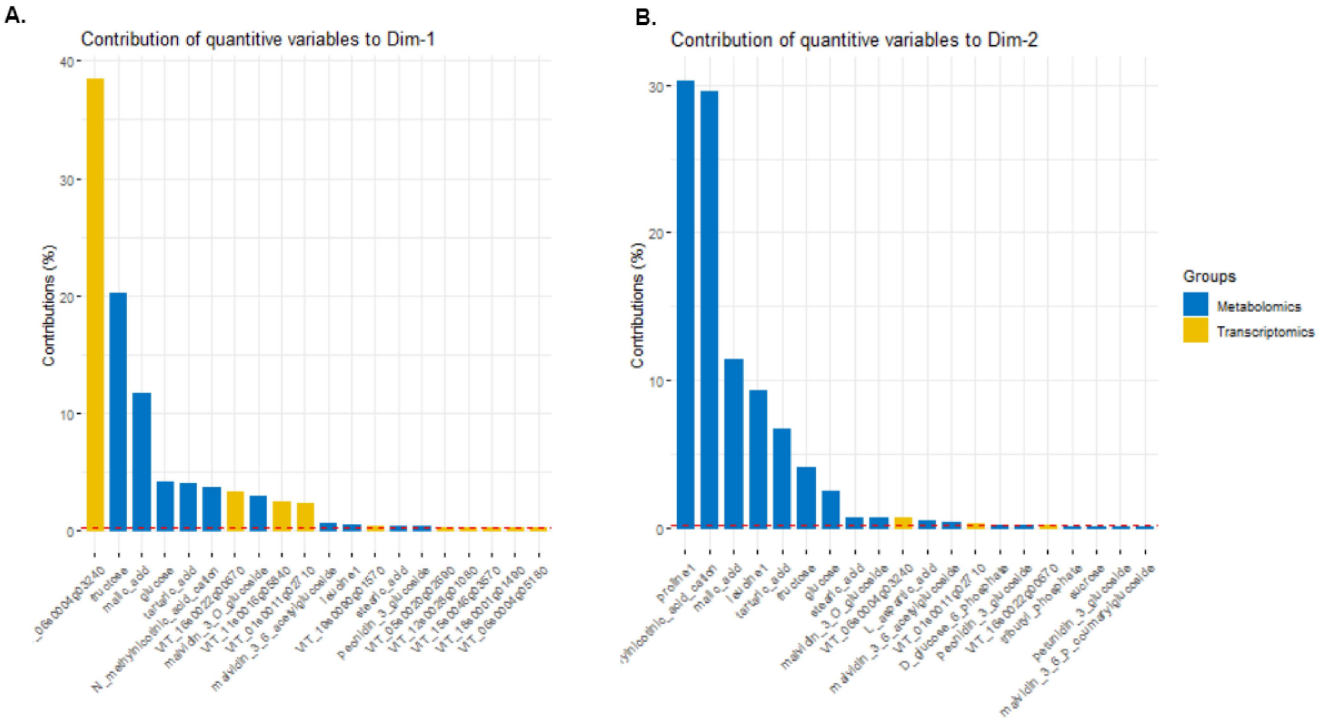


Figure 28: *Multiomics Analysis Integration. Unsupervised Learning.* Graph of variables, that shows the contribution of quantitative variables relative to the (A) dimension 1 and (B) dimension 2.

Furthermore, as we can see in the figure 28, the variables with bigger values, contribute the most to the definition of the dimensions, hence are the most important in explaining the variability in the data set. The transcriptomics and metabolomics datasets contribute both to the first dimension, the metabolomics dataset explains dimension 2 with greater detail. The features that explain most of the variability in the dataset, for dimensions 1 and 2 are illustrated in table 28.

Table 28: Features that explain the most variability in the dataset according to the MFA unsupervised model.

Dimension 1	UniProtKB	Annotation	Dimension 2
VIT06s0004g03240	A5AFS1	Elongation factor 1-alpha 1	proline
fructose	–	–	N-methylnicotinic acid cation
malic acid	–	–	malic acid
glucose	–	–	leucine
tartaric acid	–	–	tartaric acid
N-methylnicotinic acid cation	–	–	fructose
VIT16s0022g00670	F6HAU0	Vacuolar invertase 1, GIN1	glucose
malividin-3-O-glucoside	–	–	-
VIT11s0016g05840	F6HHB3	Protease inhibitor/seed storage/lipid transfer protein (LTP)	-
VIT01s0011g02710	A5AEV3	No hit	-

Lastly, Figure 29 represents the individuals, colored by their outcome variable, Berry development. Individuals with similar profiles are closer to each other on the factor map.

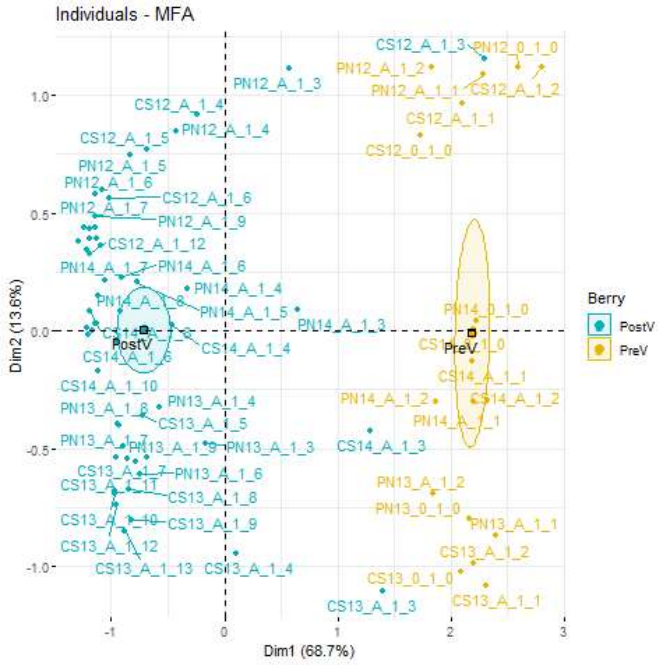


Figure 29: *Multiomics Analysis Integration. Unsupervised Learning.* Graph of variables, that shows the contribution of quantitative variables relative to the (A) dimension 1 and (B) dimension 2.

Figure 29 exhibits a clear distinction between the two levels of the outcome variable PreV and PostV, explained by dimension 1 (68.7%), which indicates that the berry development is explained by dimension 1.

Regarding dimension 1, in the resultant features that explain most of the variability, and thus, explain berry development, we found some transcripts and metabolites obtained in previous analysis. It was the case of:

- *A5AFS1 (Elongation factor 1-alpha 1)*, that enables GTP binding, translation elongation factor activity, heterocyclic compound binding and organic cyclic compound binding;
- *F6HAU0 (vacuolar invertase 1 (GIN1)*, involved in the carbohydrate metabolic process, which is responsible for the organic acid formation and other compounds important in the PreV phase; [200]
- *F6HHB3, a protease inhibitor/seed storage/lipid transfer protein (LTP)*, involved in plant growth and development, is important for fruit ripening and cell wall biogenesis, cellulose biosynthetic process and cell wall organization, key roles in berry development [200]

However, it was also identified a new transcript A5AEV3, with unknown annotation, being a good case study for future research. On the other hand, for the metabolomics dataset, we identified fructose,

malic acid, glucose and tartaric acid, key metabolites identified in the previous analysis. All these features were mentioned before and thus, are in agreement with the supervised multiomics integration analysis.

5.2 CASE STUDY II: *Arabidopsis thaliana*

5.2.1 *Pre-processing*

The second case study focused on two datasets of transcriptomics and fluxomics data that, as shown in table 5, had the same 26 samples. We successfully inputted the datasets in our pipeline. Then, proceeded with the preprocessing of the data as follows.

Missing Values

In the first step of preprocessing, we identified that none of the datasets contained missing values, hence none of the rows was removed. However, in order to further decrease the number of features in transcriptomics and fluxomics datasets, we continued with the next step, feature selection.

Feature Selection

Three filter methods were applied in both datasets. Regarding the transcriptomics dataset, the first filter, that assured that at least one transcript per cell existed, did not remove any features. The second filter, the median filter, that filtered genes expressed in at least two samples removed 14706 rows, resulting in a total of 17795 features. The last filter, the flat patterns filter, filtered genes whose maximum ratio value over the minimum value of expression was greater than 2 and deleted 17070 features, ending with a transcriptomics dataset of 725 features.

On the other hand, although the same filters were executed for the fluxomics dataset, it did not remove any features. However, since the fluxomics datasets may contain fluxes with zero values in all samples, we searched for rows with only zero values and successfully removed 1195 rows, ending with a total of 407 reactions.

Normalisation and Scaling

Furthermore, as the last step in the preprocessing stage, we executed the normalisation and scaling of the datasets. As the transcriptomics dataset was previous scaled by the authors, we scaled the fluxomics dataset with the *"scale()"* function in R, to obtain more comprehensible plots and improve the models' accuracy.

5.2.2 *Exploratory Analysis*

After completing the preprocessing stage, we advanced to a brief exploratory analysis to have a clear idea of how the data is divided and if it was in agreement with the original article.

Barplots

As the metadata contained the treatment type of the samples, we executed a barplot to see how many samples were in the control and drought groups. Figure 30 depicts a barplot of the treatment variable. Both treatments have the same number of control and drought samples, the thirteen samples as described in the article, confirming the dataset balance.

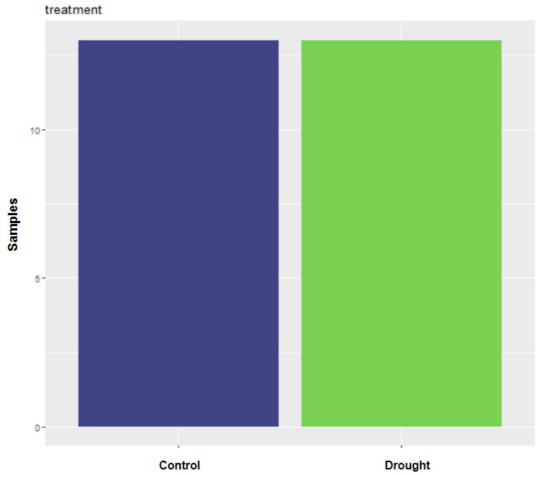


Figure 30: *Exploratory Analysis.* Barplot of the treatment variable for the second case study. Same number of samples for control and drought treatments (13).

Heatmaps

Additionally, we executed two heatmaps' analyses for each dataset to identify interesting similarities and discrepancies in our data. Heatmaps are an useful way to show relationships between two variables plotted on each axis. Figure 31 shows the heatmap for the transcriptomics dataset (A) and the fluxomics dataset (B). Figure 31A illustrates that the transcriptomics dataset allows to divide the samples according to their treatment type (control or drought) more easily when compared to the analysis of the second heatmap figure 31B, from the fluxomics dataset.

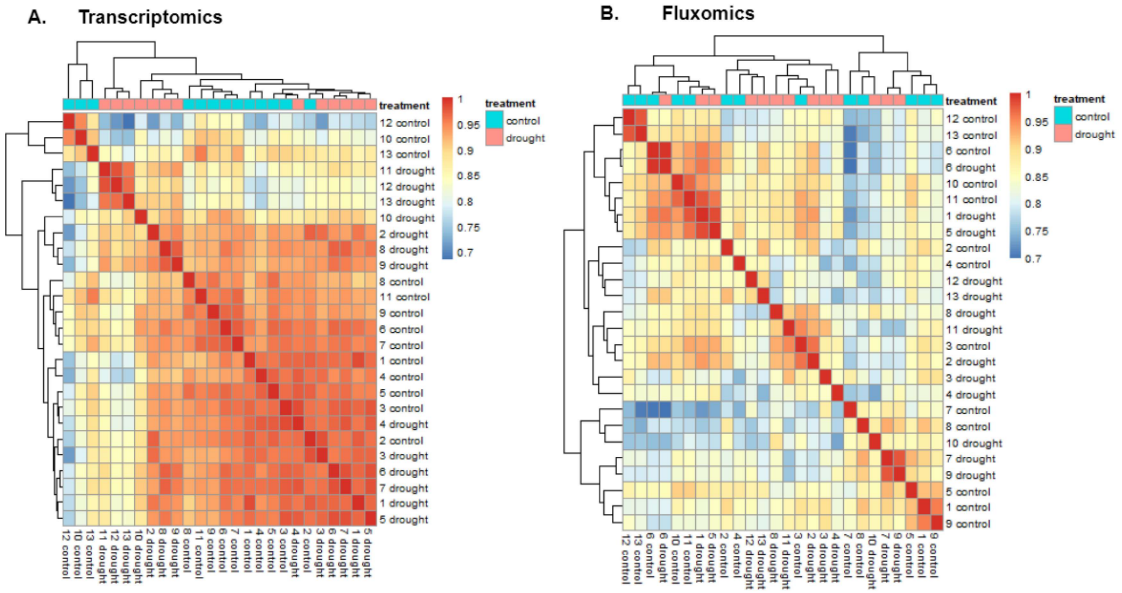


Figure 31: *Exploratory Analysis.* Heatmap analysis further inspecting the treatment variable for the second case study. (A) transcriptomics dataset. (B) Fluxomics dataset.

PCA

Regarding the PCA analysis, Figure 32 depicts a PCA in relation to the treatment variable. As shown in the heatmaps, using the PCA analysis, we can more clearly understand that the PC1(31.076%) and PC2 (29.464%) from the transcriptomics dataset (A) can explain transcript expression regarding the control and drought samples. On the other hand, the fluxomics dataset (B) cannot divide the samples according to their treatment.

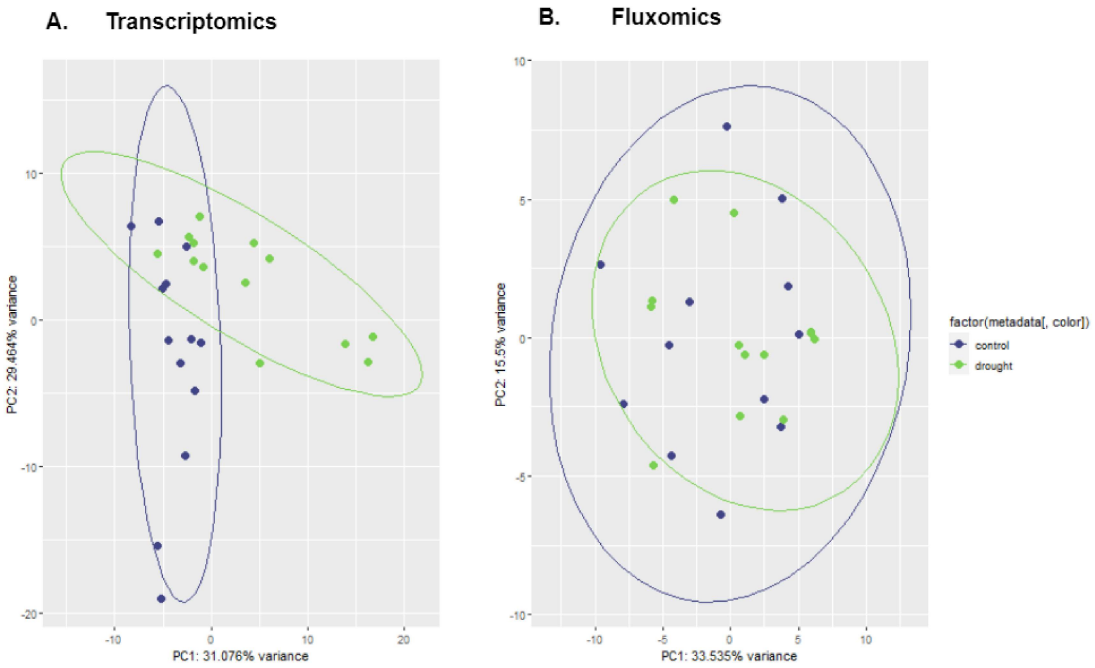


Figure 32: *Exploratory Analysis.* PCA analysis to inspect the treatment variable for the second case study. (A) transcriptomics dataset. (B) Fluxomics dataset.

Differential Expression

At last, for the final step in the exploratory analysis, we performed a differential expression analysis to the transcriptomics dataset, thus contributing to improving our knowledge of *Arabidopsis thaliana* metabolism when defending itself from the environmental stresses of drought. Additionally, this analysis allowed us to filter the transcriptomics dataset to match the number of features of fluxomics to facilitate the analysis in the multiomics integration analysis, ending with a total of 407 features in transcriptomics and fluxomics. In table 29, we have the top 10 significantly expressed transcripts regarding the contrast between the control and drought treatments, and their respective annotation obtained from the supplementary files of Sclep et al. article.

Table 29: *Case Study II:* Top 10 up-regulated and down-regulated differential expressed genes from transcriptomics dataset, explaining the drought condition.

Up-Regulated			Down-Regulated		
CATMA ID	**TAIR7**	**Annotation**	**CATMA ID**	**TAIR7**	**Annotation**
CATMA4a19840	At4g18700	Encodes CBL-interacting protein kinase 12 (CIPK12)	CATMA5a53370	At5g57660	zinc finger (B-box type) family protein
CATMA1a22270	At1g23200	pectinesterase family protein	CATMA1c71101	At1g08360	60S ribosomal protein L10A (RPL10aA)
CATMA1c72010	At1g66390	PAP2 (Production of Anthocyanin Pigment 2)	CATMA4a03190	At4g02840	small nuclear ribonucleoprotein D1 putative
CATMA4a30730	At4g29070	unknown protein	CATMA1a18630	At1g19610	Low-molecular-weight cysteine-rich 78
CATMA3c57894	At3g60910	Methyltransferase superfamily protein	CATMA3a18630	At3g18980	F-box family protein
CATMA1a28510	At1g30500	CCAAT-binding transcription factor (CBF-B/NF-YA) family protein	CATMA2a42166	At2g43760	Molybdopterin biosynthesis MoaE family protein
CATMA4a27270	At4g25580	stress-responsive protein-related	CATMA5c64749	At5g44340	TUB4 (tubulin beta-4 chain)
CATMA1a69150	At1g79970	unknown protein	CATMA5c65084	At5g62300	40S ribosomal protein S20 (RPS20C)
CATMA5a00120	At5g01090	legume lectin family protein	CATMA1c72183	At1g76680	OPR1 (12-oxophytodienoate reductase 1)
CATMA4a03330	At4g03000	protein binding / zinc ion binding	CATMA2a19460	At2g20850	SRF1; kinase,similar to leucine-rich repeat transmembrane protein kinase

Therefore, we identified interesting up-regulated transcripts involved in drought stress. The first is At4g18700 that encodes a CBL-interacting protein kinase 12 (CIPK12). This protein encodes a member of the family of SNRK3 kinases that plays an essential role in the ABA regulatory pathway [210]. The Absicic Acid (ABA) regulatory pathway is a well known regulatory circuit, also mentioned in [179], that when under a water deficit, increases the ABA levels that triggers the expression of several drought-stress related genes.

Next, we identified At1g23200, a pectinesterase protein involved in cell wall modification and organization, that has been identified in other articles, which concluded that drought stress limits cell growth and alters the cell wall [211, 212].

Additionally, we identified At1g66390 encoding the protein PAP2 (Production of Anthocyanin Pigment 2), extremely important in drought response, because the many drought-stress associated genes, triggered by ABA, result in the accumulation of protective proteins that increase the concentration of certain compatible solutes, like sugars and proline, and antioxidants, such as flavonoids and polyphenols, consequently suppressing energy-consuming pathways [179]. Anthocyanins are a main class of flavonoids that function as scavengers of Reactive Oxigen Species (ROS), therefore contributing to abiotic stress tolerance [213].

At4g29070 encoding an unknown protein was also identified as relevant as it seems to be involved in arachidonic acid secretion, according to UniProtKB. This acid is mentioned in Pan et al. article for the Italian ryegrass (*Lolium multiflorum*) species, which defended that the arachidonic acid reduced the oxidative damage in the drought-tolerant species.

At3g60910, a methyltransferase, improves tolerance to dehydration stress treatment [215]. At1g30500, encoding a transcription factor for CCAAT-binding (CBF-B/NF-YA), is described in [212] to maintain the reduced growth of plants under drought, which is an acclimation response of plants to survive prolonged drought stress.

Furthermore, we also identified At4g25580, a stress-responsive related protein, that plays a crucial role in abscisic acid response. At1g79970 is an unknown protein but is similar to a senescence-associated protein that plays an important role in regulating ABA signalling and drought tolerance through interaction with open stomata 1 (OST1) [216].

Lastly, we identified At5g01090, a legume lectin family protein, that plays a role in stress-related responses [217], and At4g03000, a zinc ion binding protein. Zinc ion binding proteins can be ZPT2-related proteins that have zinc-finger motifs in their molecules and work as transcriptional repressors, which increase stress tolerance following growth retardation under drought stress [218].

5.2.3 *Classic Machine Learning Models*

As for Case Study I, classical ML models were used for both the individual omics analysis and multiomics integration analysis, namely SVM, RF and ANN models.

For both analysis, individual and multiomics integration, we split our data into train and test datasets. For the individual omics analysis, table 30 depicts the train and test dimensions for both datasets.

Table 30: *Individual Omics Analysis*. Dimensions (samples, features) of the original transcriptomics and metabolomics datasets and their respective train and test datasets.

	Transcriptomics	**Fluxomics**
Original	(26,408)	(26,408)
Train	(18,408)	(14,408)
Test	(8,408)	(12,408)

For the different ML models, we executed 10-fold and 3 repetitions CV in R for the individual omics analysis and implemented 3-fold CV in Python for the multiomics integration analysis. Hence, by executing these classic ML models, we can verify which models better suit our data and see how well can the individual datasets predict the treatment outcome, find the most important features for the prediction, compared to the multiomics integration analysis. Table S1 shows the values of the different error metrics (PECC, Precision and Recall) obtained by the three models for both datasets, in the individual omics analysis.

Support Vector Machine

- *Individual Omics Analysis*

The SVM model has an accuracy, precision and recall of 0.75, 0.8 and 0.8 for the transcriptomics dataset and accuracy, precision and recall of 0.33, 1 and 0.33 for the fluxomics dataset. Figure 33 shows the ROC curve plot for the SVM model and respective value for each dataset. The transcriptomics dataset had an AUC of 0.96 while the fluxomics dataset had an AUC value of 0.11, which indicates that the fluxomics dataset was not a very good predictor of the outcome variable as the AUC value measures the ability that a classifier has to distinguish between the positive and negative classes.

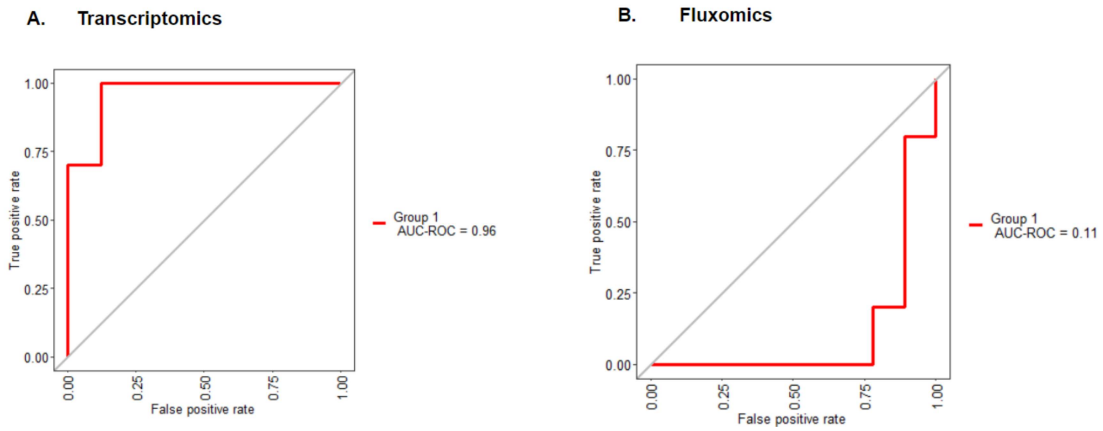


Figure 33: *Individual Omics Analysis*. ROC curve of the SVM model for the of (A) transcriptomics dataset, with an AUC value of 0.96 and (B) Fluxomics dataset, with an AUC value of 0.11.

- *Multiomics Integration Analysis*

For the SVM model, we executed grid search and random search to obtain the best estimator for our data. The best estimator hyperparameters were: C=1, gamma=0.01, kernel="linear" obtained with grid search and random search. The model's accuracy, precision and recall were 0.33 for all. Furthermore, we calculated the ROC curve, depicted in figure 34 and the top 10 most relevant features (table 31). The AUC value was 0.33.

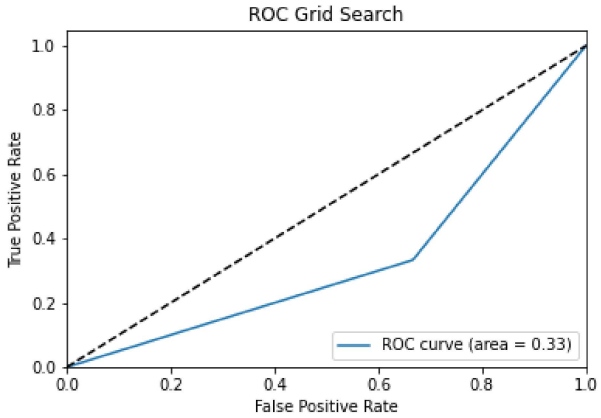


Figure 34: *Multiomics Integration Analysis.* ROC curve for the SVM model in the concatenation-based integration. Grid Search ROC curve with an AUC value of 0.33.

Table 31: Most important features regarding the concatenation dataset for the SVM model, in Case Study II.

Most Relevant features			
Variable	**TAIR7 mapping**	**Annotation**	**Score**
CATMA1c72010	At1g66390	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2)	0.122518
CATMA4c42685	At4g37980	ELI3-1 (ELICITOR-ACTIVATED GENE 3)	0.0963773
CATMA4A19840	At4g18700	Encodes CBL-interacting protein kinase 12 (CIPK12)	0.089341
CATMA3b43840	At3g50840	phototropic-responsive NPH3 family protein	0.088634
TCP7	–	Triose phosphate translocator (G3P)	0.0842384
CATMA2A35140	At2g36870	xyloglucan:xyloglucosyl transferase	0.0826423
CATMA5a00120	At5g01090	legume lectin family protein	0.0806931
CATMA5A22560	At5g24870	zinc finger (C3HC4-type RING finger) family protein	0.0777628
CATMA1A54860	At1g65560	allyl alcohol dehydrogenase	0.0744085
CATMA2A33070	At2g34960	CAT5 (CATIONIC AMINO ACID TRANSPORTER 5)	0.0688854

Random Forest

- *Individual Omics Analysis*

For the RF, the transcriptomics dataset obtained an accuracy, precision and recall of 0.5, 0.667 and 0.4, respectively, and for the fluxomics dataset, it obtained an accuracy, precision and recall of 0.5, 0.33 and 0.5, respectively. The ROC curve for both datasets is displayed in figure 35. As the SVM model, the RF model had much better results using transcriptomics (AUC value of 0.95) than fluxomics (AUC value of 0.62), indicating that these models can correctly distinguish control and drought conditions when trained with transcriptomics data alone.

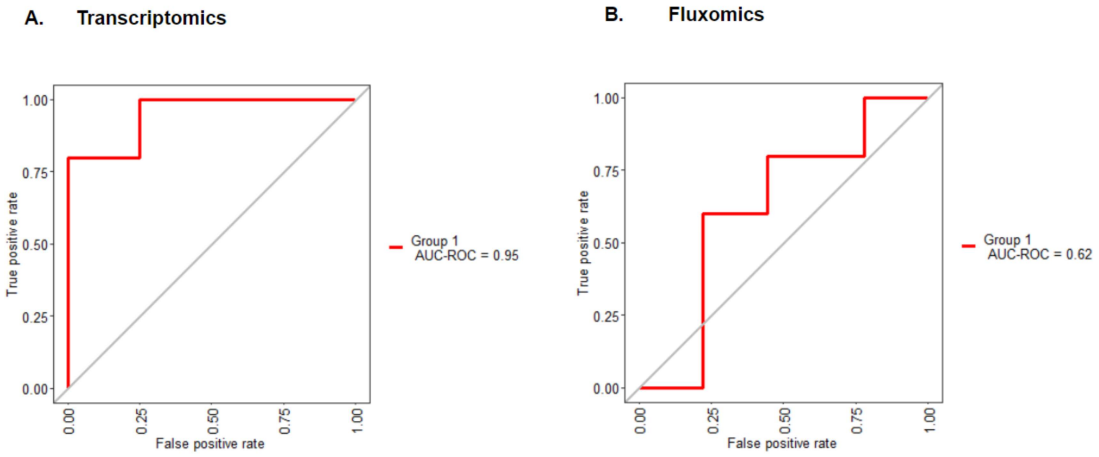


Figure 35: *Individual Omics Analysis*. ROC curve of the RF model for the of (A) transcriptomics dataset, with an AUC value of 0.95 and (B) Fluxomics dataset, with an AUC value of 0.62.

- *Multiomics Integration Analysis*

For concatenation-based integration, we also performed a RF model. By executing the grid search and random search, we obtained the best estimator with the following hyperparameters: max_depth=2, max_features=6, min_samples_leaf=2. The model's accuracy, precision and recall were 0.33, 0.33 and 0.33, respectively. Figure 36 shows the ROC curve for the RF model. Additionally, we also identified the most relevant features, as depicted in table 32.

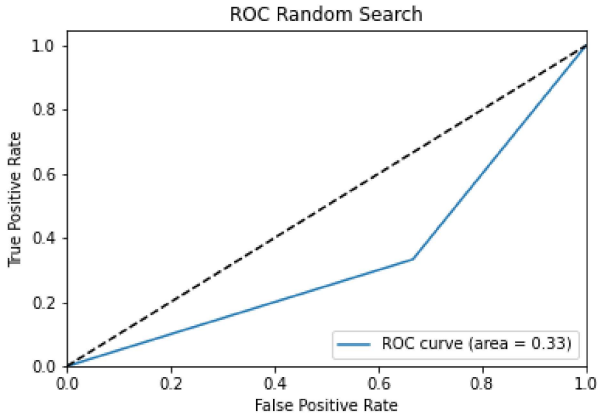


Figure 36: *Multiomics Integration Analysis.* ROC curve of the RF model in the concatenation-based integration. Obtained an AUC value of 0.33.

Table 32: Most important features regarding the concatenation dataset for the RF model, in Case Study II.

Most Relevant features			
Variable	**TAIR7 mapping**	**Annotation**	**Score**
CATMA5A56040	At5g60280	lectin protein kinase family protein	0.0255075
CATMA5A47470	At5g51545	unknown protein	0.0179174
CATMA1A22270	At1g23200	pectinesterase family protein	0.0178999
CATMA5c64749	At5g44340	TUB4 (tubulin beta-4 chain),beta tubulin gene	0.0162646
CATMA3c57192	At3g13610	oxidoreductase, 2OG-Fe(II) oxygenase family protein	0.0161255
CATMA5A53370	At5g57660	zinc finger (B-box type) family protein	0.0154991
CATMA5A47980	At5g52030	TraB protein-related	0.0143004
CATMA5a00130	At5g01100	unknown protein	0.0135263
CATMA4c42587	At4g31985	60S ribosomal protein L39 (RPL39C)	0.0126991
CATMA5a44470	At5g48490	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein	0.0124231

Artificial Neural Network

- *Individual Omics Analyses*

For the last individual omics analysis model, we opted for the ANN classifier. The accuracy, precision and recall values for the transcriptomics dataset were 0.5, 0.667 and 0.4, respectively, while for the fluxomics dataset were 0.416, 0.285 and 0.5, respectively. Figure 37 depicts the ROC curve for the transcriptomics and fluxomics datasets. The AUC value for the transcriptomics dataset was 0.92, and for the fluxomics dataset, was 0.6.

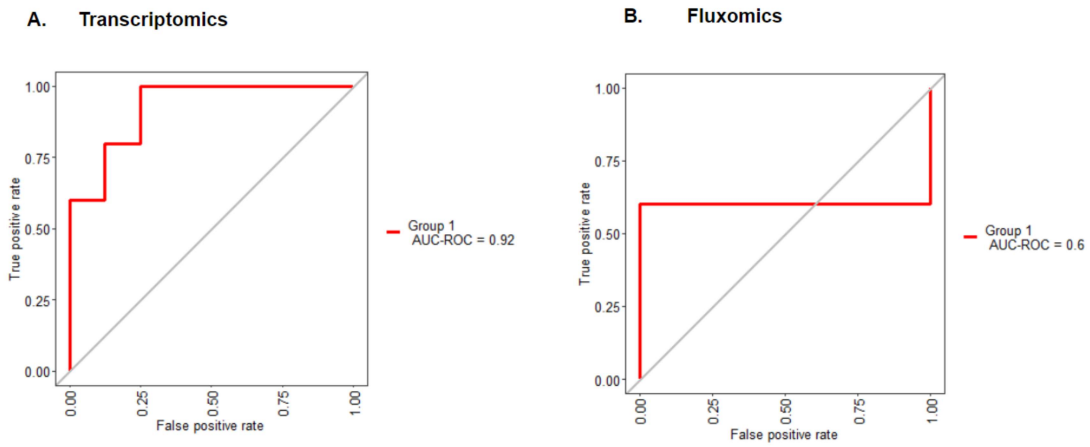


Figure 37: *Individual Omics Analysis.* ROC curve of the ANN model for the of (A) transcriptomics dataset, with an AUC value of 0.92 and (B) Fluxomics dataset, with an AUC value of 0.6.

- *Multiomics Integration Analysis*

Lastly, for the multiomics integration analysis, we implemented an ANN model. Using the random search only, since the grid search took to long to run we obtained the best estimator with the following hyperparameters: activation='logistic', alpha=0.01, hidden_layer_sizes=(25,), max_iter=900, solver='lbfgs'. The model's accuracy, precision and recall were 0.5, 0.6 and 0.75. Figure 38 displays the ROC curve, with an AUC value of 0.375. Table 33 depicts the most relevant features for the ANN model obtained using the Python package *LIME* to calculate feature relevance. However, it can only say the important features for each sample at each time. Nevertheless, it gives us the features that had more importance to make the decision.

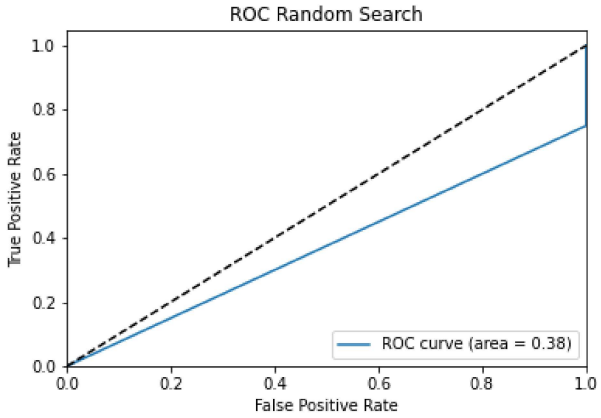


Figure 38: *Multiomics Integration Analysis.* ROC curve of the ANN model for the concatenation-based integration in Case Study II, with an AUC value of 0.375.

Table 33: Most important features regarding the concatenation dataset for the ANN model, in Case Study II.

Most Relevant features		
Variable	**TAIR7 mapping**	**Annotation**
CATMA1a24995	At1g26770	ATEXPA10 (ARABIDOPSIS THALIANA EXPANSIN A10), encodes an expansin
R01900_x	–	Glyoxylate and dicarboxylate metabolism; Reductive carboxylate cycle (CO2 fixation);Citrate cycle (TCA cycle)
TCX14	–	Citrate Transporter
R01325_x	–	Glyoxylate and dicarboxylate metabolism; Reductive carboxylate cycle (CO2 fixation);Citrate cycle (TCA cycle)

5.2.4 *Feature Relevance*

- *Individual Omics Analysis*

Besides the error metrics evaluation, we also obtained the most relevant features, used by the models in the transcriptomics dataset to predict the treatment variable, for each classifier (SVM, RF, ANN - supplementary tables S2, table S4 and table S6 respectively).

Looking at the results from the three classifiers for the transcriptomics dataset, we identified features in common with at least two models. The results are depicted in table 34, which shows the transcripts identified, the respective annotation and function they have in drought conditions, and also the models where they were identified.

Table 34: Individual Omics Analysis. Feature Relevance. Most common transcripts in the three classical ML models, their annotation, respective function in drought conditions and which models have in common.

Transcripts (TAIR mapping)	Annotation	Function	Models	Ref.
At1g66390	PAP2 (Productionof Anthocyanin Pigment 2)	Drought-stress associated genes, triggered by ABA, result in the accumulation of protective proteins that increase the concentration of certain compatible solutes, like sugars and proline,and antioxidants, such as flavonoids and polyphenoids, consequently suppressing energy-consuming pathways.	All	[179]
At1g19610	Low-molecular-weight cysteine-rich 7	Involved in Defense response of fungus, identified in control conditions	All	[200]
At1g23200	pectinesterase family protein	Involved in cell wall modification and organization, which may indicate that drought stress limits cell growth and alters the cell wall.	SVM and RF	[211, 212]
At3g60910	Generic methyltransferase	Improves tolerance to dehydration stress treatment.	SVM and RF	[215]
At1g30500	CCAAT-binding transcription factor (CBF-B/NF-YA)	It is thought to maintain the reduced growth of plants under drought, which is an acclimation response of plants to survive prolonged drought stress.	RF and ANN	[212]
At1g76680	OPR1 (12-oxophytodienoate reductase 1)	Involved in jasmonic acid biosynthesis and therefore enhances the antioxidant system. Up-regulated in the differential expression in the control samples.	SVM and RF	[219, 220]
At3g18980	F-box family protein	Essential for triggering ethylene responses in plants. The ethylene hormone modulates plant growth and development as well as plant responses to abiotic stress.	SVM and RF	[221, 222]
At4g37980	(ELI3-1: Elicitor-activated gene 3)	Oxidoreductase/ zinc ion binding that encodes a putative mannitol dehydrogenase. It is involved in lignin biosynthesis, catalyzing the final step of the production of lignin monomers, which is a important component in the plant cell wall and is associated with drought tolerance.	RF and ANN	[223]

Table 35: Individual Omics Analysis. Feature Relevance. Most common reactions in the three classical ML models, the subsystem in which they occur, and which models have it in common.

Reaction	Subsystem	Models
TCP7	Triose phosphate translocator (G3P)	ALL
R01015_p	Fructose and mannose metabolism; Glycolysis / Gluconeogenesis; Inositol metabolism;Carbon fixation (control)	SVM and RF
R01015_c	Fructose and mannose metabolism; Glycolysis / Gluconeogenesis; Inositol metabolism;Carbon fixation (drought)	
TCP8	Triose phosphate translocator (glyceroneP) (drought)	SVM and ANN
R00127_c	Purine metabolism (control)	SVM and ANN
R03321_c	Glycolysis / Gluconeogenesis (drought)	
R02739_c	Pentose phosphate pathway; Glycolysis / Gluconeogenesis (drought)	
TCX16	Isocitrate transporter (control)	
R01324_c	Citrate cycle (TCA cycle) (control)	
R01325_x	Glyoxylate and dicarboxylate metabolism; Reductive carboxylate cycle (CO2 fixation);Citrate cycle (TCA cycle) (drought)	

Looking closely at variables identified in the transcriptomics dataset, most of the functions are related with the reduced cell and plant growth induced in the drought conditions by the modification of cell wall and also drought-stress associated genes triggered in this conditions.

Regarding the fluxomics dataset (table 35), most reactions are related to the glycolysis/ gluconeogenesis pathway, the carbon fixation and the TCA cycle. Although these pathways are active in both control and drought samples, glycolysis/ gluconeogenesis is active in the drought treatment as drought conditions promote an increase in the levels of sugars, like fructose and mannose [224]. The carbon fixation pathway is hypothesized to perform major functions in drought resistance as mentioned in the Wei et al. article. Furthermore, the TCA cycle is also functional in water-stress responses since these responses increase the content of TCA cycle intermediates and total amino acid levels [224].

Other reactions, like the TCP8, and the TCP7, responsible for the triose phosphate translocator pathway were also identified; however, not much is known about the relation of this reaction to the drought stress response in *Arabidopsis thaliana*, but they were also considered relevant in the original article [179].

- *Multiomics Integration Analysis*

For the multiomics integration analysis, when looking at all the features obtained using the three classical ML models, we do not see any features in common with the models, neither in the transcriptomics or the fluxomics datasets. The poor results in the different models may be caused by the small number of samples in this case study. Nevertheless, the use of novel multiomics integration algorithms may give better results.

5.2.5 *Novel models for Multiomics Integration Analysis*

Next, we performed multiomics integration analysis, in which we studied three different integration-based approaches: concatenation-based, transformation-based and model-based integration. The multiomics integration stage was the most important since it could give us better and more holistic results than individual omics analysis. We evaluated which models could give better accuracy and more useful information. Therefore, for the same reasons indicated in **Case Study I**, we selected for this analysis only the models that obtained better results and provide more useful information. The selected models were: DIABLO, SMSPL for concatenation-based integration, SNFtool for transformation-based integration and option 1 with soft voting for the model-based integration. Table 36 shows the different algorithms for each integration and the corresponding accuracy, precision and recall values for the **Case Study II**.

Table 36: Results of the several metrics used to evaluate the performance of the different models in the **Case Study II**. PECC (accuracy), Precision and Recall for the classification algorithms, and the AUC values.

Model	Metrics			
	Classification Metrics			
	PECC	**Precision**	**Recall**	**AUC value**
DIABLO	0.83	1	1	0.833
SMSPL	0.6667	1	0.6	0.667
SNFtool	0.6667	0.6	1	0.667
Ensemble Classifier:				
Option 1 (soft Voting)	0.833	0.8	1	0.75

Concatenation-based Integration

For the concatenation-based integration, we opted for two models: DIABLO and SMSPL.

- **DIABLO**

The steps are identical to the steps executed in the **Case Study I**. First, we set our data as a list of data matrices matching the same samples in the rows. The omics datasets were named as blocks, **block omics1** corresponding to the transcriptomics dataset, and **block omics2** for the fluxomics dataset.

For the matrix design, a symmetrical matrix that indicates the correlation between the two omics as a value ranging from 0 to 1 was built, and a partial least square regression and a Sparse Partial Least Squares regression were executed to find the best value. We opted for the value of 0.78.

Regarding the tuning of the number components, we first fitted a DIABLO model without variable selection to assess the global performance and then we selected the best component numbers based on the plot from figure 39, which shows the performance of the executed DIABLO model.

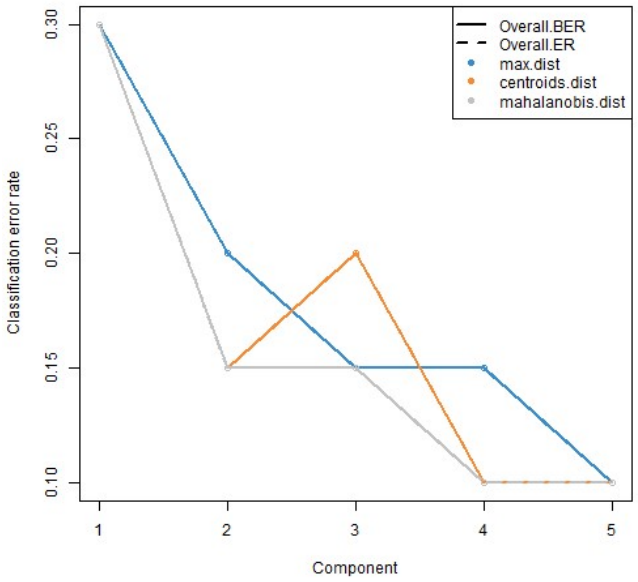


Figure 39: *Multiomics Integration Analysis.* DIABLO model performance for number components tuning.

Looking at the figure, we extracted the number of components regarding the weighted vote of the overall balanced error rate (BER) and the distance of the centroid (centroids.dist), obtaining a total of five (5) components for the final model.

Then, we executed the final model, with the selected five components, and acquired the following sample plot that projects the sample into the space extended by the components of each block, figure 40.

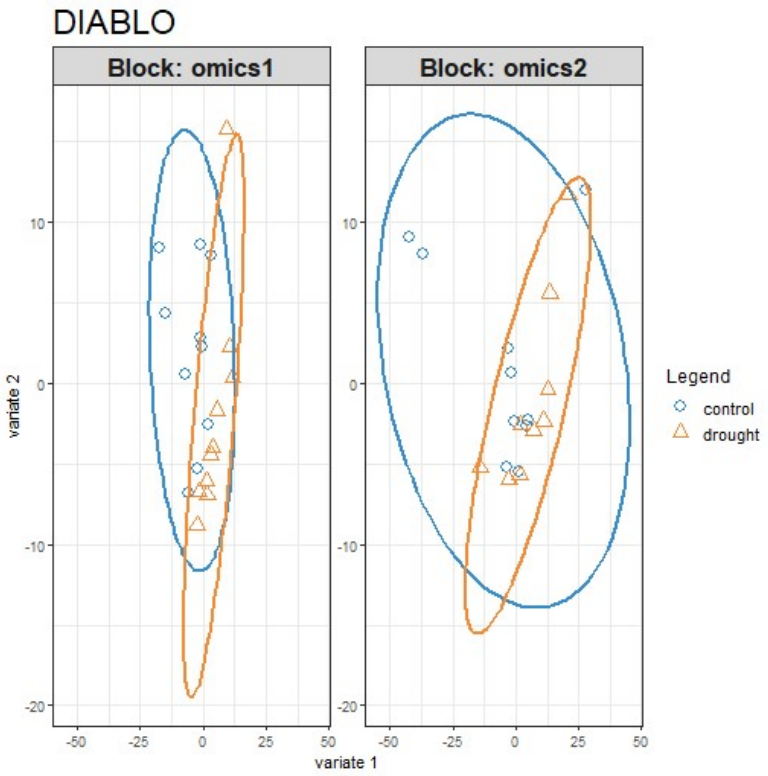


Figure 40: *Multiomics Integration Analysis.* Plot resultant of the"plotIndiv()"function of mixOmics, that projects each sample into a space extended by the components of each block. Block: omics1 concerning the transcriptomics dataset and block: omics2 regarding the metabolomics dataset.

As shown in figure 40, the results are similar to the PCA analysis executed in the exploratory analysis (figure 32). Only the transcriptomics dataset (block: omics1) is capable of differentiating the control and drought treatment to some extent.

MixOmics also allows the execution of variable plots, to visualize and analyze the associations between the selected variables. The variable plot, figure 41 depicts the variables from all blocks selected for components 1 and 2.

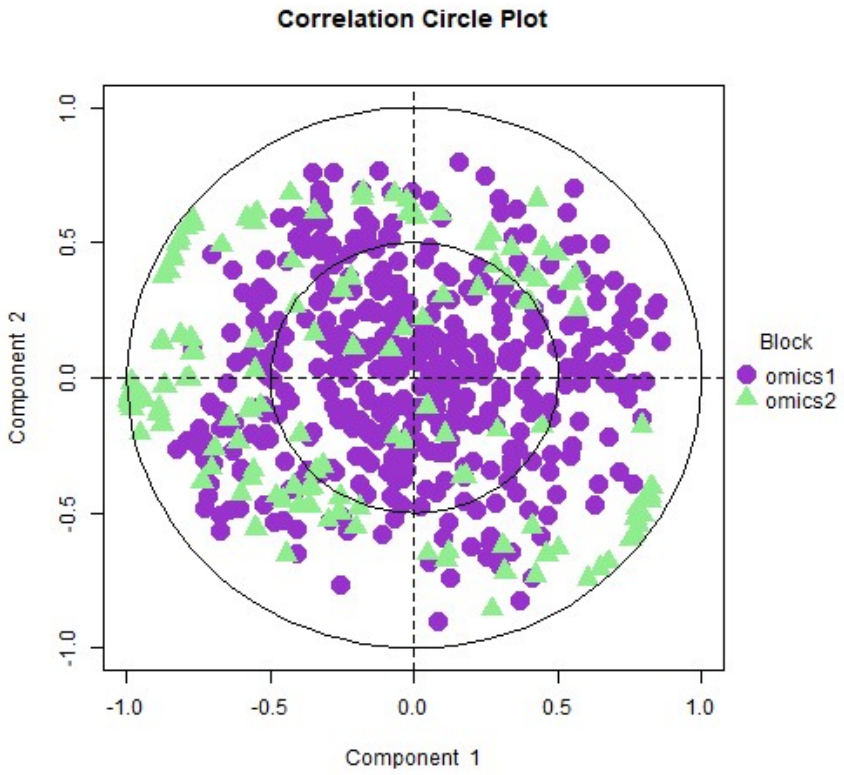


Figure 41: *Multiomics Integration Analysis.* Plot resultant from the *"plotVar()"* function of mixOmics, that allows the visualisation and analysis of the variations between selected variables.

Clusters of points indicate a strong correlation between variables. However, when looking at the resulting figure, we do not see an evident formation of clusters in the space since they are dispersed.

Furthermore, the Circos plot (figure 42) from the mixOmics library also gives a good insight into the correlations between variables of different types. Looking at the figure, we can see that features from both transcriptomics and fluxomics can explain the control and drought treatments.

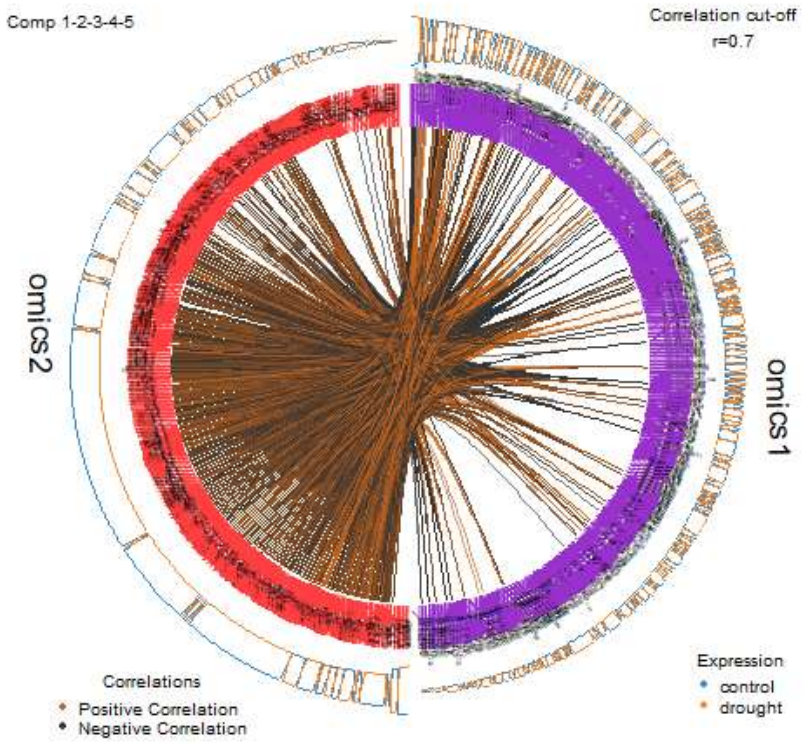


Figure 42: *Multiomics Integration Analysis.* Plot resultant from the *"circosPlot()"* function of mixOmics that allows the visualisation and analysis of the variations between selected variables.

The last plot executed was the loadings plot using the function *"plotLoadings()"*. This plot aids in visualising the loading weights of each selected variable on each component and each data set, (figure 43). As in the Circos plot, both datasets explain the control and drought treatment; however, fluxomics contributes more to the control condition.

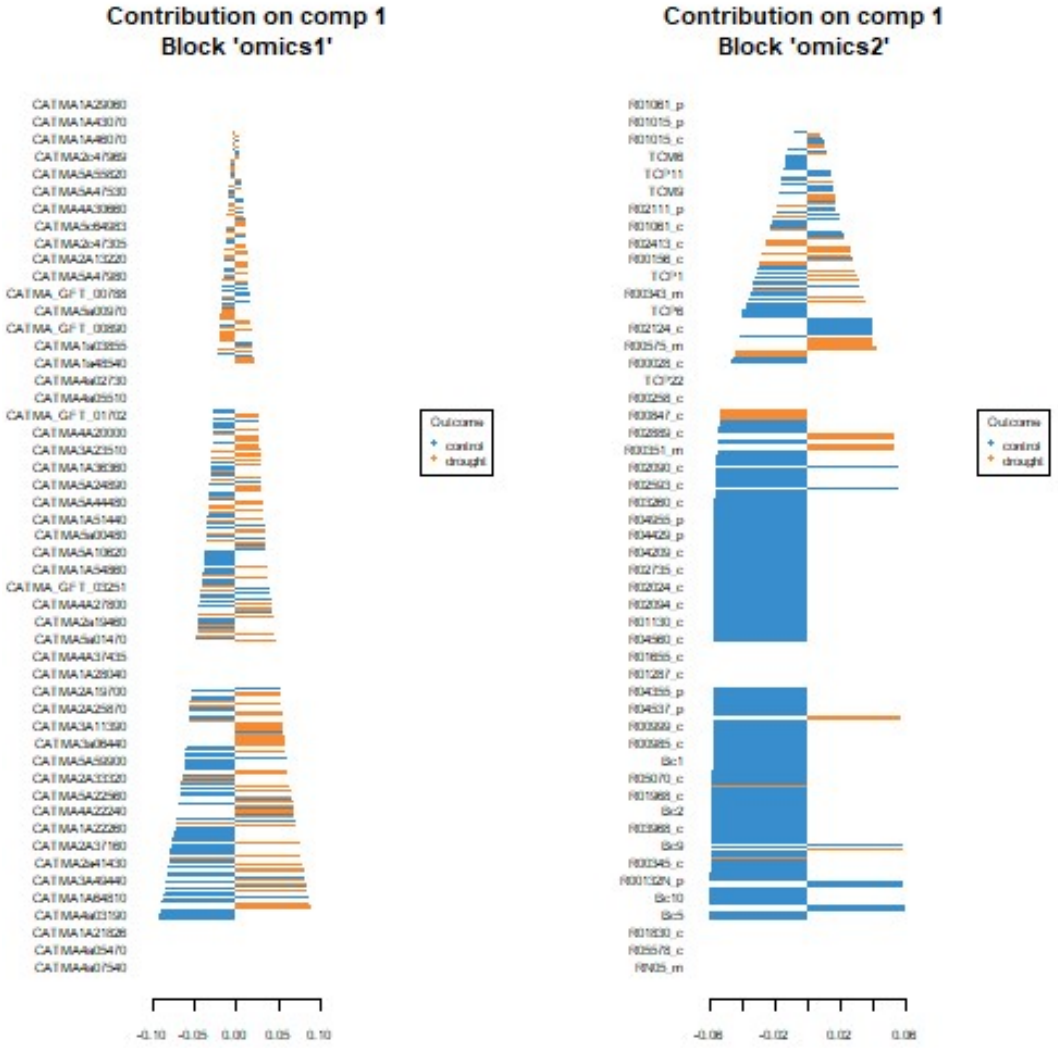


Figure 43: *Multiomics Integration Analysis.* Plot resultant from the *"plotLoadings()"* function of mixOmics, that allows the visualisation of the loading weights of each selected variable on each component and each dataset.

Finally, to evaluate the performance of the DIABLO model, we opted for 10-fold cross-validation repeated 10 times using the *"perf()"* function. We obtained the ROC curve, for both the transcriptomics and fluxomics blocks and the correspondent AUC value for the model performance. Figure 44 shows the final ROC curve. Just like depicted in table 30, DIABLO's accuracy, precision and recall are 0.83, 1, 1, respectively.

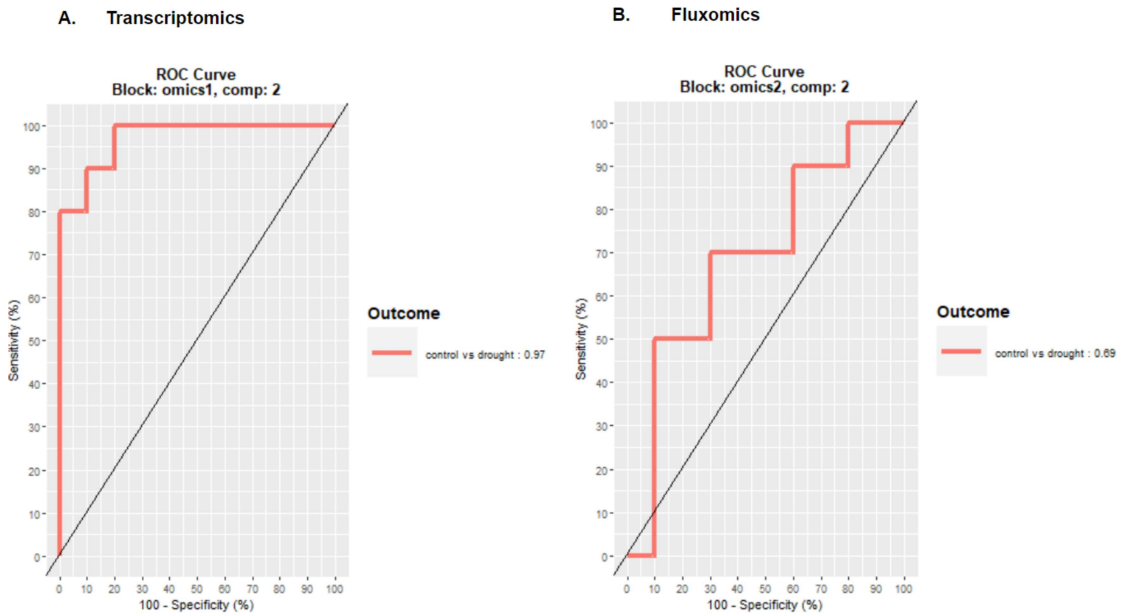


Figure 44: *Multiomics Integration Analysis.* AUC value equal to 0.9704 for the transcriptomics dataset, and 0.6686 for the fluxomics dataset. The overall AUC was of 0.83.

Furthermore, we executed the function *"selectVar()"* and obtained the following 10 most important features for transcriptomics (table 37) and fluxomics dataset (table 38) to evaluate the most relevant features in the DIABLO algorithm.

Table 37: Most relevant features obtained from the DIABLO model for the transcriptomics dataset in **Case Study II.**

CATMA ID	TAIR7 mapping	Annotation
CATMA4a07540	-	-
CATMA3A43840	At3g50840	phototropic-responsive NPH3 family protein
CATMA5a00120	At5g01090	legume lectin family protein
CATMA1A18630	At1g19610	LCR78/PDF1.4 (Low-molecular-weight cysteine-rich 78)
CATMA1A69330	At1g80150	pentatricopeptide (PPR) repeat-containing protein
CATMA1a05880	At1g06800	lipase class 3 family protein
CATMA1A22270	At1g23200	pectinesterase family protein
CATMA3A23280	At3g23280	zinc finger (C3HC4-type RING finger) family protein / ankyrin repeat family protein
CATMA4A13710	At4g13530	unknown protein
CATMA4c42582	At4g31640	transcriptional factor B3 family protein

Table 38: Most relevant features obtained from the DIABLO model for the fluxomics dataset in **Case Study II**.

Reaction	**Subsystem**
BIO_L	Biomass synthesis (Leaf)
Bc6	Glutamate drain
Bc25	Asparagine drain
R00578_c	Nitrogen metabolism;Alanine and aspartate metabolism
Bc17	Glutamine drain
R03652N_c	Glutamate metabolism;Aminoacyl-tRNA biosynthesis
R01830_c	Pentose phosphate pathway
Bc3	Cellulose drain

- **SMSPL**

For the second concatenation-based approach, we chose the SMSPL model that can predict subtypes and identify potentially multiomics signatures. We first created a list of omics matrices, one for the train and the other for the test datasets. After initializing all the parameters, we proceeded with the start of the classifier and posterior optimization stage. Lastly, we ran the best validation map and evaluated its performance. Figure 45 shows the ROC curve and AUC value for the train and test prediction.

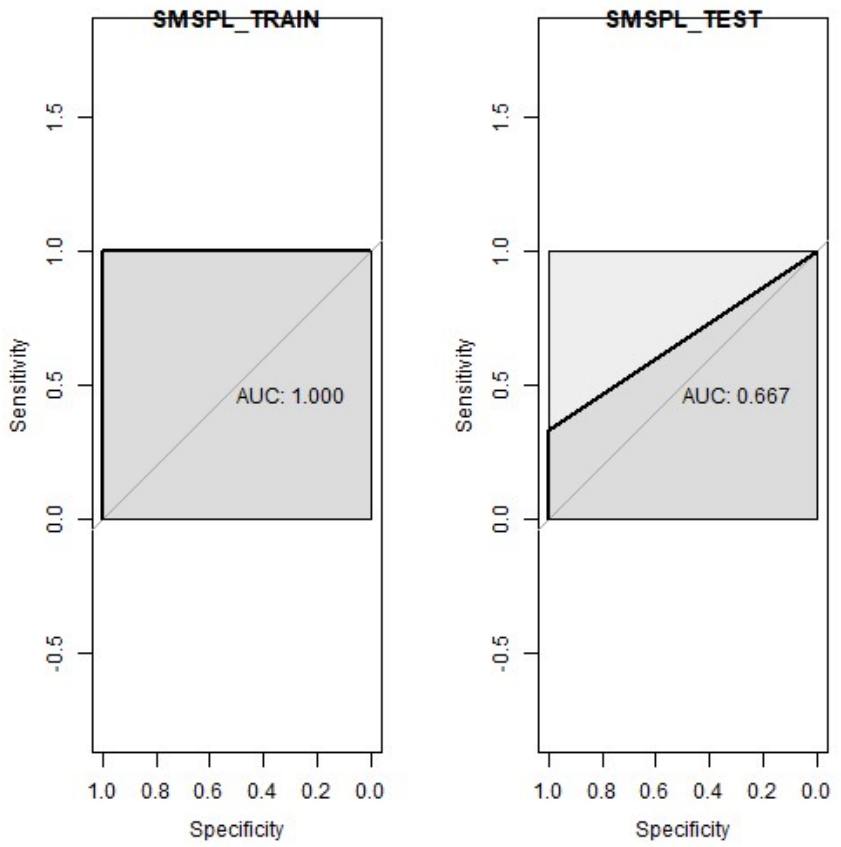


Figure 45: *Multiomics Integration Analysis.* SMSPL ROC curve. AUC value for the train train dataset is 1 and for the test dataset the value is 0.667.

The test dataset prediction, as pictured in table 30, obtained an accuracy, precision and recall of 0.6667, 1 and 0.6, respectively. Furthermore, it also calculates the most important features for the transcriptomics (table 39) and fluxomics dataset (table 40).

Table 39: Most relevant features obtained from the SMSPL model for the transcriptomics dataset in **Case Study II**.

CATMA ID	TAIR7 mapping	Annotation
CATMA1a22270	At1g23200	pectinesterase family protein
CATMA4a03190	At4g02840	small nuclear ribonucleoprotein D1
CATMA3a18630	At3g18980	F-box family protein
CATMA5c64749	At5g44340	TUB4 (tubulin beta-4 chain)
CATMA5a38240	At5g42470	unknown protein
CATMA1A51440	At1g62320	early-responsive to dehydration protein-related

Table 40: Most relevant features obtained from the SMSPL model for the fluxomics dataset in **Case Study II**.

Reaction	Subsystem
TCP7	Triose phosphate translocator (G3P)
R04780_p	Glycolysis / Gluconeogenesis; Pentose phosphate pathway;Fructose and mannose metabolism (drought)
R02950_c	Coumarine and phenylpropanoid biosynthesis (Lignin subunit; coniferyl alcohol)
R01561_c	Purine metabolism

Transformation-based Integration

- **SNFtool**

For the transformation-based integration, we used the library *SNFtool* to implement the SNFtool model. First, we calculated the pairwise distance and next, we created the similarity graphs (presented in the supplementary figure S1). Then, the graphs were fused and the overall matrix was computed using similarity network fusion. We also obtain a list containing the rank based on the normalized mutual information for each feature, indicating the features with more influence, which we can see in table 41 for the transcriptomics dataset and table 42 for the fluxomics dataset. Furthermore, we executed the spectral clustering that gave information regarding the final subtype information (supplementary figure S2). Additionally, we evaluate the accuracy of the obtained clustering results. Lastly, we predicted the new labels using label propagation. The model's accuracy, precision and recall were 0.6667, 0.6 and 1, respectively. The ROC curve is depicted in figure 46, with an AUC value of 0.667.

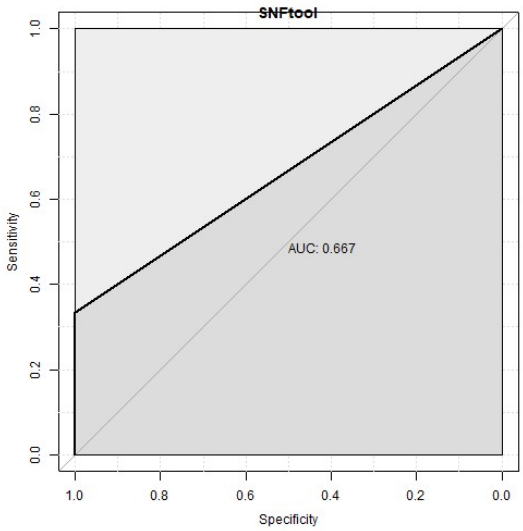


Figure 46: *Multiomics Integration Analysis.* SNFtool ROC curve. The AUC value is 0.667.

Analysing these results, it seems that if we had more samples we could obtain a better result, but we still identified some relevant features. To further validate this model, we must apply it to new data.

Table 41: Most relevant features obtained from the SNFtool model for the transcriptomics dataset in **Case Study II**.

CATMA ID	**TAIR7 mapping**	**Annotation**
CATMA5a50910	At5g55140	ribosomal protein L30 family protein
CATMA5a02390	-	-
CATMA5A43690	At5g47710	C2 domain-containing protein
CATMA3A43840	At3g50840	phototropic-responsive NPH3 family protein
CATMA4a02730	At4g02425	unknown protein
CATMA2a16900	At2g18240	RER1 protein
CATMA5a13690	At5g15430	calmodulin-binding protein-related
CATMA5a45910	At5g49990	xanthine/uracil permease family protein
CATMA1a21826	At1g22780	PFL (POINTED FIRST LEAVES)
CATMA1c71468	At1g30060	COP1-interacting protein-related

Table 42: Most relevant features obtained from the SMSPL model for the fluxomics dataset in **Case Study II**.

Reaction	Subsystem
TCX16	Glycerate transport
R01324_c	Citrate cycle (TCA cycle)
R01325_x	Glyoxylate and dicarboxylate metabolism; Reductive carboxylate cycle (CO2 fixation); Citrate cycle (TCA cycle)
R01900_x	Glyoxylate and dicarboxylate metabolism; Reductive carboxylate cycle (CO2 fixation);Citrate cycle (TCA cycle)
R00243_c	Glutamate metabolism; Nitrogen metabolism; Urea cycle and metabolism of amino groups; Arginine and proline metabolism
R03050_c	Butanoate metabolism
BIO_L	Biomass synthesis (Leaf)
TCM1	Pyruvate transporter
TCX2	Serine transporter
TCX13	Glycerate transport

Model-Based Integration

- **Ensemble Classifier with different ML algorithms (Hard and Soft Voting)**

For model-based integration we opted for the option 1 with the soft voting approach.

- **Option 1**

Option 1 executes an ensemble classifier with two SVM models, predicts the outcome of each one and calculates the final prediction using a soft voting approach on all the classifiers predictions. For this case study, we chose the soft voting strategy. The final prediction had an accuracy, precision and recall of 0.833, 0.8 and 1, respectively. The ROC curve plot is present in figure 47, with an AUC value of 0.75.

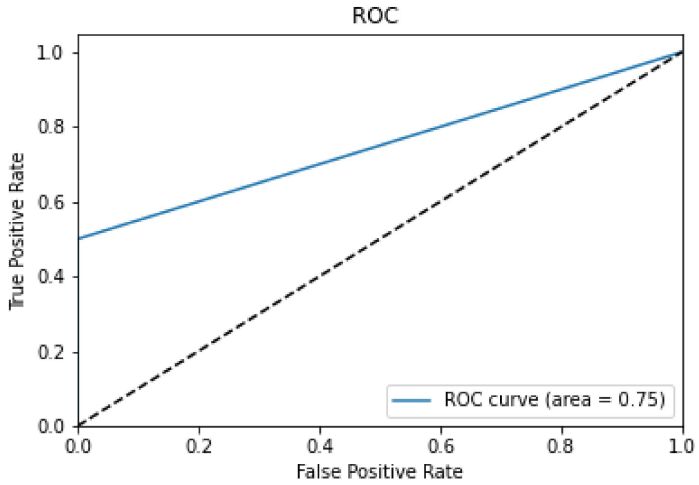


Figure 47: *Multiomics Integration Analysis.* Ensemble Classifier (option 1) roc curve. The AUC value is 0.75.

We also identified the top 4 features with more importance for the transcriptomics and fluxomics dataset, shown in table 43.

Table 43: Most relevant features obtained from the ensemble classifier (option 1) model for the model-based integration approach in **Case Study II**.

CATMA ID	**TAIR7 mapping**	**Annotation**
CATMA1c72010	At1g66390	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2)
CATMA4c42685	At4g37980	ELI3-1 (ELICITOR-ACTIVATED GENE 3)
CATMA1a28510	At1g30500	CCAAT-binding transcription factor (CBF-B/NF-YA) family protein
CATMA5A47680	At5g51750	subtilase family protein
Reaction	**Subsystem**	
R01195_p	Ferredoxin transhydrogenase	
TCP7	Triose phosphate translocator (G3P)	
R00342_p	Citrate cycle (TCA cycle); Glyoxylate and dicarboxylate metabolism; Pyruvate metabolism; Reductive carboxylate cycle (CO2 fixation);Carbon fixation (control)	
R02739_c	Pentose phosphate pathway;Glycolysis / Gluconeogenesis (drought)	

Feature Relevance

Looking at the most relevant features that prevail in most models of the multiomics integration analysis, for the transcriptomics dataset, we identified two relevant features:

- *At3g50840 (phototropic-responsive NPH3 family protein)*, identified in DIABLO and SNFtool - Up-regulated in drought conditions, although little information is identified about this protein under drought, the Cortés and Blair article mentions this protein as being overexpressed in drought conditions;
- *At1g23200 (pectinesterase family protein)*, identified in DIABLO and SMSPL - involved in cell wall modification and organization, that has been identified in other articles, which concluded that drought stress limits cell growth and alters the cell wall [211, 212].

However, when looking at all the results, we were able to identify some patterns regarding the drought condition for both transcriptomics and fluxomics data. The most relevant features of the transcriptomics dataset were related to cell wall modifications and responses against osmotic stress. These roles are important in drought conditions to limit the cell growth, since the concomitant restructuring of the cell wall allows growth processes to occur at lower water contents [227], and also to protect the cell from osmotic stress caused by drought conditions. Regarding the fluxomics dataset, only two most common reactions were identified:

- *BIO_L* - Biomass synthesis (Leaf), identified in DIABLO and SNFtool in control conditions;
- *TCP7* - Triose phosphate translocator (G3P), identified in SMSPL and Ensemble in drought conditions.

Nevertheless, looking at all the results from the fluxomics dataset, we find some pathways that repeat themselves in drought conditions. It is the case of Triosephosphate translocator (G3P); Glyoxylate and dicarboxylate metabolism, Reductive carboxylate cycle (CO_2 fixation), Citrate cycle (TCA cycle); and Pentose phosphate pathway and Glycolysis / Gluconeogenesis. Regarding the triose phosphate translocator (G3P), little is known about the relation of this reaction to the drought stress response in *Arabidopsis thaliana*, but they were also considered relevant in the original article [179]. The CO_2 fixation pathway and the TCA cycle had previously been noted as important pathways in drought tolerance in the differential expression analysis. Furthermore, we noticed the Glyoxylate and dicarboxylate metabolism, which was mentioned in some articles as a possible factor in the response to drought stress [228]. The pentose phosphate pathway and glycolysis/ gluconeogenesis are active in the drought treatment as drought conditions promote an increase in the levels of sugars, like fructose and mannose and other compounds, such as antioxidants, such as flavonoids, and polyphenols [224].

Unsupervised Learning

The last step in our project was the implementation of unsupervised learning methods for multiomics integration. We used three models for each type of integration in order to discover similarities, differences, hidden patterns or data grouping in multiomics unlabeled datasets. However, the transformation-based and model-based unsupervised models did not provide much information besides the cluster plot so they were omitted from the results and discussion.

- **Concatenation-Based Integration**

  For the concatenation-based integration, we implemented the MFA algorithm, using the library *FactoMineR*. The same steps mentioned in **Case Study I** were performed here. Looking at figure figure 48, that demonstrates the contribution of the groups to dimension 1 and dimension 2. The plot in supplementary figure S3 illustrates the contribution of the different groups regarding dimension 1 and dimension 2.

  - **MFA**

Furthermore, to visualize in detail the contribution of the quantitative variables (in %) to the definition of the dimensions, we created figure **??**.

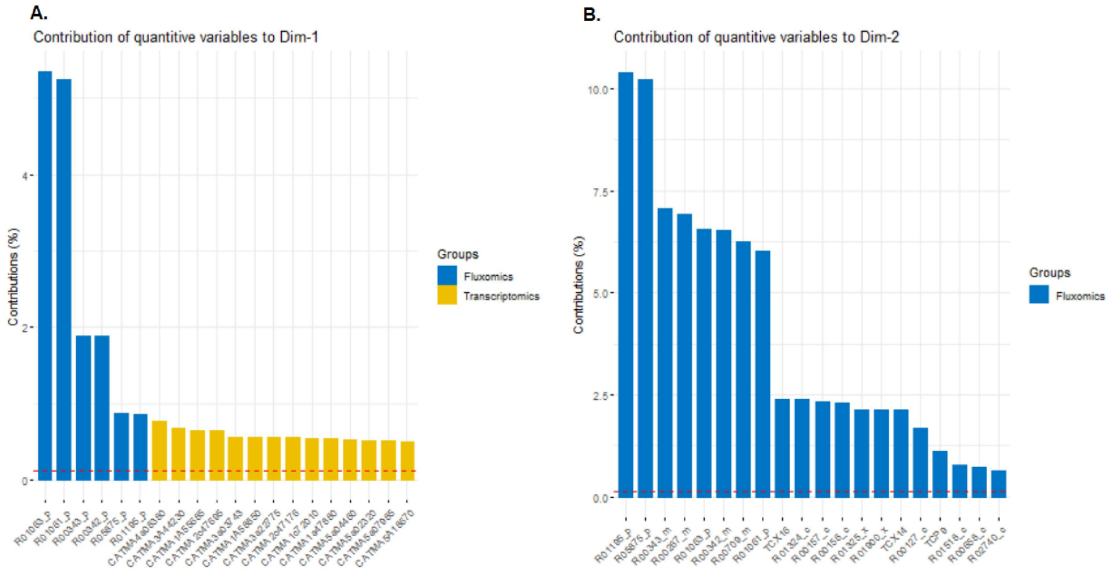


Figure 48: *Multiomics Integration Analysis with Unsupervised Learning.* Graph of the variables, that shows the contribution of quantitative variables relative to the (A) Dimension 1 and (B) Dimension 2.

As shown in figure 48, the variables with the larger value contribute the most to the definition; hence, these are the most important in explaining the variability in the data set. Table 44 shows the features that explain most of the variability in the data set, for the dimension 1 and 2 (table 44 and table 45, respectively).

Table 44: Features that explain the most variability in the dataset according to the MFA unsupervised model for Dimension 1.

Dimension 1		
Reaction	**Subsystem**	
R01063_p	Glycolysis / Gluconeogenesis;Carbon fixation	
R01061_p	Glycolysis / Gluconeogenesis;Carbon fixation (control)	
R00343_p	Pyruvate metabolism;Carbon fixation	
R00342_p	Citrate cycle (TCA cycle); Glyoxylate and dicarboxylate metabolism; Pyruvate metabolism; Reductive carboxylate cycle (CO2 fixation);Carbon fixation (control)	
R05875_p	Ferredoxin transhydrogenase	
R01195_p	Ferredoxin transhydrogenase (control)	
CATMA ID	**TAIR7 mapping**	**Annotation**
CATMA4a08380	At4g08570	heavy-metal-associated domain-containing protein / copper chaperone (CCH)-related
CATMA3a44230	At3g51240	F3H (TRANSPARENT TESTA 6)
CATMA1a55685	At1g66390	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2)
CATMA2c47695	At2g41100	TCH3 (TOUCH 3)

Table 45: Features that explain the most variability in the dataset according to the MFA unsupervised model for Dimension 2.

Dimension 2	
Reaction	**Subsystem**
R01195_p	Ferredoxin transhydrogenase (control)
R05875_p	Ferredoxin transhydrogenase
R00343_m	Pyruvate metabolism;Carbon fixation
R00267_m	Reductive carboxylate cycle (CO2 fixation); Glutathione metabolism (control)
R01063_p	Glycolysis / Gluconeogenesis;Carbon fixation
R00342_m	Citrate cycle (TCA cycle); Glyoxylate and dicarboxylate metabolism; Pyruvate metabolism; Reductive carboxylate cycle (CO2 fixation); Carbon fixation
R00709_m	Citrate cycle (TCA cycle)

Analysing the results, we see that the fluxomics dataset features are the most important for both dimension 1 and dimension 2. The top features in table 39 for dimension 1, regarding the drought condition, are related to glycolysis/gluconeogenesis, Carbon Fixation,TCA cycle, mentioned in the previous analysis. However, we also identified new pathways, like pyruvate metabolism and ferredoxin transhydrogenase. The pyruvate metabolism is important in drought responses since the pyruvate carriers in guard cells are responsible for ABA signalling [229]. Additionally, the ferredoxin transhydrogenase is identified in both control and drought conditions, because of the overall expression levels of ferredoxin-NADP+-oxidoreductase (FNR) genes that are increased upon drought [230].

Regarding the transcriptomics features identified relevant for dimension 1, the only feature described before was At1g66390 (PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2)). We identified At4g08570, a heavy-metal-associated domain-containing protein/copper chaperone (CCH)-related. In Barth et al. article, the authors mentioned that a heavy-metal-associated domain-containing protein is induced during drought stress. Furthermore, we identified At3g51240, an F3H (Transparent Testa 6) gene, that is involved in catalysing intermediates for the biosynthesis of flavonols, anthocyanidins, catechins and proanthocyanidins in plants, that function as scavengers of ROS, therefore contributing to abiotic stress tolerance [213]. The other feature, At2g41100 TCH3 (TOUCH 3), encodes a calmodulin-like protein, which acts as Ca2+ sensors in plants and is known to be involved in various stress reactions like drought [232].

On the other hand, looking at table 40, we find the same pathways that were involved with drought responses in the previous analysis, such as, ferredoxin transhydrogenase, pyruvate metabolism, carbon fixation, TCA cycle, and glyoxylate and dicarboxylate metabolism.

5.3 SUMMARY

This section, gives a brief summary of the advantages and disadvantages of each multiomics integration model performed in this project, depicted in table 46.

Table 46: Summary. Advantages and Disadvantages of every model analyzed in this project, regarding performance, attainment of feature relevance, running time and model availability and implementation.

Multiomics Integration Supervised Models	Performance	Provides Feature Relevance	Running Time	More Information	Model Availability/ Implementation
Concatenation-Based Integration	-	-	-	-	Excellent
DIABLO	Good	Yes (original code)	Slow	Yes	-
SMSPL	Good	Yes (original code)	Fast	No	-
Stack Generalization	Good	Yes (implemented)	Fast	No	-
Lasso Regression	Needs more samples	Yes (implemented)	Fast	No	-
SVM	Good	Yes (original code)	Fast	No	-
ANN	Good	Yes (implemented)	Fast	No	-
RF	Good	Yes (original code)	Fast	No	-
Transformation-Based	-	-	-	-	Medium
SNFtool	Good	Yes (original code)	Fast	Yes	-
Graph-CAN	Needs more samples	No	Fast	No	-
Kernel-Integrated RVM and Boosted-RVM	Good	No	Fast	No	-
Model-Based Integration	-	-	-	-	Difficult
Ensemble Classifier with different ML algorithms (Hard and Soft Voting)	Better in the Soft Voting strategy	Yes (implemented)	Slow, due to feature relevance	No	-

The concatenation-based integration, which is the most analysed and reviewed integration approach, has several supervised models available in publications. DIABLO, has an overall good performance, providing the most relevant features, and additional information providing an interesting look into both datasets. However, the running time, depending on the number of samples could be extensive.

The SMSPL model's performance was good, allowing to obtain the most important features fast. However, it did not provide additional information. The stack generalization model was similar to the SMSPL in terms of advantages, but we must implement feature relevance.

On the other hand, Lasso Regression due to the lower number of samples in both datasets was omitted. SVM, RF and ANN, provide an overall good performance, although performing poorly for Case Study II, due to the small number of samples. Nevertheless, they provided feature importance and a fast running time. No other information was provided except for the error metrics.

For the transformation-based integration, it was harder to obtain the models, as thre are fewer publications regarding this type of integration, especially in Python and R. The SNFtool model, was the only providing the features' importance; therefore, the other two models were omitted from the discussion. However, all the models seem to have a good performance, although this conclusion is not absolute due to the small number of samples. SNFtool also provided more information and had a fast running time.

Lastly, for the model-based integration, the available algorithms were nearly non-existent. We developed an ensemble classifier with hard and soft voting; however, feature relevance was hard to

obtain and implement. Furthermore, due to the packages used for feature relevance the model's running time was slow. Additionally, it did not provide more information besides the error metrics.

6

CONCLUSIONS AND FUTURE WORK

In this work, several models were developed to integrate multiple omics data. The main goal was to use the multiomics integration models to improve our knowledge of plants' metabolic phenotypes when facing environmental stresses and diseases and verify the advantages of performing multiomics analysis compared to individual omics analysis.

The first hurdle was the lack of plant datasets with a good number of samples and subject to more than one type of omics analysis, leading to the impossibility of drawing reliable conclusions about the performance of the models. The selected datasets were derived from extensive searches in several databases, opting for, as a Case study I, two datasets of transcriptomics and metabolomics analysis, with 73 samples of *Vitis vinifera* regarding berry development, and, for Case Study II, two datasets of transcriptomics and fluxomics analysis, with 26 samples, for *Arabidopsis thaliana*, regarding control and drought treatment, due to the lack of fluxomics analysis in *Vitis vinifera* species.

The second problem identified was that although there were a considerable amount of multiomics integration models, none of them was designed for plants, focusing on the health issue, especially in humans, with the TCGA dataset being the most used to validate the models. Within the multiomics integration models, the concatenation-based integration strategy models are the ones with the most offer, following the transformation-based integration ones, while the model-based integration models are the ones with the least offer and the most difficult to implement. All methods received different omics analyses but for the same samples.

Even so, we selected some methods to integrate into our pipeline. This integration was successful, in which we were able to read the datasets, preprocess them, do an exploratory analysis, perform an individual omics analysis, and execute the multiomics integration analysis with both supervised and unsupervised methods.

Regarding the concatenation-based integration models, they were the easiest to implement. Also, as they are one of the most studied integration strategies, some of the models offer innovative and interesting ways of looking at data and observing relationships between them. This is the case of the DIABLO model, which in addition to having a good performance, even with the small sample size, manages to identify the most important features. The only downside is that it can take some time to run. The SMSPL model, in terms of performance, thanks to its innovative self-pace learning strategy is good and can indicate the features with greater relevance. However, it does not give us more information. For the remaining models, it was possible to obtain the most relevant features using

different packages. Nonetheless, the efficiency, even with the implemented cross-validation, could not be evaluated due to the small number of samples.

Transformation-based integration models, on the other hand, were more difficult to find compared to concatenation-based models; thus, only three methods were selected. In terms of performance, The SNFtool model was able to predict well the outcome for Case Study I and II, even with the reduced number of samples. Additionally, it is capable of identifying the most relevant features and displaying the executed clusters, being also easy to implement in our pipeline. The remaining transformation-based integration models (Graph-CAN, Kernel-RVM and Boosted-RVM model), created by the same author, had a good performance. However, since they were not able to identify the most relevant features, nor provide more information, precluding the biological interpretation of the results, we omitted these models from our discussion.

Finally, the model-based integration models were the most difficult to find and implement. We did not find any model available, based on this type of integration, and easy to implement in our pipeline. However, after extensive research in publications, we chose an ensemble classifier, which combined all the predictions made by different models, using different voting strategies, to calculate the final prediction. Since it calculated the final prediction using the other predictions, the search for the most relevant features was challenging for this type of strategy. In terms of performance, the voting soft strategy was more successful compared to the hard strategy.

In the unsupervised learning models, the same was observed. The MFA model related to the concatenation-based integration, in addition, to trying to comprehend the variability of the datasets, obtained the features with greater relevance for each dimension. In turn, NEMO, referring to the transformation-based integration approach, and the BCC, for the model-based integration, only provided the plot of the executed cluster, therefore were omitted. Nevertheless, due to the small number of samples we were unable to conclude anything regarding its performance, both in Case Study I and Case Study II, only that the models could not separate the clusters based on the desired variable.

Regarding the results obtained with the individual omics analysis and the multiomics integration analysis, both obtained interesting results in both Case Study I and Case Study II. However, multiomics integration analysis uncovered a greater number of biological processes that participate in grape development and drought conditions.

In case study I, for the transcriptomics dataset, we identified key functions related to grape development, such as plant growth and development, oxidative phosphorylation, nucleotide binding, cell wall biogenesis and structure, gene information processing, auxin regulation, signal transduction, allergen and production of anthocyanins and flavonoids. The metabolites discovered were also in concordance with the results of the transcriptomics dataset, as the metabolites could be grouped by the following compounds: organic acids, flavonoids/anthocyanins for the pre-Veraison stage and sugar and aromatic compounds for the post-Veraison phase. The most common organic acids were stearic acid and malic acid, the flavonoids/anthocyanins were malvidin-3-O-glucoside and quercetin-3-glucuronide, the most relevant sugars were fructose, glucose and sucrose and lastly, for the aromatic compounds, we identified benzenemethanol and phenylalanine.

In Case Study II, the essential roles in drought condition, in the transcriptomics datasets were related with cell wall modifications and responses against osmotic stress and, for the fluxomics dataset, although most of the reactions were related to the control condition, the most relevant reactions for drought conditions were related with the Triose phosphate translocator (G3P); Glyoxylate and dicarboxylate metabolism, Reductive carboxylate cycle (CO2 fixation), Citrate cycle(TCA cycle); and Pentose phosphate pathway and Glycolysis / Gluconeogenesis.

Therefore the main goal proposed for this thesis was accomplished: develop methods and computation tools based on ML to integrate different omics data and extract knowledge to understand plant behaviour under different environmental conditions and integrate all the collected data, developed tools and algorithms into a pipeline to be integrated into an open-source computation framework. However, some steps can be improved in future work:

- Obtain datasets with a large number of samples for more than one type of omics analysis;
- Develop more models for each type of integration;
- Find ways to obtain biological relations between the different omics using the integration based models;
- Develop more models with the main focus in plants;
- Test other methods for feature selection to improve the process;
- Test other hyperparameter values to optimize the models;
- Optimize the code to speed up the training process and include more types of omics.

After this has been achieved, the next goal is to develop other multiomics integration plant models, able to obtain more accurate relations between the different omics datasets and implement them in our open-source computational framework to allow other users to understand the disease mechanisms and interactions between the plants and its pathogens and predict phenotypes of disease resistance.

BIBLIOGRAPHY

[1] Arthur Germano Fett-Neto and Fett-Neto. *Biotechnology of plant secondary metabolism*. Springer, 2016.

[2] PortugalLive. Economy.). https://www.portugal-live.net/en/portugal/facts/economy.html, . Accessed: 2021-01-29.

[3] World's Top Exports. Portugal's Top 10 Exports.). http://www.worldstopexports.com/portugals-top-10-exports/, . Accessed: 2021-01-29.

[4] Observador. Produção de vinho em Portugal dever'a cair 3% face à campanha anterior.). https://observador.pt/2020/07/30/producao-de-vinho-em-portugal-devera-cair-3-face-a-campanha-anterior/. Accessed: 2021-01-29.

[5] Bor-Sen Chen and Chia-Chou Wu. Systems biology as an integrated platform for bioinformatics, systems synthetic biology, and systems metabolic engineering. *Cells*, 2(4):635–688, 2013.

[6] Claudio Angione. Human systems biology and metabolic modelling: a review—from disease metabolism to precision medicine. *BioMed research international*, 2019, 2019.

[7] Jérôme Grimplet, Grant R Cramer, Julie A Dickerson, Kathy Mathiason, John Van Hemert, and Anne Y Fennell. Vitisnet:"omics" integration through grapevine molecular networks. *PloS one*, 4(12):e8365, 2009.

[8] Plant Metabolic Network (PMN). On www.plantcyc.org. https://pmn.plantcyc.org/organism-summary?object=GRAPE&_ga=2.33353726.1348080849.1611924605-536282905.1611924605. Accessed: 2021-01-29.

[9] Eva Collakova, Jiun Y Yen, and Ryan S Senger. Are we ready for genome-scale modeling in plants? *Plant science*, 191:53–70, 2012.

[10] Semidán Robaina Estévez and Zoran Nikoloski. Generalized framework for context-specific metabolic model extraction methods. *Frontiers in plant science*, 5:491, 2014.

[11] Daniel Machado and Markus Herrgård. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol*, 10(4): e1003580, 2014.

[12] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112):20150571, 2015.

[13] Biswapriya B Misra, Carl Langefeld, Michael Olivier, and Laura A Cox. Integrated omics: tools, advances and future approaches. *Journal of molecular endocrinology*, 62(1):R21–R45, 2019.

[14] Guido Zampieri, Supreeta Vijayakumar, Elisabeth Yaneske, and Claudio Angione. Machine and deep learning meet genome-scale metabolic modeling. *PLoS computational biology*, 15(7): e1007084, 2019.

[15] Pratip Rana, Carter Berry, Preetam Ghosh, and Stephen S Fong. Recent advances on constraint-based models by integrating machine learning. *Current opinion in biotechnology*, 64:85–91, 2020.

[16] Satish C Bhatla and Manju A Lal. *Plant physiology, development and metabolism*. Springer, 2018.

[17] Harinder PS Makkar, Perumal Siddhuraju, Klaus Becker, et al. *Plant secondary metabolites*. Springer, 2007.

[18] Tasiu Isah. Stress and defense responses in plant secondary metabolites production. *Biological research*, 52(1):39, 2019.

[19] Kambiz Baghalian, Mohammad-Reza Hajirezaei, and Falk Schreiber. Plant metabolic modeling: achieving new insight into metabolism and metabolic engineering. *The Plant Cell*, 26(10): 3847–3866, 2014.

[20] N Paul Anulika, E Osamiabe Ignatius, E Sunday Raymond, Osaro-Itota Osasere, and A Hilda Abiola. The chemistry of natural product: Plant secondary metabolites. *Int. J. Technol. Enhanc. Emerg. Eng. Res*, 4:1–8, 2016.

[21] Delphine M Pott, Sonia Osorio, and Jose G Vallarino. From central to specialized metabolism: An overview of some secondary compounds derived from the primary metabolism for their role in conferring nutritional and organoleptic characteristics to fruit. *Frontiers in plant science*, 10, 2019.

[22] Chuanying Fang, Alisdair R Fernie, and Jie Luo. Exploring the diversity of plant metabolism. *Trends in plant science*, 24(1):83–98, 2019.

[23] Ronan Sulpice and Peter C McKeown. Moving toward a comprehensive map of central plant metabolism. *Annual review of plant biology*, 66:187–210, 2015.

[24] Mark Stitt, Ronan Sulpice, and Joost Keurentjes. Metabolic networks: how to identify key components in the regulation of metabolism and growth. *Plant physiology*, 152(2):428–444, 2010.

[25] Nobukazu Shitan. Secondary metabolites in plants: transport and self-tolerance mechanisms. *Bioscience, biotechnology, and biochemistry*, 80(7):1283–1293, 2016.

[26] Distribution of the world's grapevine varieties. https://www.oiv.int/public/medias/5888/en-distribution-of-the-worlds-grapevine-varieties.pdf. Accessed: 2021-06-29.

[27] Carlos Conde, Paulo Silva, Natacha Fontes, Alberto Carlos Pires Dias, Rui M Tavares, Maria João Sousa, Alice Agasse, Serge Delrot, and Hernâni Gerós. Biochemical changes throughout grape berry development and fruit and wine quality. 2007.

[28] Dokoozlian NK. Grape berry growth and development. *In PL Christensen, ed, Raisin Production Manual*, page 30–37, 2000.

[29] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[30] Ron Caspi, Richard Billington, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Peter E Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D Karp. The metacyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic acids research*, 48(D1):D445–D453, 2020.

[31] Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 39 (suppl_1):D38–D51, 2010.

[32] C UniProt. Uniprot: a worldwide hub of protein knowledge. nucleic 944 acids res 47. *D506-D515*, 945, 2019.

[33] Lisa Jeske, Sandra Placzek, Ida Schomburg, Antje Chang, and Dietmar Schomburg. Brenda in 2019: a european elixir core data resource. *Nucleic acids research*, 47(D1):D542–D549, 2019.

[34] Milton H Saier Jr, Vamsee S Reddy, Brian V Tsu, Muhammad Saad Ahmed, Chun Li, and Gabriel Moreno-Hagelsieb. The transporter classification database (tcdb): recent advances. *Nucleic acids research*, 44(D1):D372–D379, 2016.

[35] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.

[36] Pascal Schläpfer, Peifen Zhang, Chuan Wang, Taehyong Kim, Michael Banf, Lee Chae, Kate Dreher, Arvind K Chavali, Ricardo Nilo-Poyanco, Thomas Bernard, et al. Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant physiology*, 173 (4):2041–2059, 2017.

[37] Sushma Naithani, Parul Gupta, Justin Preece, Peter D'Eustachio, Justin L Elser, Priyanka Garg, Daemon A Dikeman, Jason Kiff, Justin Cook, Andrew Olson, et al. Plant reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic acids research*, 48(D1): D1093–D1103, 2020.

[38] Falk Schreiber, Christian Colmsee, Tobias Czauderna, Eva Grafahrend-Belau, Anja Hartmann, Astrid Junker, Björn H Junker, Matthias Klapperstück, Uwe Scholz, and Stephan Weise. Metacrop 2.0: managing and exploring information about crop plant metabolism. *Nucleic acids research*, 40(D1):D1173–D1177, 2012.

[39] Noe Fernandez-Pozo, Naama Menda, Jeremy D Edwards, Surya Saha, Isaak Y Tecle, Susan R Strickler, Aureliano Bombarely, Thomas Fisher-York, Anuradha Pujar, Hartmut Foerster, et al. The sol genomics network (sgn)—from genotype to phenotype to breeding. *Nucleic acids research*, 43(D1):D1036–D1041, 2015.

[40] Tanya Z Berardini, Leonore Reiser, Donghui Li, Yarik Mezheritsky, Robert Muller, Emily Strait, and Eva Huala. The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *genesis*, 53(8):474–485, 2015.

[41] Björn Usadel, Fabien Poree, Axel Nagel, Marc Lohse, ANGELIKA CZEDIK-EYSENBERG, and Mark Stitt. A guide to using mapman to visualize and compare omics data in plants: a case study in the crop species, maize. *Plant, cell & environment*, 32(9):1211–1229, 2009.

[42] Peter D Karp, Richard Billington, Ron Caspi, Carol A Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M Keseler, Markus Krummenacker, Peter E Midford, Quang Ong, et al. The biocyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, 20 (4):1085–1093, 2019.

[43] Debmalya Barh, Muhammad Sarwar Khan, and Eric Davies. *PlantOmics: the omics of plant science*. Springer, 2015.

[44] Wan Mohd Aizat, Ismanizan Ismail, and Normah Mohd Noor. Recent development in omics studies. In *Omics Applications for Systems Biology*, pages 1–9. Springer, 2018.

[45] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating'omics' data sets. *Nature reviews Molecular cell biology*, 7(3):198–210, 2006.

[46] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[47] Rasko Leinonen, Hideaki Sugawara, Martin Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39(suppl_1): D19–D21, 2010.

[48] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[49] Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, et al. The ncbi dbgap database of genotypes and phenotypes. *Nature genetics*, 39(10):1181–1186, 2007.

[50] W Mulyasasmita and SC Heilshorn. Protein-engineered biomaterials: Synthesis and characterization. 2011.

[51] Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, et al. Arrayexpress—a public repository for microarray gene expression data at the ebi. *Nucleic acids research*, 31(1):68–71, 2003.

[52] Yasset Perez-Riverol, Attila Csordas, Jingwen Bai, Manuel Bernal-Llinares, Suresh Hewapathirana, Deepti J Kundu, Avinash Inuganti, Johannes Griss, Gerhard Mayer, Martin Eisenacher, et al. The pride database and related tools and resources in 2019: improving support for quantification data. *Nucleic acids research*, 47(D1):D442–D450, 2019.

[53] Eric W Deutsch, Attila Csordas, Zhi Sun, Andrew Jarnuczak, Yasset Perez-Riverol, Tobias Ternent, David S Campbell, Manuel Bernal-Llinares, Shujiro Okuda, Shin Kawano, et al. The proteomexchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic acids research*, page gkw936, 2016.

[54] Tobias Schmidt, Patroklos Samaras, Martin Frejno, Siegfried Gessulat, Maximilian Barnert, Harald Kienegger, Helmut Krcmar, Judith Schlegl, Hans-Christian Ehrlich, Stephan Aiche, et al. Proteomicsdb. *Nucleic acids research*, 46(D1):D1271–D1281, 2018.

[55] Frank Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N Loevenich, and Ruedi Aebersold. The peptideatlas project. *Nucleic acids research*, 34(suppl_1):D655–D658, 2006.

[56] Robertson Craig, John P Cortens, and Ronald C Beavis. Open source system for analyzing, validating, and storing protein identification data. *Journal of proteome research*, 3(6):1234–1242, 2004.

[57] Timothy K Toby, Luca Fornelli, and Neil L Kelleher. Progress in top-down proteomics and the analysis of proteoforms. *Annual review of analytical chemistry*, 9:499–519, 2016.

[58] Yuki Moriya, Shin Kawano, Shujiro Okuda, Yu Watanabe, Masaki Matsumoto, Tomoyo Takami, Daiki Kobayashi, Yoshinori Yamanouchi, Norie Araki, Akiyasu C Yoshizawa, et al. The jpost environment: an integrated proteomics data repository and database. *Nucleic acids research*, 47 (D1):D1218–D1224, 2019.

[59] Namrata S Kale, Kenneth Haug, Pablo Conesa, Kalaivani Jayseelan, Pablo Moreno, Philippe Rocca-Serra, Venkata Chandrasekhar Nainala, Rachel A Spicer, Mark Williams, Xuefei Li, et al. Metabolights: an open-access database repository for metabolomics data. *Current protocols in bioinformatics*, 53(1):14–13, 2016.

[60] Adam J Carroll, Murray R Badger, and A Harvey Millar. The metabolomeexpress project: enabling web-based processing, analysis and transparent dissemination of gc/ms metabolomics datasets. *BMC bioinformatics*, 11(1):376, 2010.

[61] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837, 2016.

[62] The metabolomics workbench. https://www.metabolomicsworkbench.org/. Accessed: 2020-12-29.

[63] Hajime Ohyanagi, Tomoyuki Takano, Shin Terashima, Masaaki Kobayashi, Maasa Kanno, Kyoko Morimoto, Hiromi Kanegae, Yohei Sasaki, Misa Saito, Satomi Asano, et al. Plant omics data center: an integrated web repository for interspecies gene expression networks with nlp-based curation. *Plant and Cell Physiology*, 56(1):e9–e9, 2015.

[64] Toru Kudo, Shin Terashima, Yuno Takaki, Ken Tomita, Misa Saito, Maasa Kanno, Koji Yokoyama, and Kentaro Yano. Plantexpress: a database integrating oryzaexpress and arthaexpress for single-species and cross-species gene expression network analyses with microarray-based transcriptome data. *Plant and Cell Physiology*, 58(1):e1–e1, 2017.

[65] Qi Sun, Boris Zybailov, Wojciech Majeran, Giulia Friso, Paul Dominic B Olinares, and Klaas J Van Wijk. Ppdb, the plant proteomics database at cornell. *Nucleic acids research*, 37(suppl_1): D969–D974, 2009.

[66] M Udayakumar, D Prem Chandar, N Arun, J Mathangi, K Hemavathi, and R Seenivasagam. Pmdb: Plant metabolome database—a metabolomic approach. *Medicinal Chemistry Research*, 21(1):47–52, 2012.

[67] Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 44(D1):D67–D72, 2016.

[68] Nuala A O'Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[69] Yoshio Tateno, Tadashi Imanishi, Satoru Miyazaki, Kaoru Fukami-Kobayashi, Naruya Saitou, Hideaki Sugawara, and Takashi Gojobori. Dna data bank of japan (ddbj) for genome scale research in life science. *Nucleic acids research*, 30(1):27–30, 2002.

[70] Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, et al. The european nucleotide archive. *Nucleic acids research*, 39(suppl_1):D28–D31, 2010.

[71] Robert Petryszak, Maria Keays, Y Amy Tang, Nuno A Fonseca, Elisabet Barrera, Tony Burdett, Anja Füllgrabe, Alfonso Munoz-Pomer Fuentes, Simon Jupp, Satu Koskinen, et al. Expression atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic acids research*, 44(D1):D746–D752, 2016.

[72] Charles E Cook, Mary Todd Bergman, Robert D Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. The european bioinformatics institute in 2016: data growth and integration. *Nucleic acids research*, 44(D1):D20–D26, 2016.

[73] Massive: Mass spectrometry interactive virtual environment. *Center for Computational Mass Spectrometry*.

[74] NODE (The National Omics Data Encyclopedia). https://www.biosino.org/node/. Accessed: 2020-12-28.

[75] Minseung Kim and Ilias Tagkopoulos. Data integration and predictive modeling methods for multi-omics datasets. *Molecular omics*, 14(1):8–25, 2018.

[76] F Provost. Glossary of terms special issue on applications of machine learning and the knowledge discovery process. *Machine Learning*, 30:271–274, 1998.

[77] Difference Between Algorithm and Model in Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/. Accessed: 2021-02-01.

[78] Test, Training and Validation Sets. BrainsToBytes.

[79] Miroslava Cuperlovic-Culf. Machine learning methods for analysis of metabolic data and metabolic pathway modeling. *Metabolites*, 8(1):4, 2018.

[80] Javaid Nabi. Machine learning — fundamentals. *Basic theory underlying the field of Machine Learning*, 2018.

[81] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[82] M Sanjay. Why and how to cross validate a model. *Towards Data Science, Nov*, 12:2018, 2020.

[83] 11 Important Model Evaluation Metrics for Machine Learning Everyone should know. Analytis Vidhya. https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/. Accessed: 2021-02-01.

[84] Claude Sammut and Geoffrey I Webb. Leave-one-out cross-validation. *Encyclopedia of machine learning*, pages 600–601, 2010.

[85] Chien-Fu Jeff Wu et al. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986.

[86] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[87] H Jabbar and Rafiqul Zaman Khan. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, pages 163–172, 2015.

[88] Jason Brownlee. An introduction to feature selection. *Machine learning process*, 6, 2014.

[89] Ensemble Learning to Improve Machine Learning Results.How ensemble methods work: bagging, boosting and stacking. https://blog.statsbot.co/ensemble-learning-d1dcd548e936. Accessed: 2021-02-01.

[90] Ajitesh Kumar. Hard vs Soft Voting Classifier Python Example. Data Analytics. https://vitalflux.com/hard-vs-soft-voting-classifier-python-example/, 2020. Accessed: 2021-10-07.

[91] Jason Brownlee. A tour of the most popular machine learning algorithms. *Retrieved from Machine Learning Mastery: https://machinelearningmastery. com/a-tour-of-machine-learning-algorithms*, 2019.

[92] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, José A Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

[93] Jose Cleydson F Silva, Ruan M Teixeira, Fabyano F Silva, Sergio H Brommonschenkel, and Elizabeth PB Fontes. Machine learning approaches and their current application in plant molecular biology: A systematic review. *Plant Science*, 284:37–47, 2019.

[94] Thales Francisco Mota Carvalho, José Cleydson F Silva, Iara Pinheiro Calil, Elizabeth Pacheco Batista Fontes, and Fabio Ribeiro Cerqueira. Rama: a machine learning approach for ribosomal protein prediction in plants. *Scientific reports*, 7(1):1–13, 2017.

[95] José Cleydson F Silva, Thales FM Carvalho, Elizabeth PB Fontes, and Fabio R Cerqueira. Fangorn forest (f2): a machine learning approach to classify genes and genera in the family geminiviridae. *BMC bioinformatics*, 18(1):431, 2017.

[96] Ali Moghimi, Ce Yang, Marisa E Miller, Shahryar F Kianian, and Peter M Marchetto. A novel approach to assess salt stress tolerance in wheat using hyperspectral imaging. *Frontiers in plant science*, 9:1182, 2018.

[97] Salvador Gutiérrez, Juan Fernández-Novales, Maria P Diago, and Javier Tardaguila. On-the-go hyperspectral imaging under field conditions and machine learning for the classification of grapevine varieties. *Frontiers in plant science*, 9:1102, 2018.

[98] Mónica Pineda, María L Pérez-Bueno, and Matilde Barón. Detection of bacterial infection in melon plants by classification methods based on imaging data. *Frontiers in plant science*, 9:164, 2018.

[99] Ping Xuan, Maozu Guo, Xiaoyan Liu, Yangchao Huang, Wenbin Li, and Yufei Huang. Plantmirnapred: efficient classification of real and pseudo plant pre-mirnas. *Bioinformatics*, 27(10): 1368–1376, 2011.

[100] Yonggan Wu, Bo Wei, Haizhou Liu, Tianxian Li, and Simon Rayner. Mirpara: a svm-based software tool for prediction of most probable microrna coding regions in genome scale sequences. *BMC bioinformatics*, 12(1):1–14, 2011.

[101] Ping Xuan, Maozu Guo, Yangchao Huang, Wenbin Li, and Yufei Huang. Maturepred: efficient identification of microrna within novel plant pre-mirnas. *PloS one*, 6(11):e27422, 2011.

[102] Nestoras Karathanasis, Ioannis Tsamardinos, and Panayiota Poirazi. Mirduplexsvm: a high-performing mirna-duplex prediction and evaluation methodology. *PloS one*, 10(5):e0126151, 2015.

[103] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee, and Byoung-Tak Zhang. mitarget: microrna target gene prediction using a support vector machine. *BMC bioinformatics*, 7(1): 1–12, 2006.

[104] Haibo Cui, Jingjing Zhai, and Chuang Ma. mirlocator: machine learning-based prediction of mature micrornas within plant pre-mirna sequences. *PLoS One*, 10(11):e0142753, 2015.

[105] Ilham A Shahmuradov, Viktor V Solovyev, and AJ Gammerman. Plant promoter prediction with confidence estimation. *Nucleic acids research*, 33(3):1069–1076, 2005.

[106] Firoz Anwar, Syed Murtuza Baker, Taskeed Jabid, Md Mehedi Hasan, Mohammad Shoyaib, Haseena Khan, and Ray Walshe. Pol ii promoter prediction using characteristic 4-mer motifs: a machine learning approach. *BMC bioinformatics*, 9(1):414, 2008.

[107] AKM Azad, Saima Shahid, Nasimul Noman, and Hyunju Lee. Prediction of plant promoters based on hexamers and random triplet pair analysis. *Algorithms for Molecular Biology*, 6(1):19, 2011.

[108] Ilham A Shahmuradov, Ramzan Kh Umarov, and Victor V Solovyev. Tssplant: a new tool for prediction of plant pol ii promoters. *Nucleic acids research*, 45(8):e65–e65, 2017.

[109] Dustin T Holloway, Mark Kon, and Charles De Lisi. Integrating genomic data to predict transcription factor binding. *Genome informatics*, 16(1):83–94, 2005.

[110] Bo Jiang, Michael Q Zhang, and Xuegong Zhang. Oscar: one-class svm for accurate recognition of cis-elements. *Bioinformatics*, 23(21):2823–2828, 2007.

[111] Xinbin Dai, Ji He, and Xuechun Zhao. A new systematic computational approach to predicting target genes of transcription factors. *Nucleic acids research*, 35(13):4433–4440, 2007.

[112] Song Cui, Eunseog Youn, Joohyun Lee, and Stephan J Maas. An improved systematic approach to predicting transcription factor target genes using support vector machine. *PloS one*, 9(4): e94519, 2014.

[113] Hanjun Dai, Ramzan Umarov, Hiroyuki Kuwahara, Yu Li, Le Song, and Xin Gao. Sequence2vec: a novel embedding approach for modeling transcription factor binding affinity landscape. *Bioinformatics*, 33(22):3575–3583, 2017.

[114] Douglas B Kell, Robert M Darby, and John Draper. Genomic computing. explanatory analysis of plant expression profiling data using machine learning. *Plant Physiology*, 126(3):943–951, 2001.

[115] Lenwood S Heath, Naren Ramakrishnan, Ronald R Sederoff, Ross W Whetten, Boris I Chevone, Craig A Struble, Vincent Y Jouenne, Dawei Chen, Leonel Van Zyl, and Ruth Grene. Studying the functional genomics of stress responses in loblolly pine with the expresso microarray experiment management system. *Comparative and Functional Genomics*, 3(3):226–243, 2002.

[116] Yunhai Li, Kee Khoon Lee, Sean Walsh, Caroline Smith, Sophie Hadingham, Karim Sorefan, Gavin Cawley, and Michael W Bevan. Establishing glucose-and aba-regulated transcription networks in arabidopsis by microarray analysis and promoter classification using a relevance vector machine. *Genome research*, 16(3):414–427, 2006.

[117] Gema Ancillo, J Gadea, J Forment, José Guerri, and Luis Navarro. Class prediction of closely related plant varieties using gene expression profiling. *Journal of experimental botany*, 58(8): 1927–1933, 2007.

[118] Chuang Ma, Mingming Xin, Kenneth A Feldmann, and Xiangfeng Wang. Machine learning–based differential network analysis: A study of stress-responsive transcriptomes in arabidopsis. *The Plant Cell*, 26(2):520–537, 2014.

[119] Ying Ni, Delasa Aghamirzaie, Haitham Elmarakeby, Eva Collakova, Song Li, Ruth Grene, and Lenwood S Heath. A machine learning approach to predict gene regulatory networks in seed development in arabidopsis. *Frontiers in plant science*, 7:1936, 2016.

[120] Sandeep K Kushwaha, Pallavi Chauhan, Katarina Hedlund, and Dag Ahrén. Nbspred: a support vector machine-based high-throughput pipeline for plant resistance protein nbslrr prediction. *Bioinformatics*, 32(8):1223–1225, 2016.

[121] Tarun Pal, Varun Jaiswal, and Rajinder S Chauhan. Drppp: A machine learning based tool for prediction of disease resistance proteins in plants. *Computers in biology and medicine*, 78: 42–48, 2016.

[122] Jana Sperschneider, Ann-Maree Catanzariti, Kathleen DeBoer, Benjamin Petre, Donald M Gardiner, Karam B Singh, Peter N Dodds, and Jennifer M Taylor. Localizer: subcellular localization prediction of both plant and effector proteins in the plant cell. *Scientific reports*, 7 (1):1–14, 2017.

[123] Jana Sperschneider, Peter N Dodds, Karam B Singh, and Jennifer M Taylor. Apoplastp: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytologist*, 217(4):1764–1778, 2018.

[124] David Toubiana, Rami Puzis, Lingling Wen, Noga Sikron, Assylay Kurmanbayeva, Aigerim Soltabayeva, Maria del Mar Rubio Wilhelmi, Nir Sade, Aaron Fait, Moshe Sagi, et al. Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data. *Communications biology*, 2(1):1–13, 2019.

[125] Jordan Ubbens, Mikolaj Cieslak, Przemyslaw Prusinkiewicz, and Ian Stavness. The use of plant models in deep learning: an application to leaf counting in rosette plants. *Plant methods*, 14 (1):6, 2018.

[126] Konstantinos P Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018.

[127] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2018.

[128] Anita Sathyanarayanan, Rohit Gupta, Erik W Thompson, Dale R Nyholt, Denis C Bauer, and Shivashankar H Nagaraj. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Briefings in bioinformatics*, 21(6):1920–1936, 2020.

[129] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

[130] Introduction to Dimensionality Reduction Technique. https://www.javatpoint.com/dimensionality-reduction-technique. Accessed: 2021-01-20.

[131] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[132] Ignacio González, Sébastien Déjean, Pascal Martin, and Alain Baccini. Cca: An r package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12):1–14, 2008.

[133] Ali-Reza Mohammadi-Nejad, Gholam-Ali Hossein-Zadeh, and Hamid Soltanian-Zadeh. Structured and sparse canonical correlation analysis as a brain-wide multi-modal data fusion approach. *IEEE transactions on medical imaging*, 36(7):1438–1448, 2017.

[134] Peng-Bo Zhang and Zhi-Xin Yang. Robust matrix elastic net based canonical correlation analysis: An effective algorithm for multi-view unsupervised learning. *arXiv preprint arXiv:1711.05068*, 2017.

[135] Jun Chen, Frederic D Bushman, James D Lewis, Gary D Wu, and Hongzhe Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2013.

[136] Kefei Liu, Qi Long, and Li Shen. Grouping effects of sparse cca models in variable selection. *arXiv preprint arXiv:2008.03392*, 2020.

[137] Sijia Huang, Kumardeep Chaudhary, and Lana X Garmire. More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84, 2017.

[138] Irene Sui Lan Zeng and Thomas Lumley. Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinformatics and biology insights*, 12: 1177932218759292, 2018.

[139] Kim-Anh Lê Cao, Debra Rossouw, Christele Robert-Granié, and Philippe Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

[140] Thomas Verron, Robert Sabatier, and Richard Joffre. Some theoretical properties of the o-pls method. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(2):62–68, 2004.

[141] Wenyuan Li, Shihua Zhang, Chun-Chi Liu, and Xianghong Jasmine Zhou. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, 28(19): 2458–2466, 2012.

[142] D Hervas, José Manuel Prats-Montalbán, A Lahoz, and Alberto Ferrer. Sparse n-way partial least squares with r package snpls. *Chemometrics and Intelligent Laboratory Systems*, 179: 54–63, 2018.

[143] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11): e1005752, 2017.

[144] Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.

[145] Wu Nai-Qi Liang Yong Zhang Hui Yang, Zi-Yi and Ren Yan-Qiong. Smspl: Robust multimodal approach to integrative analysis of multi-omics data. *IEEE Transactions on Cybernetics*, 2020.

[146] Hervé Abdi, Dominique Valentin, et al. Multiple factor analysis (mfa). *Encyclopedia of measurement and statistics*, pages 1–14, 2007.

[147] Bohyun Lee, Shuo Zhang, Aleksandar Poleksic, and Lei Xie. Heterogeneous multi-layered network model for omics data integration and analysis. *Frontiers in Genetics*, 10:1381, 2020.

[148] Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, and Luciano Milanesi. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17(S2):S15, 2016.

[149] Noemi Di Nanni, Matteo Bersanelli, Luciano Milanesi, and Ettore Mosca. Network diffusion promotes the integrative analysis of multiple omics. *Frontiers in Genetics*, 11:106, 2020.

[150] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333, 2014.

[151] Kang K Yan, Hongyu Zhao, and Herbert Pang. A comparison of graph-and kernel-based–omics data integration algorithms for classifying complex traits. *BMC bioinformatics*, 18(1):1–13, 2017.

[152] Nimrod Rappoport and Ron Shamir. Nemo: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, 35(18):3348–3356, 2019.

[153] Yinyin Yuan, Richard S Savage, and Florian Markowetz. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol*, 7(10):e1002227, 2011.

[154] Ronglai Shen, Qianxing Mo, Nikolaus Schultz, Venkatraman E Seshan, Adam B Olshen, Jason Huse, Marc Ladanyi, and Chris Sander. Integrative subtype discovery in glioblastoma using icluster. *PloS one*, 7(4):e35236, 2012.

[155] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[156] Eric F Lock and David B Dunson. Bayesian consensus clustering. *Bioinformatics*, 29(20): 2610–2616, 2013.

[157] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12): i268–i275, 2015.

[158] Saúl Fraire-Velázquez, Raúl Rodríguez-Guerra, and Lenin Sánchez-Calderón. Abiotic and biotic stress response crosstalk in plants. *Abiotic stress response in plants—physiological, biochemical and genetic perspectives*, pages 3–26, 2011.

[159] Dhivyaa Rajasundaram and Joachim Selbig. More effort—more results: recent advances in integrative 'omics' data analysis. *Current opinion in plant biology*, 30:57–61, 2016.

[160] Ryan Ghan, Steven C Van Sluyter, Uri Hochberg, Asfaw Degu, Daniel W Hopper, Richard L Tillet, Karen A Schlauch, Paul A Haynes, Aaron Fait, and Grant R Cramer. Five omic technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (vitis vinifera l.) cultivars. *BMC genomics*, 16(1):1–26, 2015.

[161] Paulo A Zaini, Rafael Nascimento, Hossein Gouran, Dario Cantu, Sandeep Chakraborty, My Phu, Luiz R Goulart, and Abhaya M Dandekar. Molecular profiling of pierce's disease outlines the response circuitry of vitis vinifera to xylella fastidiosa infection. *Frontiers in plant science*, 9: 771, 2018.

[162] Stefania Savoi, Darren CJ Wong, Panagiotis Arapitsas, Mara Miculan, Barbara Bucchetti, Enrico Peterlunger, Aaron Fait, Fulvio Mattivi, and Simone D Castellarin. Transcriptome and metabolite profiling reveals that prolonged drought modulates the phenylpropanoid and terpenoid pathway in white grapes (vitis vinifera l.). *BMC plant biology*, 16(1):1–17, 2016.

[163] Millie Rådjursöga, Helen M Lindqvist, Anders Pedersen, B Göran Karlsson, Daniel Malmodin, Lars Ellegård, and Anna Winkvist. Nutritional metabolomics: postprandial response of meals relating to vegan, lacto-ovo vegetarian, and omnivore diets. *Nutrients*, 10(8):1063, 2018.

[164] Dhivyaa Rajasundaram, Jean-Luc Runavot, Xiaoyuan Guo, William GT Willats, Frank Meulewaeter, and Joachim Selbig. Understanding the relationship between cotton fiber properties and non-cellulosic cell wall polysaccharides. *PloS one*, 9(11):e112168, 2014.

[165] Max Bylesjö, Daniel Eriksson, Miyako Kusano, Thomas Moritz, and Johan Trygg. Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. *The Plant Journal*, 52(6):1181–1191, 2007.

[166] Anita Zamboni, Mariasole Di Carli, Flavia Guzzo, Matteo Stocchero, Sara Zenoni, Alberto Ferrarini, Paola Tononi, Ketti Toffali, Angiola Desiderio, Kathryn S Lilley, et al. Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. *Plant Physiology*, 154(3):1439–1459, 2010.

[167] Vaibhav Srivastava, Ogonna Obudulu, Joakim Bygdell, Tommy Löfstedt, Patrik Rydén, Robert Nilsson, Maria Ahnlund, Annika Johansson, Pär Jonsson, Eva Freyhult, et al. Onpls integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipi-superoxide dismutase populus plants. *Bmc Genomics*, 14(1): 1–16, 2013.

[168] Andrea Anesi, Matteo Stocchero, Silvia Dal Santo, Mauro Commisso, Sara Zenoni, Stefania Ceoldo, Giovanni Battista Tornielli, Tracey E Siebert, Markus Herderich, Mario Pezzotti, et al. Towards a scientific interpretation of the terroir concept: plasticity of the grape berry metabolome. *BMC plant biology*, 15(1):1–17, 2015.

[169] Animesh Acharjee, Bjorn Kloosterman, Richard GF Visser, and Chris Maliepaard. Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC bioinformatics*, 17(5):363–373, 2016.

[170] Darren CJ Wong and José Tomás Matus. Constructing integrated networks for identifying new secondary metabolic pathway regulators in grapevine: recent applications and future opportunities. *Frontiers in plant science*, 8:505, 2017.

[171] Jing Jiang, Fei Xing, Chunyu Wang, Xiangxiang Zeng, and Quan Zou. Investigation and development of maize fused network analysis with multi-omics. *Plant Physiology and Biochemistry*, 141:380–387, 2019.

[172] Liang Jiang, Takuya Yoshida, Sofia Stiegert, Yue Jing, Saleh Alseekh, Michael Lenhard, Francisco Pérez-Alfocea, and Alisdair R Fernie. Multi-omics approach reveals the contribution of klu to leaf longevity and drought tolerance. *Plant Physiology*, 2020.

[173] Nam D Nguyen, Ian K Blaby, and Daifeng Wang. Maninetcluster: a novel manifold learning approach to reveal the functional links between gene networks. *BMC genomics*, 20(12):1–14, 2019.

[174] Kitiporn Plaimas, Jan-Phillip Mallm, Marcus Oswald, Fabian Svara, Victor Sourjik, Roland Eils, and Rainer König. Machine learning based analyses on metabolic networks supports high-throughput knockout screens. *BMC systems biology*, 2(1):1–11, 2008.

[175] Balázs Szappanos, Károly Kovács, Béla Szamecz, Frantisek Honti, Michael Costanzo, Anastasia Baryshnikova, Gabriel Gelius-Dietrich, Martin J Lercher, Márk Jelasity, Chad L Myers, et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature genetics*, 43(7):656–662, 2011.

[176] Minseung Kim, Navneet Rai, Violeta Zorraquino, and Ilias Tagkopoulos. Multi-omics integration accurately predicts cellular state in unexplored conditions for escherichia coli. *Nature communications*, 7(1):1–12, 2016.

[177] Peter DePaulo. Sample size for qualitative research. *Quirks Marketing Research Review*, 1202, 2000.

[178] Marianna Fasoli, Chandra L Richter, Sara Zenoni, Edoardo Bertini, Nicola Vitulo, Silvia Dal Santo, Nick Dokoozlian, Mario Pezzotti, and Giovanni Battista Tornielli. Timing and order of the molecular events marking the onset of berry ripening in grapevine. *Plant physiology*, 178 (3):1187–1206, 2018.

[179] Ratklao Siriwach, Fumio Matsuda, Kentaro Yano, and Masami Yokota Hirai. Drought stress responses in context-specific genome-scale metabolic models of arabidopsis thaliana. *Metabolites*, 10(4):159, 2020.

[180] Olivier Jaillon, Jean-Marc Aury, Benjamin Noel, Alberto Policriti, Christian Clepet, Alberto Casagrande, Nathalie Choisne, Sébastien Aubourg, Nicola Vitulo, Claire Jubin, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *nature*, 449(7161):463, 2007.

[181] Inaki Inza, Borja Calvo, Rubén Armañanzas, Endika Bengoetxea, Pedro Larranaga, and José A Lozano. Machine learning: an indispensable tool in bioinformatics. In *Bioinformatics methods in clinical research*, pages 25–48. Springer, 2010.

[182] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

[183] Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. mlr: Machine learning in r. *Journal of Machine Learning Research*, 17(170):1–5, 2016. URL http://jmlr.org/papers/v17/15-066.html.

[184] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

[185] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[186] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[187] Spencer Aiello, Eric Eckstrand, Anqi Fu, Mark Landry, and Patrick Aboyoun. Machine learning with r and h2o. *H2O booklet*, 550, 2016.

[188] Trevor Hastie and Junyang Qian. Glmnet vignette. *Retrieved June*, 9(2016):1–30, 2014.

[189] Eugene Lin and Hsien-Yuan Lane. Machine learning and systems genomics approaches for multi-omics data. *Biomarker research*, 5(1):1–6, 2017.

[190] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[191] Parminder S Reel, Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, page 107739, 2021.

[192] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[193] H. Abdi and D. Valentin. Multiple factor analysis (mfa). 2009.

[194] Charu Sudan, Shiva Prakash, Prasanna Bhomkar, Shalu Jain, and Neera Bhalla-Sarin. Ubiquitous presence of β-glucuronidase (gus) in plants and its regulation in some model plants. *Planta*, 224(4):853–864, 2006.

[195] Anton Pieter Nel. Tannins and anthocyanins: From their origin to wine analysis-a review. *South African Journal of Enology and Viticulture*, 39(1):1–20, 2018.

[196] Alessandra Ferrandino and Silvia Guidoni. Anthocyanins, flavonols and hydroxycinnamates: an attempt to use them to discriminate vitis vinifera l. cv 'barbera'clones. *European Food Research and Technology*, 230(3):417–427, 2010.

[197] Pengbao Shi, Bing Li, Haiju Chen, Changzheng Song, Jiangfei Meng, Zhumei Xi, and Zhenwen Zhang. Iron supply affects anthocyanin content and related gene expression in berries of vitis vinifera cv. cabernet sauvignon. *Molecules*, 22(2):283, 2017.

[198] Silvia Dal Santo, Giovanni Battista Tornielli, Sara Zenoni, Marianna Fasoli, Lorenzo Farina, Andrea Anesi, Flavia Guzzo, Massimo Delledonne, and Mario Pezzotti. The plasticity of the grapevine berry transcriptome. *Genome biology*, 14(6):1–18, 2013.

[199] Francisco Goes da Silva, Alberto Iandolino, Fadi Al-Kayal, Marlene C Bohlmann, Mary Ann Cushman, Hyunju Lim, Ali Ergul, Rubi Figueroa, Elif K Kabuloglu, Craig Osborne, et al. Characterizing the grape transcriptome. analysis of expressed sequence tags from multiple vitis species and development of a compendium of gene expression during berry development. *Plant physiology*, 139(2):574–597, 2005.

[200] Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.

[201] Pingzhou Du, Manoj Kumar, Yuan Yao, Qiaoli Xie, Jinyan Wang, Baolong Zhang, Siming Gan, Yuqi Wang, and Ai-Min Wu. Genome-wide analysis of the tpx2 family proteins in eucalyptus grandis. *BMC genomics*, 17(1):1–11, 2016.

[202] Changzheng Xu, Feng Luo, and Frank Hochholdinger. Lob domain proteins: beyond lateral organ boundaries. *Trends in plant science*, 21(2):159–167, 2016.

[203] Sara Zenoni, Silvia Dal Santo, Giovanni B Tornielli, Erica D'Incà, Ilaria Filippetti, Chiara Pastore, Gianluca Allegro, Oriana Silvestroni, Vania Lanari, Antonino Pisciotta, et al. Transcriptional responses to pre-flowering leaf defoliation in grapevine berry from different growing sites, years, and genotypes. *Frontiers in plant science*, 8:630, 2017.

[204] Victor Quesada, Raquel Sarmiento-Manus, Rebeca Gonzalez-Bayon, Andrea Hricova, María Rosa Ponce, and José Luis Micol. Porphobilinogen deaminase deficiency alters vegetative and reproductive development and causes lesions in arabidopsis. *PLoS One*, 8(1):e53378, 2013.

[205] Yihe Yu, Weirong Xu, Shengyi Wang, Yan Xu, Hui'e Li, Yuejin Wang, and Shuxiu Li. Vprfp1, a novel c4c4-type ring finger protein gene from chinese wild vitis pseudoreticulata, functions as a transcriptional activator in defence response of grapevine. *Journal of experimental botany*, 62 (15):5671–5682, 2011.

[206] Nunzio D'Agostino, Martina Buonanno, Joëlle Ayoub, Amalia Barone, Simona Maria Monti, and Maria Manuela Rigano. Identification of non-specific lipid transfer protein gene family members in solanum lycopersicum and insights into the features of sola l 3 protein. *Scientific reports*, 9(1):1–16, 2019.

[207] Fan-Hsuan Yang, Lisa W DeVetter, Bernadine C Strik, and David R Bryla. Stomatal functioning and its influence on fruit calcium accumulation in northern highbush blueberry. *HortScience*, 55 (1):96–102, 2020.

[208] Amy R Rinaldo, Erika Cavallini, Yong Jia, Sarah MA Moss, Debra AJ McDavid, Lauren C Hooper, Simon P Robinson, Giovanni B Tornielli, Sara Zenoni, Christopher M Ford, et al. A grapevine anthocyanin acyltransferase, transcriptionally regulated by vvmyba, can produce most acylated anthocyanins present in grape skins. *Plant physiology*, 169(3):1897–1916, 2015.

[209] Gert Sclep, Joke Allemeersch, Robin Liechti, Björn De Meyer, Jim Beynon, Rishikesh Bhalerao, Yves Moreau, Wilfried Nietfeld, Jean-Pierre Renou, Philippe Reymond, et al. Catma, a comprehensive genome-scale resource for silencing and transcript profiling of arabidopsis genes. *BMC bioinformatics*, 8(1):1–13, 2007.

[210] Tie Liu, Adam D Longhurst, Franklin Talavera-Rauh, Samuel A Hokin, and M Kathryn Barton. The arabidopsis transcription factor abig1 relays aba signaled growth inhibition and drought induced senescence. *Elife*, 5:e13768, 2016.

[211] Monica De Caroli, Elisa Manno, Gabriella Piro, and Marcello S Lenucci. Ride to cell wall: Arabidopsis xth11, xth29 and xth33 exhibit different secretion pathways and responses to heat and drought stress. *The Plant Journal*, 107(2):448, 2021.

[212] Amal Harb, Arjun Krishnan, Madana MR Ambavaram, and Andy Pereira. Molecular and physiological analysis of drought stress in arabidopsis reveals early responses leading to acclimation in plant growth. *Plant physiology*, 154(3):1254–1271, 2010.

[213] Pan Li, Yan-Jie Li, Feng-Ju Zhang, Gui-Zhi Zhang, Xiao-Yi Jiang, Hui-Min Yu, and Bing-Kai Hou. The arabidopsis udp-glycosyltransferases ugt79b2 and ugt79b3, contribute to cold, salt and drought stress tolerance via modulating anthocyanin accumulation. *The Plant Journal*, 89 (1):85–103, 2017.

[214] Ling Pan, Zhongfu Yang, Jianping Wang, Pengxi Wang, Xiao Ma, Meiliang Zhou, Ji Li, Nie Gang, Guangyan Feng, Junming Zhao, et al. Comparative proteomic analyses reveal the proteome response to short-term drought in italian ryegrass (lolium multiflorum). *PloS one*, 12 (9):e0184289, 2017.

[215] Chulhyun Ahn, Uhnmee Park, and Phun Bum Park. Increased salt and drought tolerance by d-ononitol production in transgenic arabidopsis thaliana. *Biochemical and Biophysical Research Communications*, 415(4):669–674, 2011.

[216] Qianqian Wang, Qianli Guo, Yuanyuan Guo, Jieshu Yang, Min Wang, Xiaoke Duan, Jiayu Niu, Shuai Liu, Jianzhen Zhang, Yanke Lu, et al. Arabidopsis subtilase sasp is involved in the regulation of aba signaling and drought tolerance by interacting with open stomata 1. *Journal of experimental botany*, 69(18):4403–4417, 2018.

[217] Neha Vaid, Prashant Kumar Pandey, and Narendra Tuteja. Genome-wide analysis of lectin receptor-like kinase family from arabidopsis and rice. *Plant molecular biology*, 80(4):365–388, 2012.

[218] Hideki Sakamoto, Kyonoshin Maruyama, Yoh Sakuma, Tetsuo Meshi, Masaki Iwabuchi, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. Arabidopsis cys2/his2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. *Plant physiology*, 136(1):2734–2746, 2004.

[219] Linchuan Fang, Lingye Su, Xiaoming Sun, Xinbo Li, Mengxiang Sun, Sospeter Karanja Karungo, Shuang Fang, Jinfang Chu, Shaohua Li, and Haiping Xin. Expression of vitis amurensis nac26 in arabidopsis enhances drought tolerance by modulating jasmonic acid synthesis. *Journal of experimental botany*, 67(9):2829–2845, 2016.

[220] Carsten Müssig, Christian Biesgen, Janina Lisso, Ursula Uwer, Elmar W Weiler, and Thomas Altmann. A novel stress-inducible 12-oxophytodienoate reductase from arabidopsis thaliana provides a potential link between brassinosteroid-action and jasmonic-acid synthesis. *Journal of plant physiology*, 157(2):143–152, 2000.

[221] Hong Qiao, Katherine N Chang, Junshi Yazaki, and Joseph R Ecker. Interplay between ethylene, etp1/etp2 f-box proteins, and degradation of ein2 triggers ethylene responses in arabidopsis. *Genes & development*, 23(4):512–521, 2009.

[222] Jinrui Shi, Jeffrey E Habben, Rayeann L Archibald, Bruce J Drummond, Mark A Chamberlin, Robert W Williams, H Renee Lafitte, and Ben P Weers. Overexpression of argos genes modifies plant sensitivity to ethylene, leading to improved drought tolerance in both arabidopsis and maize. *Plant physiology*, 169(1):266–282, 2015.

[223] Tong Li, Ying Huang, Ahmed Khadr, Ya-Hui Wang, Zhi-Sheng Xu, and Ai-Sheng Xiong. Dcdreb1a, a dreb-binding transcription factor from daucus carota, enhances drought tolerance in transgenic arabidopsis thaliana and modulates lignin levels by regulating lignin-biosynthesis-related genes. *Environmental and Experimental Botany*, 169:103896, 2020.

[224] Marcel V Pires, Adilson A Pereira Júnior, David B Medeiros, Danilo M Daloso, Phuong Anh Pham, Kallyne A Barros, Martin KM Engqvist, Alexandra Florian, Ina Krahnert, Veronica G

Maurino, et al. The influence of alternative pathways of respiration that utilize branched-chain amino acids following water shortage in arabidopsis. *Plant, Cell & Environment*, 39(6): 1304–1319, 2016.

[225] Tao Wei, Kejun Deng, Hongbin Wang, Lipeng Zhang, Chunguo Wang, Wenqin Song, Yong Zhang, and Chengbin Chen. Comparative transcriptome analyses reveal potential mechanisms of enhanced drought tolerance in transgenic salvia miltiorrhiza plants expressing atdreb1a from arabidopsis. *International journal of molecular sciences*, 19(3):827, 2018.

[226] Andrés J Cortés and Matthew W Blair. Genotyping by sequencing and genome–environment associations in wild common bean predict widespread divergent adaptation to drought. *Frontiers in Plant Science*, 9:128, 2018.

[227] John P Moore, Mäite Vicré-Gibouin, Jill M Farrant, and Azeddine Driouich. Adaptations of higher plant cell walls to water loss: drought vs desiccation. *Physiologia plantarum*, 134(2): 237–245, 2008.

[228] Xinbo Wang, Yanhua Xu, Jingjing Li, Yongzhe Ren, Zhiqiang Wang, Zeyu Xin, and Tongbao Lin. Identification of two novel wheat drought tolerance-related proteins by comparative proteomic analysis combined with virus-induced gene silencing. *International journal of molecular sciences*, 19(12):4020, 2018.

[229] Chun-Long Li, Mei Wang, Xiao-Yan Ma, and Wei Zhang. Nrga1, a putative mitochondrial pyruvate carrier, mediates aba regulation of guard cell ion channels and drought stress responses in arabidopsis. *Molecular plant*, 7(10):1508–1521, 2014.

[230] Nina Lehtimäki, Minna Lintala, Yagut Allahverdiyeva, Eva-Mari Aro, and Paula Mulo. Drought stress-induced upregulation of components involved in ferredoxin-dependent cyclic electron transfer. *Journal of plant physiology*, 167(12):1018–1022, 2010.

[231] Olaf Barth, Sebastian Vogt, Ria Uhlemann, Wiebke Zschiesche, and Klaus Humbeck. Stress induced and nuclear localized hipp26 from arabidopsis thaliana interacts via its heavy metal associated domain with the drought stress related zinc finger transcription factor athb29. *Plant molecular biology*, 69(1):213–226, 2009.

[232] Sandra S Scholz, Michael Reichelt, Jyothilakshmi Vadassery, and Axel Mithöfer. Calmodulin-like protein cml37 is a positive regulator of aba during drought stress in arabidopsis. *Plant signaling & behavior*, 10(6):e1011951, 2015.

A

SUPPLEMENTARY FIGURES

A.1 CASE STUDY I

Metadata include some useful information, such as the berry weight, *"Berry Weight (g / berry)"*, the concentration of malic acid, *"Malic Acid (mg / l)"*, the accumulation of RS, *"RS (g /100 ml ± 5%)"*, the time point, *"Time Point"*, the days after veraison in which the samples were extracted, *"Days after veraison"*, and the variety and vintage they belonged.

In supplementary figure S1A, it is clear that the samples were divided by two types of variety: *Cabernet Sauvignon* and *Pinot Noir*. The total number of samples was 73, 40 for *Cabernet Sauvignon* and 33 for *Pinot Noir*. Regarding the vintage variable, 23 samples were taken in the first year (2012) and 25 samples were collected in the two consecutive years (2013-2014, see supplementary figure S1B). In supplementary figure S1C), the samples are divided by our selected outcome, 24 samples in pre-veraison and 49 samples in post-veraison stage, which shows dataset imbalance.

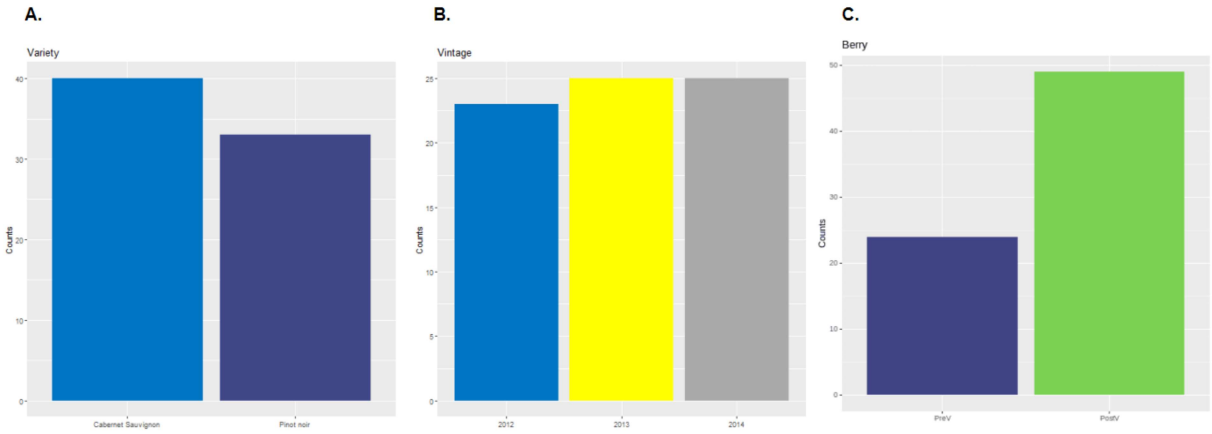


Figure S1: *(A)Variety, (B)Vintage and (C)Berry*. (A)The samples were divided as 40 samples for *Cabernet Sauvignon* (light blue) and 33 for *Pinot Noir* (dark blue). (B) In the three consecutive years, samples were extracted as follows: 23 samples in 2012 (light blue), 25 samples in 2013 (yellow) and 25 samples in 2014 (grey). (C) For our selected outcome, the division is 24 in pre-veraison (dark blue) and 49 in post-veraison stage (green).

Supplementary figure S2 shows a summary of the berry weight by grape variety. The berry weight had a mean of 1.238g/ berry, however if we looked at the berry weight depending on the variety of grape, we could see that the berry weight in the *Pinot Noir* cultivar was slightly bigger than the *Cabernet Sauvignon* grape. This proposed an interesting difference between the both cultivar types and how they developed through time.

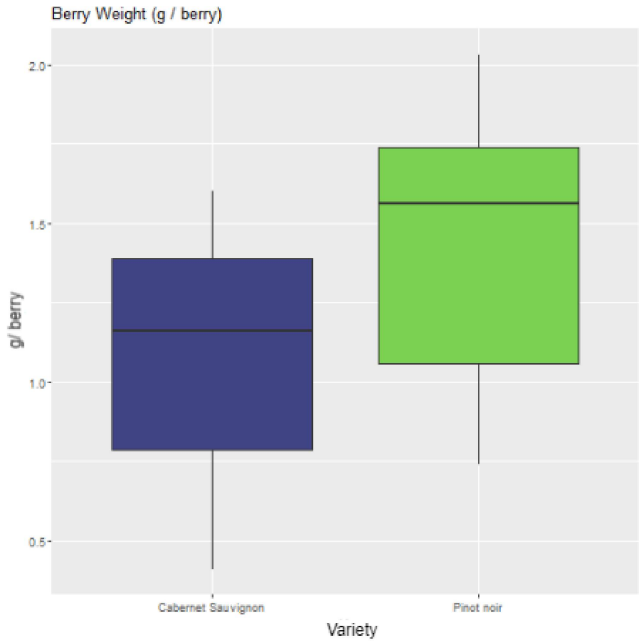


Figure S2: *Exploratory Analysis.* Boxplot of berry weight by grape variety.

Supplementary figure S3 replicated the plots made in the original article that explained how the different varieties behaved in terms of berry weight, concentration of malic acid and accumulation of RS in the different days after veraison.

As we can see in figure S3, both varieties have similar patterns of development in all three variables, however some discrepancies can be identified. It is the case of berry weight, as mentioned above, and RS that accumulates faster in the *Pinot Noir* grapes than in *Cabernet Sauvignon*, resulting in a shorter development time and early harvest. In addition, malic acid initial concentration is also higher in *Pinot Noir* berries than *Cabernet Sauvignon*, yet when the concentration matches both varieties the degradation is also faster.

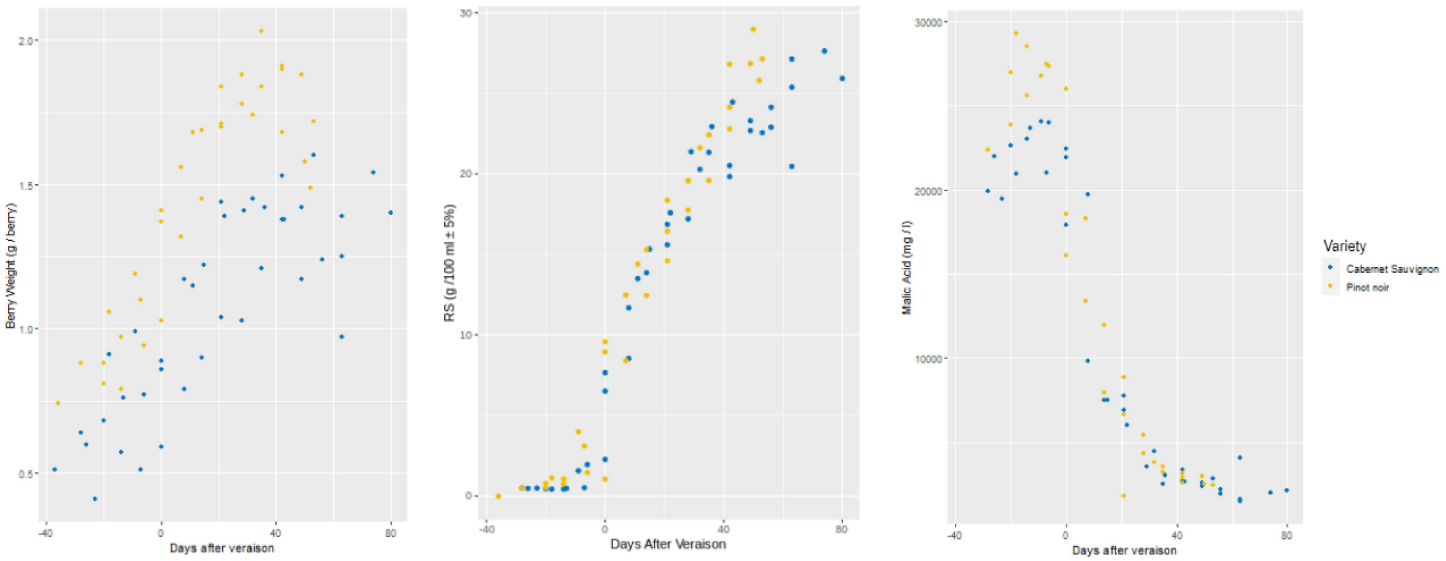


Figure S3: *Exploratory Analysis.* Progression of grape berry ripening. Grape berry development is shown by berry weight in the first plot, reducing sugar accumulation in the second plot, and malic acid (MA) accumulation in the third plot, from fruit set to harvest.

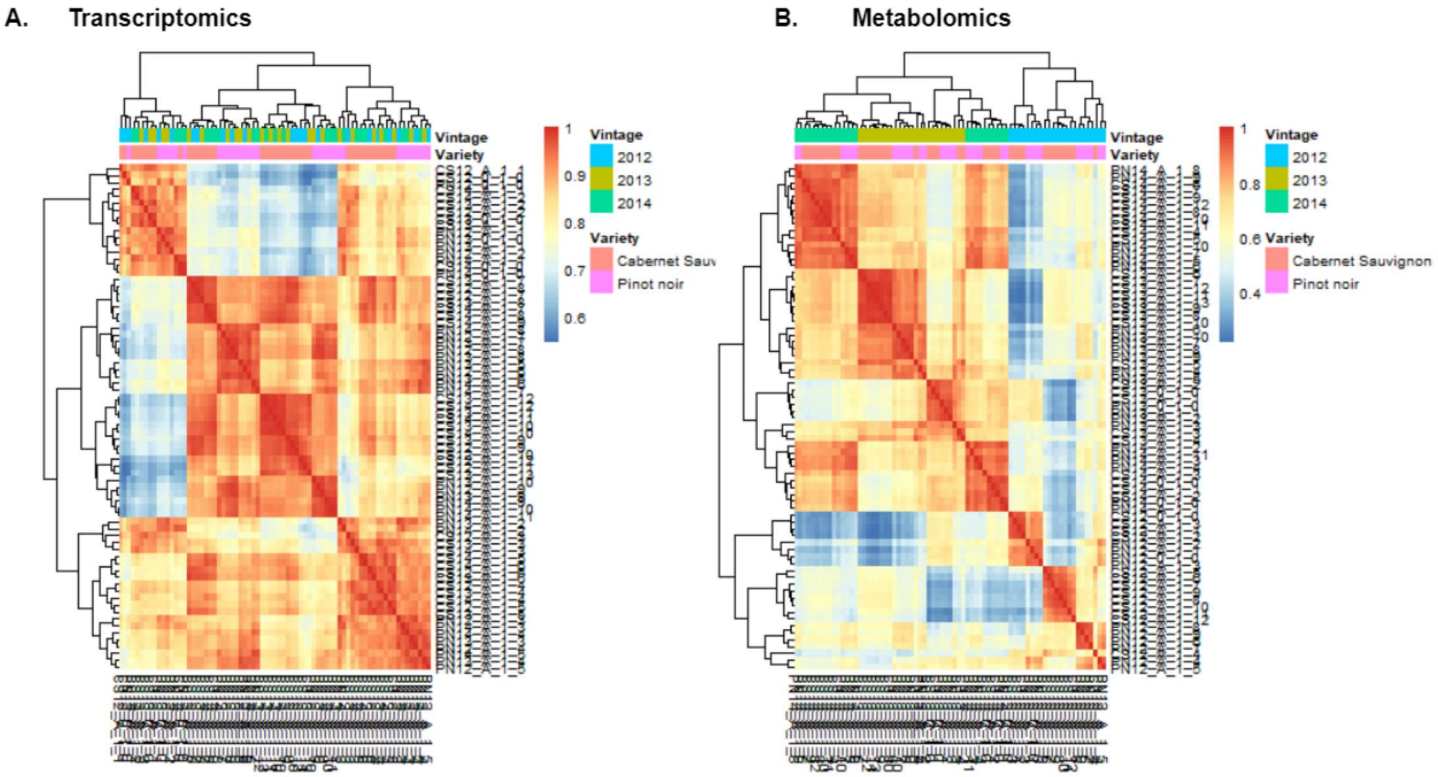


Figure S4: *Exploratory Analysis*. Heatmap of the (A) transcriptomics dataset and (B) metabolomics dataset, regarding vintage and variety.

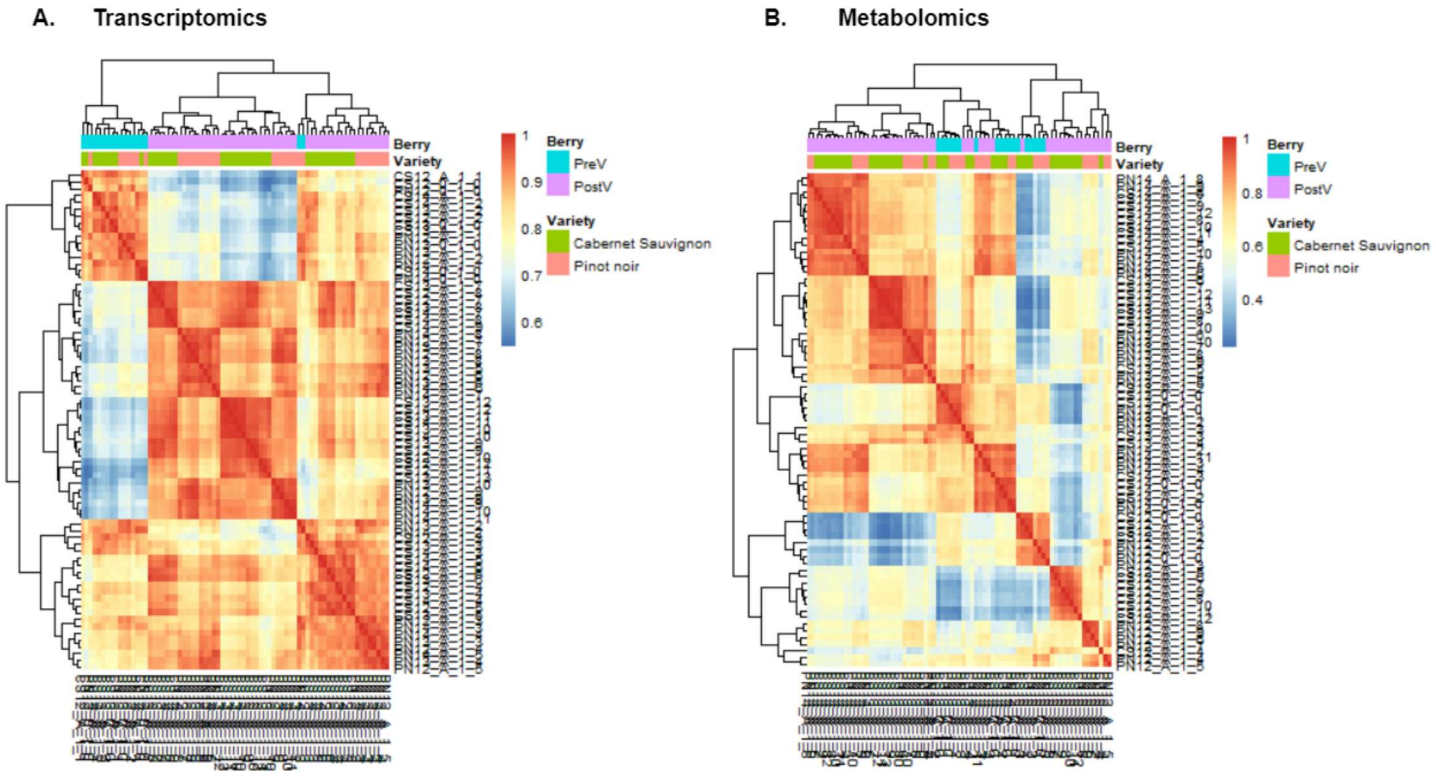


Figure S5: *Exploratory Analysis*. Heatmap of the (A) transcriptomics dataset and (B) metabolomics dataset, regarding berry development stage (PreV and PostV and variety.

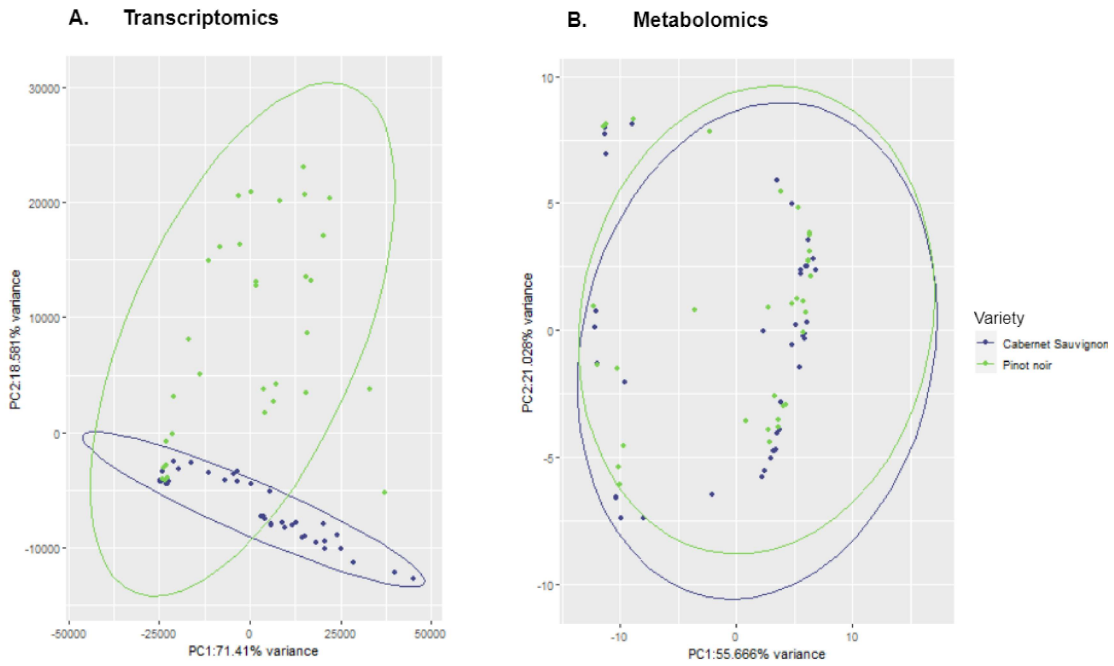


Figure S6: *Exploratory Analysis*. PCA analysis of (A) transcriptomics and (B) metabolomics regarding Variety.

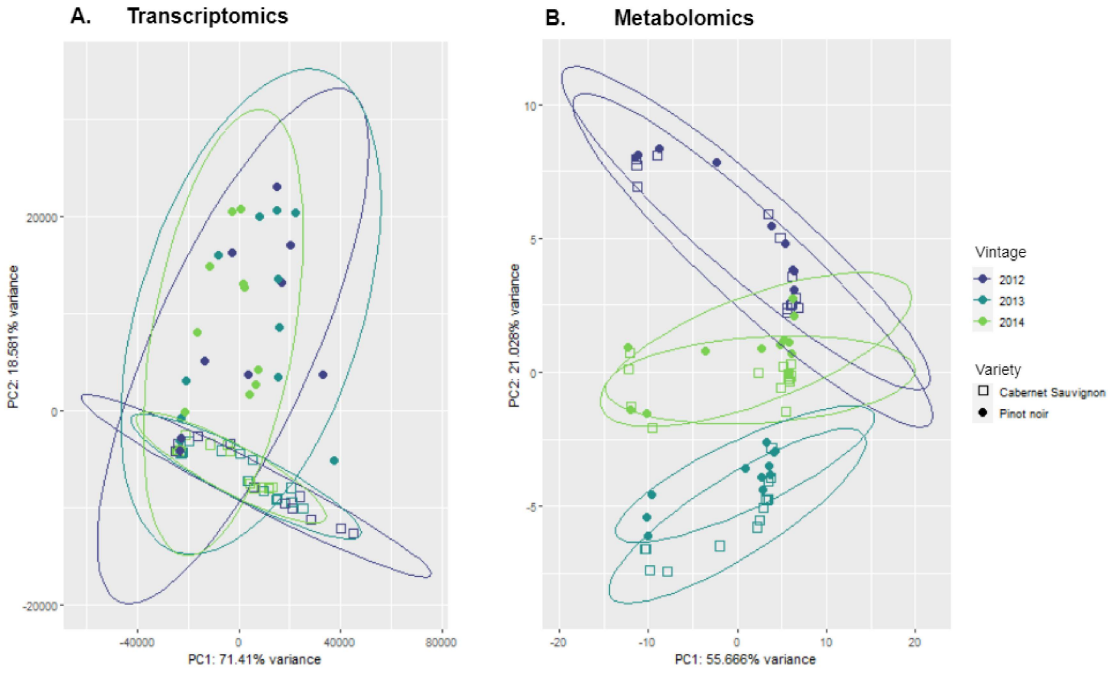


Figure S7: *Exploratory Analysis*. PCA analysis of (A) transcriptomics and (B) metabolomics regarding Variety and Vintage.

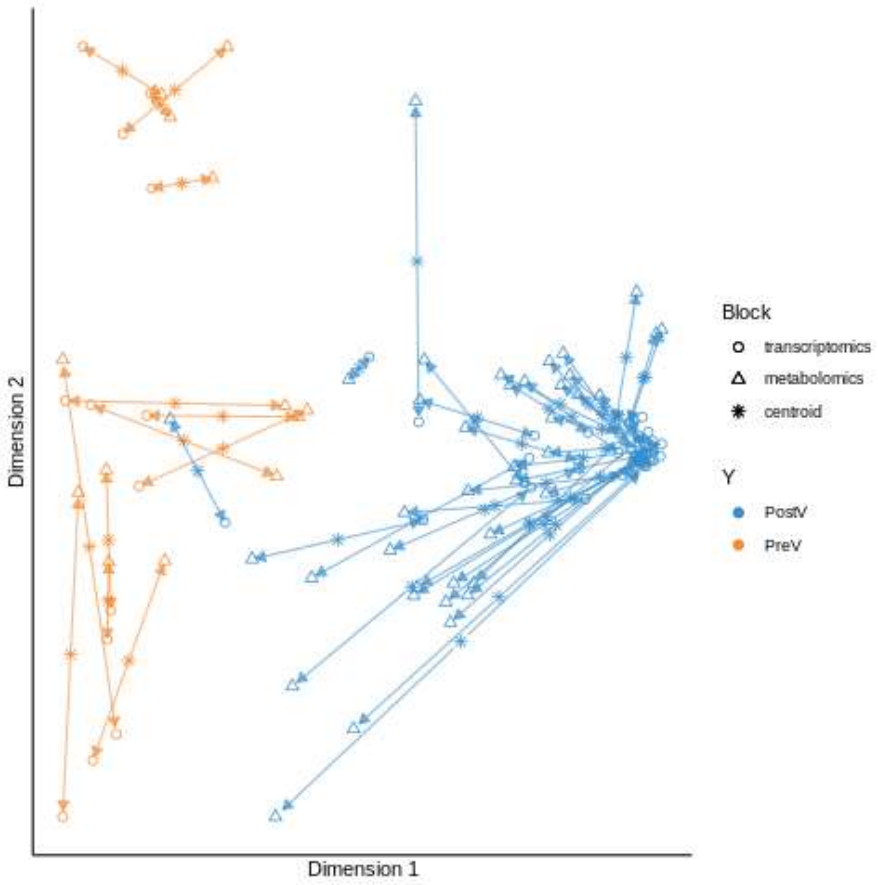


Figure S8: *Multiomics Integration*. Arrow plot in which the start of the arrow indicates the centroid among the two datasets for a given sample and the tip of the arrows indicates the location of that sample in each block.

Figure S9: *Multiomics Integration*. Relevance network plot in which we visualize the correlation regarding the two different types of variables built on the similarity matrix.

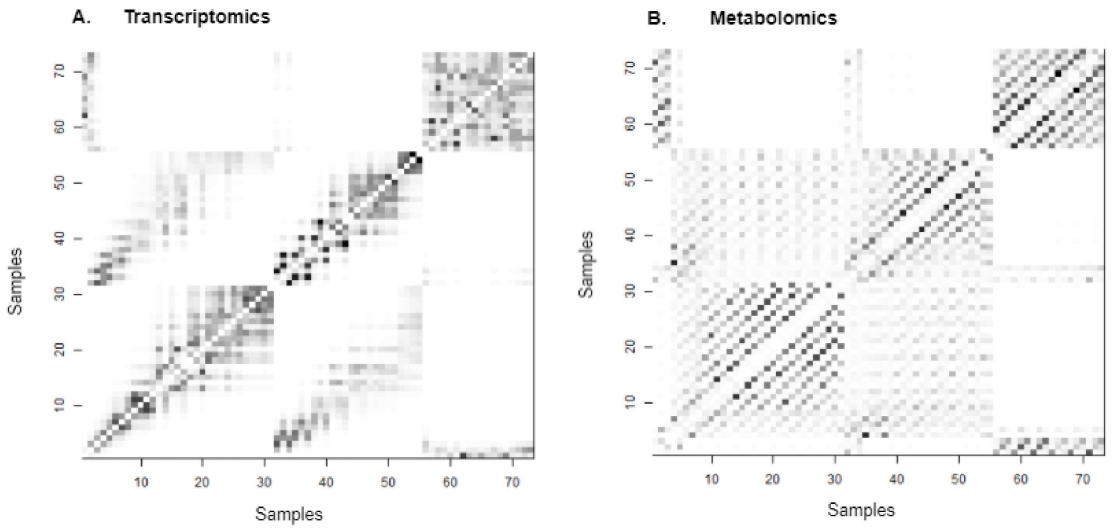


Figure S10: *Multiomics Integration*. Transformation-based integration. Display of the two similarity graphs for each omics dataset that have complementary information about clusters.(A) Transcriptomics dataset. (B) Metabolomics dataset.

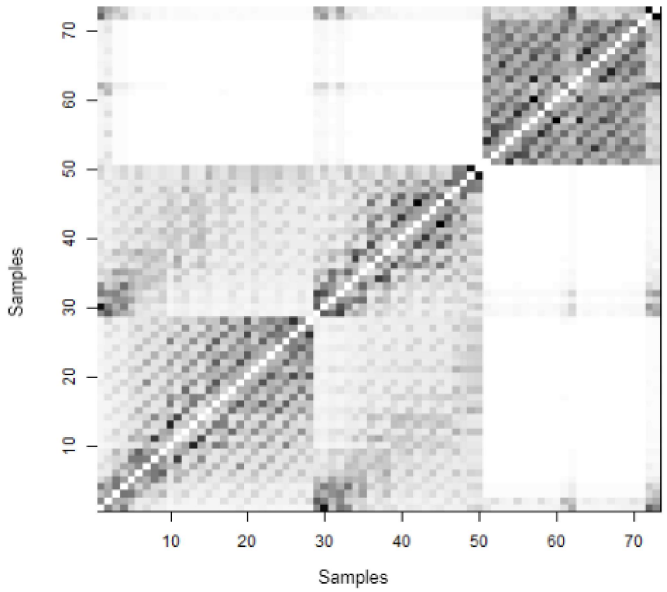


Figure S11: *Multiomics Integration*. Transformation-based integration. Display of the fusion of the two similarity graphs.

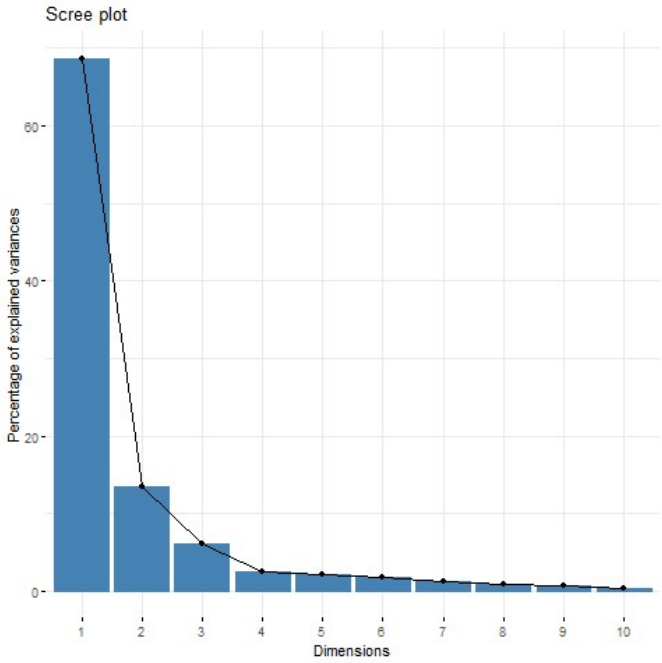


Figure S12: *Multiomics Analysis Integration. Unsupervised Learning.* Scree plot of the eigen values obtained by the MFA model.

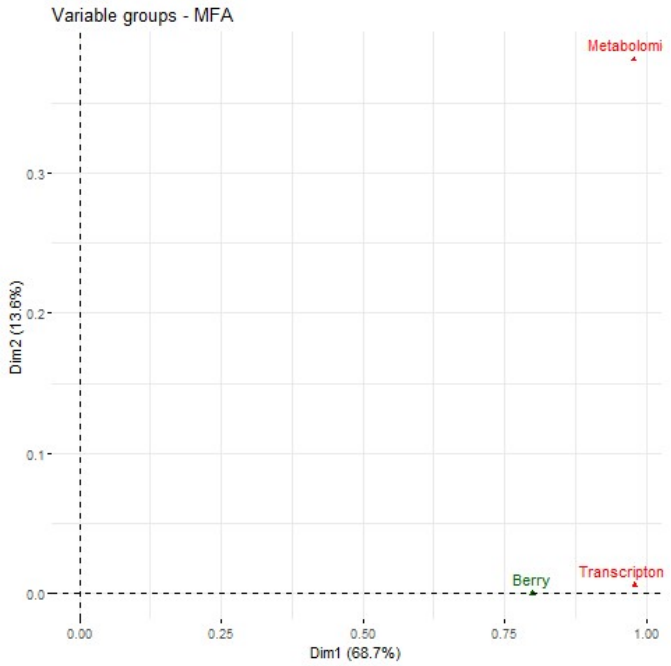


Figure S13: *Multiomics Analysis Integration. Unsupervised Learning.* Plot of the group variables using the MFA model. It illustrates the correlation between the groups and dimensions. We can see that the green groups indicate the supplementary groups of variables and that the red groups represent the active groups of variables. Therefore, our active groups correspond to the metabolomics and transcriptomics dataset. Additionally, we can see that both datasets contribute similarly to the first dimension. Concerning the second dimension, the metabolomics dataset had higher coordinates indicating a highest contribution to the second dimension. The Berry variable contributes only for the dimension 1.

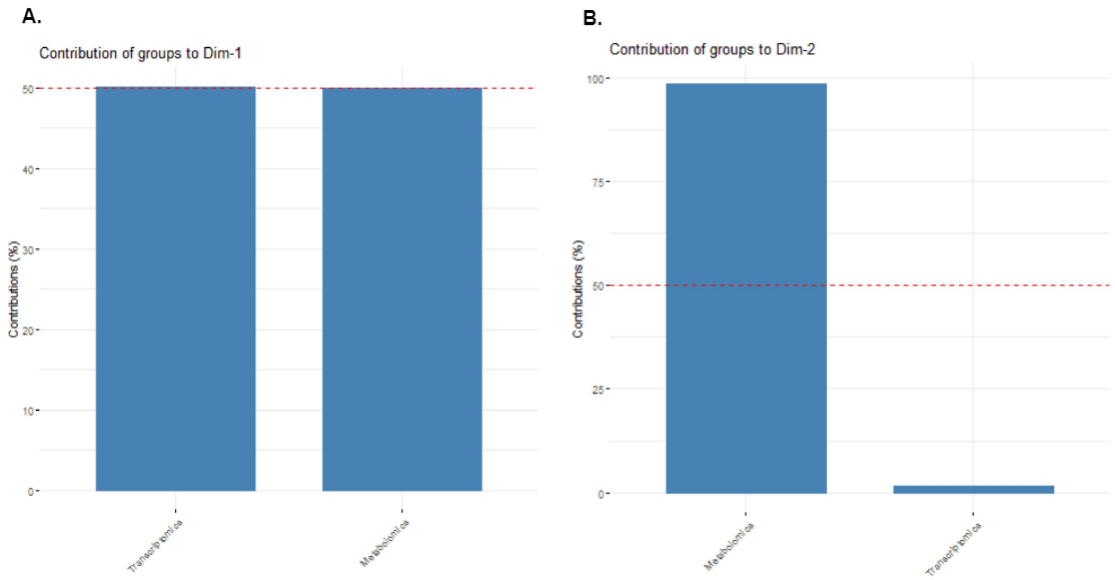


Figure S14: *Multiomics Analysis Integration. Unsupervised Learning.* Plot of the contribution of the different groups regarding the (A) Dimension 1 and (B) Dimension 2.

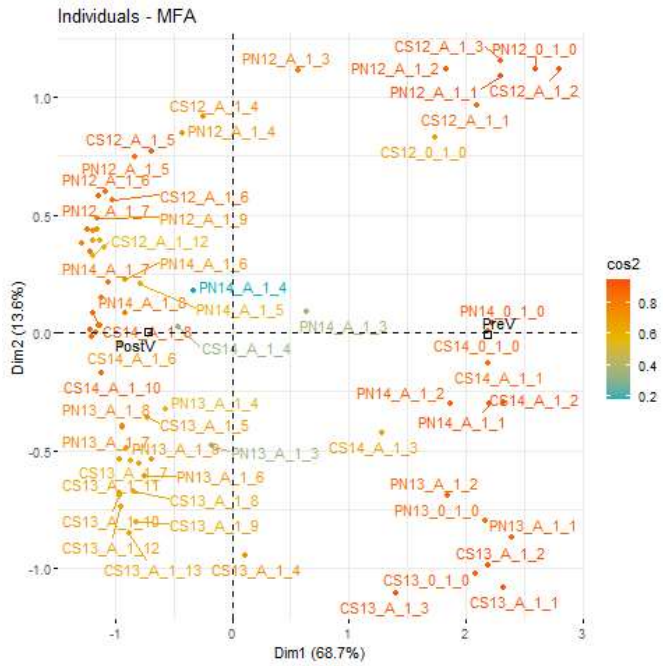


Figure S15: *Multiomics Integration.* Unsupervised Learning - Concatenation-Based Integration. Plot of the individuals colored by their cos2 value.

A.2 CASE STUDY II

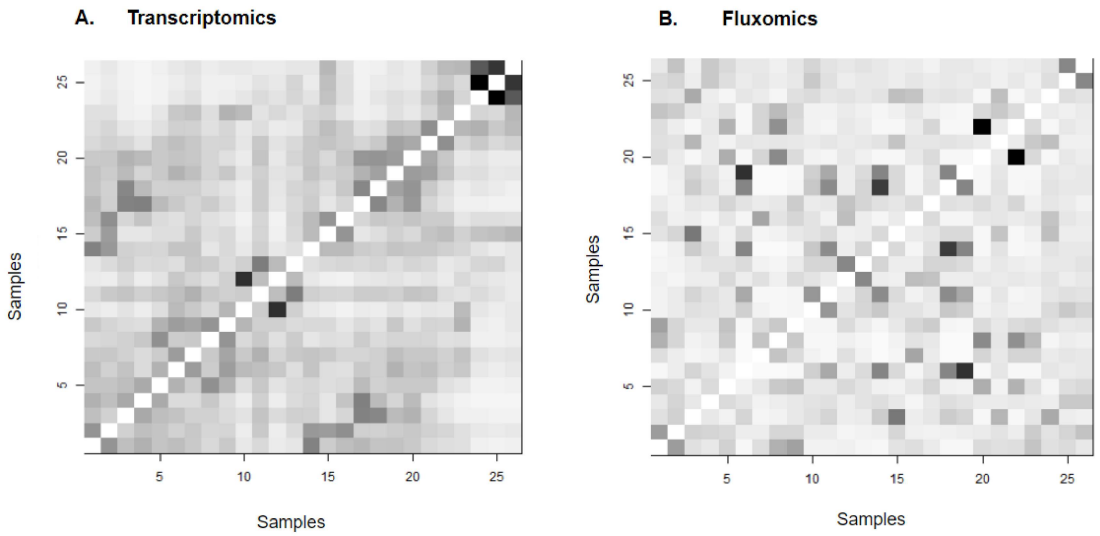


Figure S1: *Multiomics Integration*. Transformation-based integration. Display of the two similarity graphs for each omics dataset that have complementary information about clusters.(A) Transcriptomics dataset. (B) Fluxomics dataset.

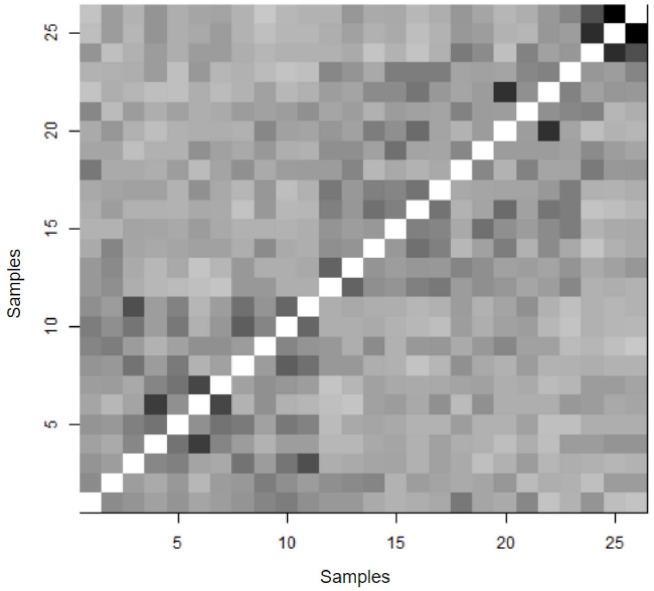


Figure S2: *Multiomics Integration*. Transformation-based integration. Display of the fusion of the two similarity graphs.

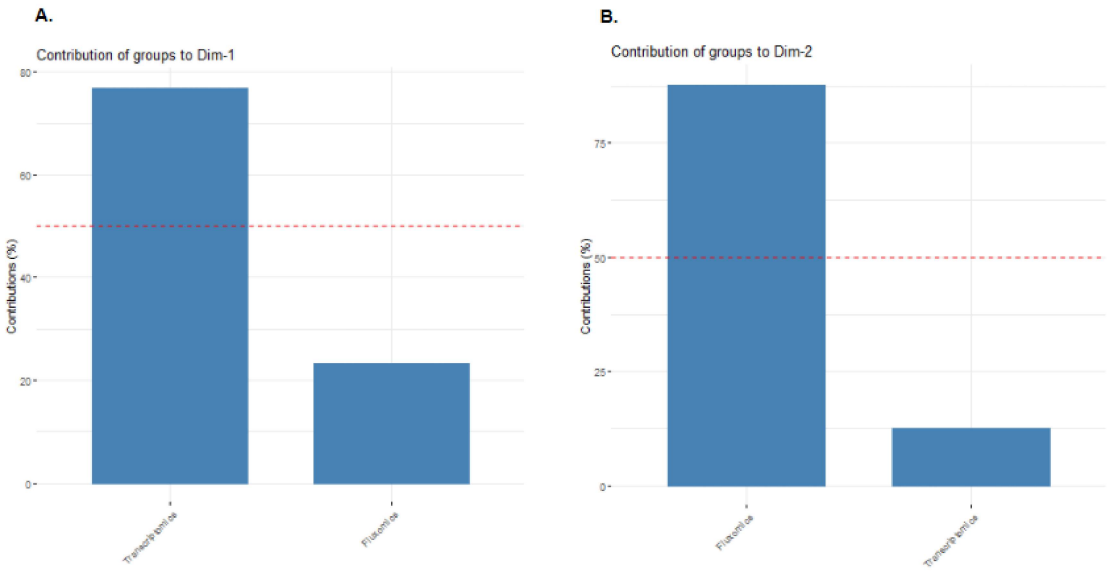


Figure S3: *Multiomics Integration Analysis with Unsupervised Learning.* Plot of the contribution of the different groups regarding the (A) Dimension 1 and (B) Dimension 2. The transcriptomics dataset contributes more to dimension 1 and the fluxomics dataset contributes more for dimension 2.

B

SUPPLEMENTARY TABLES

B.1 CASE STUDY I

Table S1: Individual Omics Analysis. Most relevant features obtained from the SVM model for the transcriptomics dataset.

Transcript	UniProtKB	Annotation	Importance
VIT_15s0046g00230	F6I6F6	Lateral organ boundaries protein 1	1.0000
VIT_13s0019g03040	F6HNN8	Indole-3-acetate beta-glucosyltransferase	1.0000
VIT_00s0759g00010	D7ST07	Porphobilinogen deaminase, chloroplast precursor	1.0000
VIT_13s0064g01030	D7T2Z6	Zinc finger (C3HC4-type ring finger)BIG BROTHER	1.0000
VIT_03s0063g01560	F6HQG7	CYP82C1p	1.0000
VIT_04s0044g01870	F6I0B3	Auxin efflux carrier	1.0000
VIT_01s0011g02330	D7T9F1	Unknown protein	1.0000
VIT_14s0060g01090	D7UA38	LNG1 (LONGIFOLIA1)	1.0000
VIT_08s0007g02170	A5C287	Yippee	1.0000
VIT_02s0154g00350	D7TN36	L-lactate dehydrogenase A	1.0000
VIT_13s0064g00890	F6HB61	Cellulose synthase CESA3	1.0000
VIT_12s0028g02160	F6H4Z7	Ribulose bisphosphate carboxylase	1.0000
VIT_14s0083g01100	F6GVV2	Alpha-1,4-glucan-protein synthase 1	1.0000
VIT_11s0016g05840	F6HHB3	Protease inhibitor/seed storage/lipid transfer protein (LTP)	1.0000
VIT_12s0028g03100	F6H5G9	GPRI1 (GOLDEN2 1)	1.0000
VIT_03s0038g02470	F6I101	Nickel ion transporter	1.0000
VIT_01s0026g00330	D7TNP7	NHL repeat-containing protein	1.0000
VIT_09s0002g04080	F6HYD4	IAA9	1.0000
VIT_05s0020g04880	D7T7A2	Seed specific protein Bn15D14A	1.0000
VIT_18s0001g15520	E0CQN6	Leaf senescence protein	1.0000

Table S2: Individual Omics Analysis. Most relevant features obtained from the SVM model for the metabolomics dataset.

Metabolites	Importance
petunidin-3-glucoside	1.0000
fructose	0.9980
tartaric acid	0.9919
malic acid	0.9898
malvidin-3-O-glucoside	0.9898
glucose	0.9878
delphinidin 3-O-glucoside	0.9817
peonidin-3-glucoside	0.9797
sucrose	0.9715
citric acid	0.9695
stearic acid	0.9634
1,2-anhydro-myo-inositol NIST	0.9329
quercetin-3-glucuronide	0.9329
threonic acid	0.9329
cyanidin 3-glucoside	0.9228
myo-inositol	0.9146
myricetin-3-glucoside	0.8984
palmitic acid	0.8882
benzenemethanol	0.8841
benzoic acid	0.8841

Table S3: Individual Omics Analysis. Most relevant features obtained from the RF model for the transcriptomics dataset.

Transcripts	UniProtKB	Annotation	Importance
VIT_17s0000g08900	F6GSJ2	LRR receptor-like kinase 2	0.3642
VIT_15s0046g00230	F6I6F6	Lateral organ boundaries protein 1	0.3200
VIT_01s0244g00070	D7TYT5	Unknown protein	0.2838
VIT_14s0060g01090	D7UA38	LNG1 (LONGIFOLIA1)	0.2783
VIT_12s0028g03100	F6H5G9	GPRI1 (GOLDEN2 1)	0.2691
VIT_14s0083g01100	F6GVV2	Alpha-1,4-glucan-protein synthase 1	0.2634
VIT_03s0063g01560	F6HQG7	CYP82C1p	0.2567
VIT_01s0026g01780	F6HPE9	Leucine-rich repeat transmembrane	0.2545
VIT_05s0020g04880	D7T7A2	Seed specific protein Bn15D14A	0.2411
VIT_00s0759g00010	D7ST07	Porphobilinogen deaminase, chloroplast precursor	0.2397
VIT_13s0064g01030	D7T2Z6	Zinc finger (C3HC4-type ring finger)BIG BROTHER	0.2391
VIT_13s0019g00240	D7TLZ6	Glycosyltransferase family 14 Beta-1-3-galactosyl-O-glycosyl-glycoprotein	0.2202
VIT_13s0064g00890	F6HB61	Cellulose synthase CESA3	0.2197
VIT_08s0007g04510	F6HKE6	RPG related protein 1 RR1	0.2165
VIT_13s0019g03040	F6HNN8	Indole-3-acetate beta-glucosyltransferase	0.2083
VIT_07s0031g02160	F6H4D2	Protein phosphatase 2C DBP	0.2081
VIT_00s0802g00020	D7ST21	Unknown protein	0.2044
VIT_00s0203g00040	D7UDP4	Vesicle-associated membrane protein	0.1944
VIT_16s0050g02630	F6H6F7	FtsH protease that is localized to the chloroplast	0.1879
VIT_09s0002g04080	F6HYD4	IAA9	0.1744

Table S4: Individual Omics Analysis. Most relevant features obtained from the RF model for the metabolomics dataset.

Metabolites	Importance
glucose	0.6228
malic acid	0.5795
peonidin-3-glucoside	0.5158
tartaric acid	0.4253
sucrose	0.4049
benzenemethanol	0.3792
delphinidin 3-O-glucoside	0.3737
threonic acid	0.3500
fructose	0.3218
stearic acid	0.3132
citric acid	0.3093
cyanidin 3-glucoside	0.2984
phenylalanine	0.2967
myricetin-3-glucoside	0.2933
quercetin-3-glucuronide	0.2843
malvidin-3-O-glucoside	0.2764
palmitic acid	0.2344
spirotetramat	0.2334
peonidin 3-(6''-acetylglucoside)	0.2188
petunidin-3-glucoside	0.2152

Table S5: Individual Omics Analysis. Most relevant features obtained from the ANN model for the transcriptomics dataset.

Transcripts	UniProtKB	Annotation	Importance
VIT_18s0001g01490	A5ASV7	Oxidoreductase N-terminal domain-containing	5.247
VIT_05s0020g02690	F6HDL7	Copper-binding family protein	4.647
VIT_13s0064g00890	F6HB61	Cellulose synthase CESA3	4.066
VIT_14s0083g01110	D7SMP3	Brassinosteroid-6-oxidase	3.724
VIT_19s0014g03850	A5BX41	Cytochrome B6-F complex iron-sulfur subunit, PETC	3.724
VIT_03s0091g00500	F6H673	Unknown protein	3.719
VIT_11s0016g05840	F6HHB3	Protease inhibitor/seed storage/lipid transfer protein (LTP)	3.647
VIT_16s0022g00670	F6HAU0	Vacuolar invertase 1, GIN1	3.172
VIT_06s0004g03240	A5AFS1	Elongation factor 1-alpha 1	3.062
VIT_05s0020g04880	D7T7A2	Seed specific protein Bn15D14A	3.001
VIT_10s0003g00980	F6HM77	Unknown protein	2.501
VIT_12s0028g01080	A5B1D3	Photosystem II oxygen-evolving complex precursor,23kda PSBP	2.071
VIT_05s0020g04490	F6HDW2	No hit	1.477
VIT_06s0004g00200	F6GUP1	Splicing factor YT521-B	1.413
VIT_04s0023g03010	F6GWQ0	fructose-bisphosphate aldolase,chloroplast precursor	1.363
VIT_01s0011g02710	A5AEV3	No hit	1.350
VIT_16s0039g02550	F6HEH4	Seed specific protein Bn15D1B	1.182
VIT_16s0050g02530	A8VPW6	Myb Triptychon	1.140
VIT_06s0009g00410	F6HAC5	BLH1 (embryo sac development arrest 29)	1.098
VIT_06s0004g07880	D7SJK7	Allergen	1.060

Table S6: Individual Omics Analysis. Most relevant features obtained from the ANN model for the metabolomics dataset.

Metabolites	Importance
fructose	13.9159
proline	13.1387
malic acid	11.5559
leucine	7.2458
N-methylnicotinic acid cation	7.1200
tartaric acid	6.4139
glucose	4.2651
stearic acid	3.0876
1-methylgalactose NIST	3.0555
malvidin-3-O-glucoside	2.9699
1,3,5-benzenetriol	2.7066
peonidin-3-glucoside	2.6081
hydroxylamine	2.1694
malvidin 3-(6"-acetylglucoside)	1.2324
erythritol	1.1846
benzenemethanol	0.9632
13-docosenamide	0.9469
lactulose	0.7142
sucrose	0.6676
peonidin 3-(6"-acetylglucoside)	0.6353

B.2 CASE STUDY II

Table S1: Values of the different error metrics (Accuracy, Recall and Precision) for each model (SVM, RF and ANN) for both the transcriptomics and fluxomics datasets in Case Study II.

Model	Metrics	Transcriptomics	Fluxomics
SVM	Accuracy	0.75	0.33
	Recall	0.8	1
	Precision	0.8	0.33
RF	Accuracy	0.5	0.5
	Recall	0.4	0.5
	Precision	0.67	0.33
ANN	Accuracy	0.5	0.41
	Recall	0.4	0.5
	Precision	0.67	0.28

Table S2: Individual Omics Analysis. Most relevant features obtained from the SVM model for the transcriptomics dataset.

CATMA ID	TAIR7 mapping	Annotation	Importance
CATMA1a22270	At1g23200	pectinesterase family protein	0.9467
CATMA4a19840	At4g18700	Encodes CBL-interacting protein kinase 12 (CIPK12)	0.9290
CATMA1c72010	At1g66390	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2)	0.9172
CATMA1a18630	At1g19610	LCR78/PDF1.4 (Low-molecular-weight cysteine-rich 78)	0.8817
CATMA3c57894	At3g60910	catalytic,Generic methyltransferase	0.8817
CATMA4a30730	At4g29070	unknown protein, Phospholipase A2, PLA2	0.8817
CATMA1c72183	At1g76680	OPR1 (12-oxophytodienoate reductase 1)	0.8757
CATMA3a18630	At3g18980	F-box family protein	0.8698
CATMA4c42685	At4g37980	ELI3-1 (ELICITOR-ACTIVATED GENE 3)	0.8639
CATMA1c71101	At1g08360	60S ribosomal protein L10A (RPL10aA)	0.8521

Table S3: Individual Omics Analysis. Most relevant features obtained from the SVM model for the fluxomics dataset.

Reaction	Overall Importance	Subsystem
TCP7	4.973	Triose phosphate translocator (G3P)
R01015_p	3.195	Fructose and mannose metabolism; Glycolysis / Gluconeogenesis; Inositol metabolism;Carbon fixation
R01015_c	3.181	Fructose and mannose metabolism; Glycolysis / Gluconeogenesis; Inositol metabolism;Carbon fixation
TCP8	2.095	Triose phosphate trasnlocator (glyceroneP)
R00127_c	1.614	Purine metabolism
R03321_c	1.500	Glycolysis / Gluconeogenesis
R02739_c	1.500	Pentose phosphate pathway;Glycolysis / Gluconeogenesis
TCX16	1.393	Isocitrate transporter
R01324_c	1.393	Citrate cycle (TCA cycle)
R01325_x	1.315	Glyoxylate and dicarboxylate metabolism; Reductive carboxylate cycle (CO2 fixation);Citrate cycle (TCA cycle)

Table S4: Individual Omics Analysis. Most relevant features obtained from the RF model for the transcriptomics dataset.

CATMA ID	TAIR7 mapping	Annotation	Importance
CATMA1a22270	At1g23200	pectinesterase family protein	0.16446
CATMA3a18630	At3g18980	F-box family protein	0.13316
CATMA1a28510	At1g30500	CCAAT-binding transcription factor (CBF-B/NF-YA) family protein	0.12914
CATMA5c65084	At5g62300	40S ribosomal protein S20 (RPS20C)	0.12542
CATMA1c72183	At1g76680	OPR1 (12-oxophytodienoate reductase 1)	0.12003
CATMA4a19460	At4g18390	TCP family transcription factor, putative	0.11641
CATMA1a18630	At1g19610	LCR78/PDF1.4 (Low-molecular-weight cysteine-rich 78)	0.11298
CATMA3c57894	At3g60910	catalytic, domain Generic methyltransferase	0.10418
CATMA1c72010	At1g66390	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2)	0.10086
CATMA2a45990	At2g47530	unknown protein	0.10045

Table S5: Individual Omics Analysis. Most relevant features obtained from the RF model for the fluxomics dataset.

Reaction	Importance	Subsystem
R01070N_c	0.07819	Fructose and mannose metabolism; Glycolysis / Gluconeogenesis; Pentose phosphate pathway; Carbon fixation
TCP21	0.07721	ADP transporter
R02950_c	0.07013	coniferyl alcohol; Coumarine and phenylpropanoid biosynthesis (Lignin subunit)
R03968_c	0.06976	Valine, leucine and isoleucine biosynthesis
TCP20	0.06973	ATP transporter
TCP6	0.06895	alpha-D-Glucose 6-phosphate transporter
R00709_m	0.06879	Citrate cycle (TCA cycle)
R01943_c	0.06609	Stilbene, coumarine and lignin biosynthesis
R00948_p	0.06504	Starch and sucrose metabolism
TCP7	0.06449	Triose phosphate translocator (G3P)

Table S6: Individual Omics Analysis. Most relevant features obtained from the ANN model for the transcriptomics dataset.

CATMA ID	TAIR7 mapping	Annotation	Importance
CATMA1c72010	At1g66390	PAP2 (PRODUCTION OF ANTHOCYANIN PIGMENT 2)	1.5292
CATMA4a03190	At4g02840	small nuclear ribonucleoprotein D1	1.0633
CATMA5c64551	At5g35525	unknown protein	0.9745
CATMA1A24995	At1g26770	ATEXPA10 (ARABIDOPSIS THALIANA EXPANSIN A10)	0.9137
CATMA4c42685	At4g37980	ELI3-1 (ELICITOR-ACTIVATED GENE 3)	0.8881
CATMA3A38785	At3g45780	PHOT1 (phototropin 1)	0.8847
CATMA1A18630	At1g19610	LCR78/PDF1.4 (Low-molecular-weight cysteine-rich 78)	0.8814
CATMA5A47680	At5g51750	subtilase family protein	0.8731
CATMA1a28510	At1g30500	CCAAT-binding transcription factor (CBF-B/NF-YA) family protein	0.8669
CATMA1a08610	At1g09750	chloroplast nucleoid DNA-binding protein-related	0.8590

Table S7: Individual Omics Analysis. Most relevant features obtained from the ANN model for the metabolomics dataset.

Reaction	Importance	Subsystem
TCP7	4.973	Triose phosphate translocator (G3P)
R01015_p	3.195	Fructose and mannose metabolism; Glycolysis / Gluconeogenesis; Inositol metabolism; Carbon fixation
R01015_c	3.181	Fructose and mannose metabolism; Glycolysis / Gluconeogenesis; Inositol metabolism; Carbon fixation
TCP8	2.095	Triose phosphate trasnlocator (glyceroneP)
R00127_c	1.614	Purine metabolism
R03321_c	1.500	Glycolysis / Gluconeogenesis
R02739_c	1.500	Pentose phosphate pathway; Glycolysis / Gluconeogenesis
TCX16	1.393	Isocitrate transporter
R01324_c	1.393	Citrate cycle (TCA cycle)
R01325_x	1.315	Glyoxylate and dicarboxylate metabolism; Reductive carboxylate cycle (CO2 fixation);Citrate cycle (TCA cycle)