Universidade do Minho
Escola de Engenharia

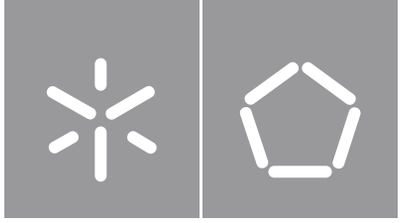
Marta Sofia Costa Sampaio

M3L: Exploring synergies between metabolic modelling and machine learning for studying plants

April, 2024

M3L: Exploring synergies between metabolic modelling and machine learning for studying plants
Marta Sofia Costa Sampaio
UMinho | 2024

FCT Fundação para a Ciência e a Tecnologia
REPÚBLICA PORTUGUESA
CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
NORTE2020 PROGRAMA OPERACIONAL REGIONAL DO NORTE
PORTUGAL 2020
UNIÃO EUROPEIA
Fundo Social Europeu



Universidade do Minho
Escola de Engenharia

Marta Sofia Costa Sampaio

M3L: Exploring synergies between metabolic modelling and machine learning for studying plants

Doctorate Thesis
Doctorate in Informatics

Work developed under the supervision of:
Professor Oscar Dias
Professor Miguel Rocha

April, 2024

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.18) [1].

Acknowledgements

I would like to thank Fundação para a Ciência e Tecnologia for the Ph.D. studentship I was awarded (SFRH/BD/144643/2019). This work is only possible due to their support. I would also like to thank the Centre of Biological Engineering at the University of Minho for hosting me and providing the resources, facilities, and an excellent working environment.

I also want to express my gratitude to my supervisors, Oscar Dias and Miguel Rocha, for their invaluable guidance, support, and encouragement throughout this journey. Their expertise, patience, and mentorship have been instrumental in shaping this thesis and my academic growth.

I am immensely grateful to my friends, especially Fernando, Diogo, Fernanda, Alexandre, Emanuel, Sequeira, Capela, Nuno, Pacheco, Daniela and Carla, for all the support, companionship, and shared experiences over these years, which have enriched my academic and personal life.

I am also deeply grateful to my family for their unwavering support, love, and understanding throughout this challenging journey.

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 Universidade do Minho.

_____, _____

(Place) (Date)

_____

(Marta Sofia Costa Sampaio)

Resumo

M3L: Explorar sinergias entre modelação metabólica e aprendizagem automática no estudo das plantas

Vitis vinifera (videira) é uma das principais árvores de fruto do mundo, sendo amplamente cultivada e comercializada, principalmente para produção de vinho. Como a qualidade do vinho está diretamente relacionada com a qualidade da uva, o estudo do metabolismo da videira é importante para esclarecer os processos subjacentes à composição da uva. Modelos metabólicos à escala genómica têm sido usados no estudo do metabolismo das plantas e vários avanços foram feitos, permitindo a integração de dados ómicos. Por outro lado, abordagens de aprendizagem automática têm sido usadas para analisar e integrar diferentes dados ómicos ao mesmo tempo que a combinação da aprendizagem automática com modelos metabólicos tem mostrado resultados promissores. Neste trabalho, um repositório de dados metabólicos de plantas (*iplants*) foi criado para centralizar e integrar dados de metabolismo e sequências de genes e proteínas de plantas. A partir deste repositório, o primeiro modelo metabólico à escala genómica da videira foi reconstruído, incluindo 7199 genes, 5399 reações e 5141 metabolitos distribuídos por 8 compartimentos. Modelos metabólicos específicos da folha, caule e uva foram criados a partir do modelo genérico através da integração de dados ómicos do cultivar Cabernet Sauvignon. Adicionalmente, para capturar as mudanças no metabolismo da uva durante o seu desenvolvimento, dois modelos de uva foram criados para representar a uva verde e a uva madura. Estes modelos foram agrupados, criando um modelo de múltiplos tecidos para estudar as interações entre tecidos durante o dia e a noite. O potencial de combinar modelos metabólicos e aprendizagem automática foi explorada ao usar modelos de aprendizagem automática para analisar os dados de fluxómica gerados pelos modelos metabólicos da uva e identificar as reações que mais contribuem para distinguir entre o estado verde e maduro. Além disso, uma análise multi-ómicas foi realizada ao integrar a fluxómica prevista com os dados de RNA-Seq e metabolómica para identificar relações entre genes, reações e metabolitos que podem explicar os diferentes fenótipos. Assim, os modelos metabólicos desenvolvidos aqui são ferramentas úteis para explorar diferentes aspetos do metabolismo das videiras e os fatores que influenciam a qualidade da uva. A integração de dados ómicos, modelos metabólicos e aprendizagem automática realizada mostrou o potencial de aplicar esta estratégia para análises mais completas quando mais dados estiverem disponíveis.

Palavras-chave: aprendizagem automática, dados ómicos, metabolismo da videira, modelos metabólicos à escala genómica, modelos multi-tecidos dia-noite

Abstract

M3L: Exploring synergies between metabolic modelling and machine learning for studying plants

Vitis vinifera, also known as grapevine, is one of the most important crops in the world, being widely cultivated and commercialized, mainly to produce wine. As wine quality is directly linked to fruit quality, the study of grapevine metabolism is important to understand the processes underlying grape composition. Towards similar aims, genome-scale metabolic models (GSMMs) have been used for the study of plant metabolism and several advances have been made, also allowing the integration of omics datasets with GSMMs. In this context, Machine Learning (ML) has been used to analyse and integrate omics data, while the combination of ML with GSMMs has shown promising results in different studies but is still scarcely used to study plants.

In this work, a repository of plant metabolic data (*iplants*) was created to centralise and integrate metabolic and sequence data from diverse plants. Then, the first GSMM of *V. vinifera* was reconstructed from *iplants*, comprising 7199 genes, 5399 reactions, and 5141 metabolites across 8 compartments. Tissue-specific models for the stem, leaf, and berry of the Cabernet Sauvignon cultivar were generated from the original model, through the integration of RNA-Seq data. Furthermore, to capture the changes in grape metabolism, two separate models were created representing the grape in both its green and mature states. These models have been merged into diel multi-tissue models to study the interactions between tissues at both light and dark phases. The potential of combining ML with GSMMs was explored by using ML to analyse the fluxomics data generated by the green and mature grape models and provide insights regarding the reactions that contribute the most to distinguishing these states. In addition, a multi-omics analysis was performed by integrating the generated fluxomics data with RNA-Seq and metabolomics data to identify connections between genes, reactions, and metabolites that can explain different phenotypes. Therefore, the models developed in this work are useful tools to explore different aspects of grapevine metabolism and the factors influencing grape quality. The integration of omics data, GSMMs, and ML for studying plant metabolism has shown the potential of applying this approach for more insightful analyses as more data becomes available.

Keywords: diel multi-tissue models, grapevine metabolism, genome-scale metabolic models, machine learning, omics data

Contents

List of Figures

List of Tables

Acronyms

ACHR Artificial Co-ordinate Hit and Run *(pp. 94, 95, 115, 117, 118)*
AdaM Adaptation of Metabolism *(p. 25)*
ANN Artificial Neural Network *(pp. 38, 53)*
API Application Programming Interface *(pp. 59–62, 65, 69, 71, 73, 75, 76, 86, 150)*
ATP Adenosine Triphosphate *(pp. 6, 15)*
AUC Area under the ROC Curve *(p. 39)*

BER Balanced Error Rate *(pp. 142–144)*
BLAST Basic Local Alignment Search Tool *(p. 12)*
BRENDA BRaunschweig Enzyme Database *(pp. 11, 13, 14)*

CBM Constraint-Based Modelling *(pp. 47–49, 51, 53–57)*
CCA Canonical Correlation Analysis *(pp. 43, 45–47)*
CDS Central Data Storage *(pp. 61, 62, 69, 74, 77)*
COBRA Constraint-Based Reconstruction and Analysis *(p. 11)*
CORDA Cost Optimization Reaction Dependency Assessment *(p. 24)*
CSV Comma Separated Values *(pp. 68, 71, 76, 88)*

DBMS DataBase Management System *(pp. 60–62, 64)*
DDBJ DNA DataBank of Japan *(pp. 19, 20)*
dFBA Dynamic Flux Balance Analysis *(pp. 18, 25, 30, 33, 49, 53)*
DIABLO Data Integration Analysis for Biomarker discovery using Latent cOmponents *(pp. 43, 141–143, 145, 146, 149, 151, 182)*

EC Enzyme Commission *(pp. 12–14, 65)*
EFMs Elementary Flux Modes *(pp. 25, 48, 49, 53)*
ENA European Nucletoide Archive *(pp. 19, 20)*

ETL Extract, Transform and Load *(pp. 60–62, 74)*
EXAMO EXploration of Alternative Metabolic Optima *(p. 26)*

FALCON Flux Assignment with Least absolute deviation Convex Objectives and Normalization *(p. 25)*
FastCORE Fast Reconstruction of Compact Context-specific Metabolic Network Model *(pp. 24, 27, 93)*
FBA Flux Balance Analysis *(pp. 16–18, 22, 23, 33, 49–51, 53, 54, 56, 86)*
FCa Flux Capacity *(pp. 130, 134)*
FCGs Flux-Coupled Genes *(p. 25)*
FVA Flux Variability Analysis *(pp. 18, 49, 53, 86, 125, 130)*

GBMDB Global Proteome Machine Database *(pp. 19, 21)*
GDM Golm Metabolome Database *(pp. 20, 21)*
GEMsplice Genome-scale metabolic modelling with splice-isoform integration *(p. 27)*
GEO Gene Expression Omnibus *(pp. 19, 20, 59, 68, 69, 71, 73, 74, 76, 77, 92)*
GIM3E Gene Inactivation Moderated by Metabolism, Metabolomics and Expression *(p. 22)*
GIMME Gene Inactivity Moderated by Metabolism and Expression *(pp. 22, 25, 27)*
GPR Gene-Protein-Reaction *(pp. 12, 25, 51, 68, 86–89, 91, 104–106, 108, 135, 153)*
GREAT GRape Expression ATlas *(pp. 93, 110–112)*
GSMM Genome-Scale Metabolic Model *(pp. 1–4, 7, 11, 12, 15–19, 21, 23, 27–30, 32–37, 48–52, 55, 59, 61, 63, 69, 71, 72, 74–77, 79, 85, 86, 91–93, 96, 99, 109, 111, 112, 117, 129, 130, 132, 150–153, 182, 183)*
GX-FBA Gene-Expression FBA *(pp. 23, 27)*

iMAT Integrated Metabolic Analysis Tool *(pp. 23, 24, 26, 27)*
iNIT Integrative Network Inference for Tissues *(pp. 23, 24, 27)*
IOMA Integrative Omics Metabolic Analysis *(p. 27)*
IPP isopentenyl pyrophosphate *(pp. 7, 9)*

JSON JavaScript Object Notation *(pp. 61, 62, 64, 65, 69–71, 73–76, 87)*

KBase The Department of Energy Systems Biology Knowledgebase *(p. 11)*
KEGG Kyoto Encyclopedia of Genes and Genomes *(pp. 11, 12, 14, 59, 78, 79, 104)*
KNN k-Nearest Neighbors *(pp. 38, 53, 131, 134, 135, 137, 139)*

LASSO least absolute shrinkage and selection operator *(p. 54)*
lbFBA Linear-bounds FBA *(pp. 26, 27)*

MADE Metabolic Adjustment by Differential Expression *(pp. 25, 27)*

MassIVE Mass Spectrometry Interactive Virtual Environment *(pp. 19, 21)*
MBA Model Building Algorithm *(pp. 24, 26, 27)*
mCADRE Metabolic Context-specificity Assessed by Deterministic Reaction Evaluation *(p. 24)*
MCMC Markov chain Monte Carlo *(p. 18)*
MEP methylerythritol phosphate *(pp. 9, 135, 136, 149, 151)*
merlin Metabolic Models Reconstruction Using Genome-Scale Information *(p. 11)*
METRADE MEabolic and TRanscriptomics ADaptation Estimator *(pp. 22, 27)*
MEV mevalonic acid *(p. 9)*
miRNA micro RNA *(p. 46)*
MKL Multiple Kernel learning *(p. 45)*
ML Machine Learning *(pp. 1–4, 37, 39–44, 47–57, 129–132, 136–138, 141, 149–151, 153, 180, 182)*
MOMA Minimization of Metabolic Adjustment *(p. 18)*
MS Mass Spectrometry *(p. 19)*

NCBI National Center for Biotechnology Information *(pp. 2, 11, 12, 19, 69, 72, 73, 75, 150)*
NMR Nuclear Magnetic Resonance Spectrometry *(p. 19)*

ODM Object Document Mapper *(pp. 61, 68)*
OGM Object Graph Mapper *(pp. 61, 64)*
oPLS orthogonal PLS *(p. 43)*
ORF Open Reading Frame *(p. 12)*

PCA Principal Component Analysis *(pp. 37, 38, 45–48, 51, 53, 54, 56)*
PEMA Principal Elementary mode Analysis *(p. 48)*
pFBA Parsimonious Flux Balance Analysis *(pp. 18, 27, 92, 96, 109, 116, 119)*
PFMA Principal Metabolic Flux Mode Analysis *(p. 48)*
PGDBs Pathway/Genome databases *(p. 12)*
PLS Partial Least Squares *(pp. 43, 45–47, 141, 142)*
PMN Plant Metabolic Network *(pp. 12, 13, 59)*
PODC Plant Omics Data Center *(pp. 19, 20)*
PPDB Plant Proteomics Database *(pp. 19, 21)*
PRIDE PRoteomics IDEntifications *(pp. 19, 20)*
PROM Probabilistic Regulation Of Metabolism *(p. 23)*
PTMs post-translational modifications *(p. 19)*

RefSeq NCBI's Reference Sequence Database *(pp. 19, 20)*
RegrEx Regularised Context-specific model Extraction *(p. 26)*

RELATCH RELATive CHange *(pp. 26, 27)*
RF Random Forests *(pp. 38, 53, 131, 132, 134, 135, 138, 140)*
RIPTiDe Reaction Inclusion by Parsimony and Transcript Distribution *(p. 27)*
RMF Required Metabolic Function *(pp. 21, 22, 24)*
ROOM Regulatory on/off Minimization of Metabolic flux changes *(p. 18)*
ROS Reactive Oxygen Species *(pp. 147, 149)*

SBML Systems Biology Markup Language *(pp. 11, 16, 68, 71, 91, 182, 183)*
sCCA sparse CCA *(p. 43)*
SDRF Sample and Data Relationship Format *(pp. 68, 69, 71, 76)*
SHAP SHAPley Additive exPlanations *(pp. 131, 134, 135, 137–140)*
SOFT Simple Omnibus Format in Text *(pp. 61, 68, 69, 71, 76)*
sPLS sparse PLS *(pp. 43, 46, 47)*
SQL Structured Query Language *(p. 61)*
SRA Sequence Read Archive *(pp. 19, 20)*
SUBA SUBcellular location database for *Arabidopsis* proteins *(p. 14)*
SVM Support Vector Machine *(pp. 38, 45, 53, 131, 132, 137, 139)*

t-SNE T-distributed Stochastic Neighbor Embedding *(pp. 37, 38, 93, 94, 111, 112, 115, 117, 118, 131, 132)*
TAIR The Arabidopsis Information Resource *(pp. 11, 14, 70, 72, 78)*
TC Transport Classification *(pp. 12–14)*
TCA Tricarboxylic acid *(pp. 6, 8, 109, 110, 118, 119, 121)*
TCDB Transporter Classification Database *(pp. 11, 13, 14)*
TEAM Temporal Expression-based Analysis of Metabolism *(p. 25)*
tFBA Transcriptionally Controlled FBA *(p. 25)*
tINIT Task-driven INIT *(pp. 23, 24)*
TPM transcripts per million *(pp. 93, 111, 112)*
TRN transcriptional regulatory networks *(p. 23)*
Troppo Tissue-specific RecOnstruction and Phenotype Prediction using Omics data *(p. 27)*
TXT Text File *(p. 68)*

uFBA unsteady-state flux balance analysis *(p. 51)*
UniProt Universal Protein Resource *(pp. 2, 11, 13, 14, 52, 59, 66, 70, 72, 75, 78, 79, 150)*

XML Extensible Markup Language *(pp. 61, 64, 68)*

1 Introduction

1.1 Context and Motivation

Plants are multicellular eukaryotic photosynthetic organisms indispensable for human life. They are the ultimate food source for almost all animals, including humans (legumes, fruits, cereals, among others), maintain the atmosphere balance by consuming carbon dioxide and releasing oxygen, and provide many materials for human use such as wood, fibres for clothing, drugs, and fuels [2]. As plants are sessile organisms, they can not escape from environmental stresses or pathogens, being forced to face a wide range of adverse environmental conditions and interact with several pathogenic or beneficial organisms. As a result, they have the most complex metabolic networks that produce an enormous diversity of metabolites to help them grow, adapt to the environment, and defend against pathogens. These metabolites are poorly explored and can have potential biotechnological applications.

In addition, plants have a significant impact on the economy. Grapevine (*Vitis vinifera*) is cultivated worldwide and has high economic value, mainly due to wine production. In 2022, wine exports reached a value of 37600 million euros [3]. In Portugal, wine exports reached 994 million euros, representing almost 7% of total exports that year [4]. Therefore, as grapevines have high economic interest and fruit quality is intrinsically linked to metabolism, the study of grapevine metabolism is essential for understanding its responses to different environmental conditions that may affect grape metabolic composition.

Metabolism has been studied by Systems Biology approaches, which use a Genome-Scale Metabolic Model (GSMM) to represent the metabolic networks of an organism and perform *in silico* phenotype predictions. GSMMs have been reconstructed mainly for unicellular organisms, like bacteria and yeasts. However, several models are available for plants, including *Arabidopsis thaliana*, *Oryza sativa* (rice) and *Zea mays* (maize) [5]. Despite the challenges in plant genome-scale metabolic modelling, new approaches have emerged to reconstruct more realistic models through the integration of omics data, mainly transcriptomics, and build diel multi-tissue models that allow the study of metabolic interactions between plant tissues in a day-night cycle.

Despite the existence of several methods for integrating omics into GSMMs, this is still a challenging and inefficient task. As omics datasets are complex and heterogeneous, Machine Learning (ML) has been widely used to process, analyse and integrate different types of omics and to extract biological knowledge

from data [6].

Recently, ML and metabolic modelling approaches have been combined to improve the model's predictions and interpretability, and this strategy has shown promising results [7–11]. ML can be used to extract knowledge from the fluxomics data generated by the models or to integrate the predicted fluxomics data with experimental omics. Thus far, these studies have been mainly applied to bacteria, yeasts, and human cells, but not to plants.

Additionally, the existence of a centralised repository to include and organise relevant plant data is of utmost importance for facilitating the study of plant metabolism and the development of new bioinformatics tools. Besides integrating information from plant databases, the repository should contain all relevant data, such as genomes, annotations, metabolic models and omics data to allow the comparison, analysis and integration of information from different sources.

In this work, computational tools were developed to assist the reconstruction and analysis of the GSMM of *V. vinifera*. In the first phase, a repository of plant metabolic data named *iplants* was built to centralise and integrate the data available in the PlantCyc and MetaCyc databases and plant GSMMs. As the first step in a GSMM reconstruction is genome annotation, sequence data was retrieved from Universal Protein Resource (UniProt) and National Center for Biotechnology Information (NCBI) for the proteins and genes in the repository. The second phase consisted of reconstructing the GSMM of *V. vinifera* using *iplants* as the source of metabolic data. This model was tested for the different metabolic processes of plants: photosynthesis, photorespiration, and respiration. After reconstructing the generic model, tissue-specific models were created for the stem, leaf, and berry by integrating omics data to explore the metabolic differences between tissues. Furthermore, to capture the changes in grape metabolism, we created separate models representing the grape in its green and mature states. These models were then merged into a multi-tissue model and the day-night cycle was included to study the interactions between tissues and the light and dark phases. In the last phase, the phenotype predictions obtained with berry models were analysed by ML to identify the reactions that most contribute to the model's predictions of the grape developmental state. In addition, a multi-view method was used to integrate the flux predictions generated by the models with RNA-Seq and metabolomics data to identify connections between genes, reactions, and metabolites that can help discover biomarkers of the grape developmental state.

1.2 Research Objectives

The main goal of this thesis is to develop and apply computational tools towards the reconstruction and analysis of the GSMM of *V. vinifera* by exploring synergies between ML, omics and metabolic modelling approaches. These tools and the results from their application may improve the knowledge of grapevine's metabolic phenotypes and factors affecting grape quality.

The following scientific and technological objectives will be pursued to accomplish this goal:

- Review state-of-art methods for studying plant metabolism, including plant metabolic modelling approaches, methods integrating omics data within GSMMs, tools for integrating and analysing

multi-omics datasets and studies employing both ML and metabolic modelling approaches;

- Create an integrated and unified repository of plant metabolic data that will include the data available in PlantCyc and MetaCyc, the sequence of genes and proteins, and the content of the published plant GSMMs to assist in the model reconstruction;
- Reconstruct a generic GSMM of *V. vinifera* using the genome sequence and the created repository;
- Select the best omics datasets to represent the main tissues of *V. vinifera*, leaf, stem and berry;
- Explore different methods and strategies to integrate omics data with GSMMs, including the reconstruction of tissue-specific models by integrating omics data;
- Create a multi-tissue model by merging the tissue models and a diel model by integrating a day-night cycle in the multi-tissue model;
- Validate the created models by predicting the metabolism under photosynthesis, photorespiration, and respiration, and predicting the behaviour of *V. vinifera* under different environmental conditions.
- Generate fluxomics data using the reconstructed models and apply ML approaches to this data to explore grapevine metabolism.
- Perform multi-omics integration using experimental omics and phenotype predictions from the models to identify connections between the variables of the different datasets, providing a comprehensive understanding of the relation between the genotype and phenotype of the grape.

1.3 Report Outline

This document is divided into six chapters and outlined as follows:

- The current chapter consists of the context and motivation for the reconstruction and analysis of plant GSMMs, mainly *V. vinifera*. In addition, the main objectives of this work are briefly described as well as the thesis outline;
- Chapter 2 introduces the fundamental concepts for understanding plant metabolism, the reconstruction and simulation of GSMMs, the analyses of omics datasets, and the creation and evaluation of ML models. In addition, it provides a detailed description of the available plant GSMMs and their applications, state-of-art methods for omics integration with GSMMs and multi-omics integration, and the recent studies combining ML and metabolic modelling for studying the metabolism. Finally, it discusses the limitations of plant metabolic modelling and the relevance of combining these two approaches in the study of plant metabolism.
- Chapter 3 addresses the creation of a unified and integrated repository of plant metabolic data and describes the pipelines created for data extraction, transformation, and update.

- Chapter 4 describes the steps taken during the reconstruction of the GSMM of *V. vinifera* and the strategies used for the creation of the tissue-specific and diel multi-tissue models. It also includes the analysis of phenotype predictions generated by these models.
- Chapter 5 comprehends the application of ML in the analysis of the fluxomics data predicted by GSMMs to understand the metabolic changes during grape development. In addition, it describes the strategy used to integrate this fluxomics data with experimental omics, mainly transcriptomics and metabolomics data, to identify biomarkers of grape development stages.
- Chapter 6 is a brief synopsis of this project, providing an assessment of the results obtained and future tasks that can improve the computational tools and analyses developed in this work.

2

State of the Art

The work presented in this chapter has also been partially published in the publication:

Sampaio M, Rocha M, Dias O. Exploring synergies between plant metabolic modelling and machine learning. Comput Struct Biotechnol J. 2022 Apr 16;20:1885-1900. doi: 10.1016/j.csbj.2022.04.016. PMID: 35521559; PMCID: PMC9052043.

2.1 Plant metabolism

Metabolism is defined as the set of all biochemical reactions taking place in an organism. Metabolic reactions are mostly catalysed by enzymes and provide the energy required for growth, reproduction, structural maintenance, and adaptation to the environment. Enzymes couple energy-consuming reactions (anabolism) with energy-releasing reactions (catabolism) to allow the biosynthesis of new metabolites and can be regulated to increase or decrease the rate of metabolic reactions, in response to internal signals or changes in environmental conditions [12, 13].

As plants are sessile organisms, they are not able to escape from environmental stresses or pathogens, being forced to face a wide range of environmental conditions and interact with several pathogenic or beneficial organisms. As a result, plants have the most complex metabolic networks that produce an enormous diversity of metabolites to help them grow, adapt to the environment, and defend against pathogens [14]. The specialised metabolites are generally unique to a defined species and are only produced in a specific organ, tissue, or cell at a particular developmental stage or under certain environmental conditions. Additionally, the extensive compartmentalisation of the interconnected metabolic pathways in plants, which does not exist in prokaryotic organisms, makes plant metabolism even more complex [15].

According to their roles, metabolites can be divided into primary and secondary metabolites. Primary metabolites, such as sugars and amino acids, are typically present in all plants and play an essential role in their growth, reproduction, and development [16]. In plants, the primary metabolism starts when light energy is converted into chemical energy, through photosynthesis, which produces the sugars needed for starting cellular respiration. The sugars are then broken down through glycolysis and Tricarboxylic acid (TCA) cycle pathways to produce energy in the form of Adenosine Triphosphate (ATP) molecules. Glucose can also be oxidised through the pentose phosphate pathway. The intermediates of these pathways are the precursors for the biosynthesis of more complex compounds, such as starch, amino acids, fatty acids, and structural compounds forming the cell wall and membranes.

Secondary metabolites are not crucial for plants to live and grow but play an important role in their adaptation to the environment. Therefore, plants use these metabolites as signal compounds to attract pollinators or animals for seed-dispersing, to communicate with symbiotic organisms, and as defence agents against herbivores, pathogenic fungi, bacteria or viruses, and abiotic stresses, such as ultraviolet (UV) light or temperature [17]. Figure 1 provides a simplified overview of the central metabolic pathways of plants, which produce the precursors for the biosynthesis of secondary metabolites.

Secondary metabolites are synthesised from primary metabolites and are often specific to individual organisms or species. They are characterised by an enormous chemical diversity and varied properties that make them compounds of high interest for pharmaceutical and biotechnological applications. Despite their diversity, most secondary metabolites are derived from central metabolic pathways, like the TCA, the isoprenoid pathway, which leads to the biosynthesis of several isoprenoids, such as sterols, and the shikimate pathway, which is used for the biosynthesis of folates and aromatic amino acids [2, 17]. Based on their origin, these metabolites can be divided into three major groups: terpenoids, phenylpropanoids, and alkaloids (Figure 2).

Terpenoids derive from the five-carbon precursor isopentenyl pyrophosphate (IPP) from the isoprenoid pathway and have been used mainly as fragrances and flavours (Figure 3). Phenylpropanoids are biosynthesised from phenylalanine and tyrosine, two amino acids produced in the shikimate pathway (Figure 4). These include, for instance, the flavonoids that are involved in plant pigmentation, pollination, and protection from UV radiation. Alkaloids are derived from different amino acids and have many clinical purposes, including analgesic, such as morphine, and anticancer, like vinblastine. In most cases, the biosynthesis of secondary metabolites only happens in a single organ, but they can be transported long distances to other tissues [2, 17].

Thereby, in addition to being the ultimate source of food, plants represent an enormous source of bioactive compounds with biotechnological and pharmaceutical relevance, having a significant impact on the economy. As plants face diverse environmental stresses and answer to these stresses through changes in their metabolism, the study of plant metabolism is of utmost importance for understanding phenotypes of disease resistance and survival under harsh environmental conditions and improving the production of fruits or metabolites of interest.

2.2 Constraint-based Modelling

2.2.1 Reconstruction of genome-scale metabolic models

2.2.1.1 Background

A GSMM is defined as an *in silico* representation of all metabolic reactions taking place within an organism and is mainly derived from genome annotation [18]. The first GSMM was reconstructed over 20 years ago (1999) for *Haemophilus influenza* [19]. Ever since, with the advance and improvement of high-throughput technologies and systems biology approaches, GSMMs have been reconstructed and are widely used for understanding the metabolism of many organisms, including bacteria, fungi, animals, plants, and humans. The applications of metabolic models are vast and include the understanding of metabolic diseases, prediction of enzymatic functions, identification of essential genes and drug targets, strain design for overproduction of compounds of interest, and modelling interactions between cells or organisms [5].

The reconstruction of GSMMs is an iterative and time-consuming process that requires the integration of multiple data obtained from different databases and bioinformatics tools. This process has been standardised and simplified by several authors, who have divided it into four main steps [18, 20], as represented in Figure 5.

In short, the reconstruction starts with genome annotation, which provides the information needed to build a draft metabolic network. Then, metabolic knowledge is retrieved from biochemical databases and literature to refine the metabolic network, which is converted to a GSMM by formulating the biomass production equation and other constraints specific to the organism. Finally, the resulting model is validated by comparing its phenotype predictions with experimental data. The annotation and formulation steps are revised iteratively until the predictions are in accordance with the experimental data [20]. The final model

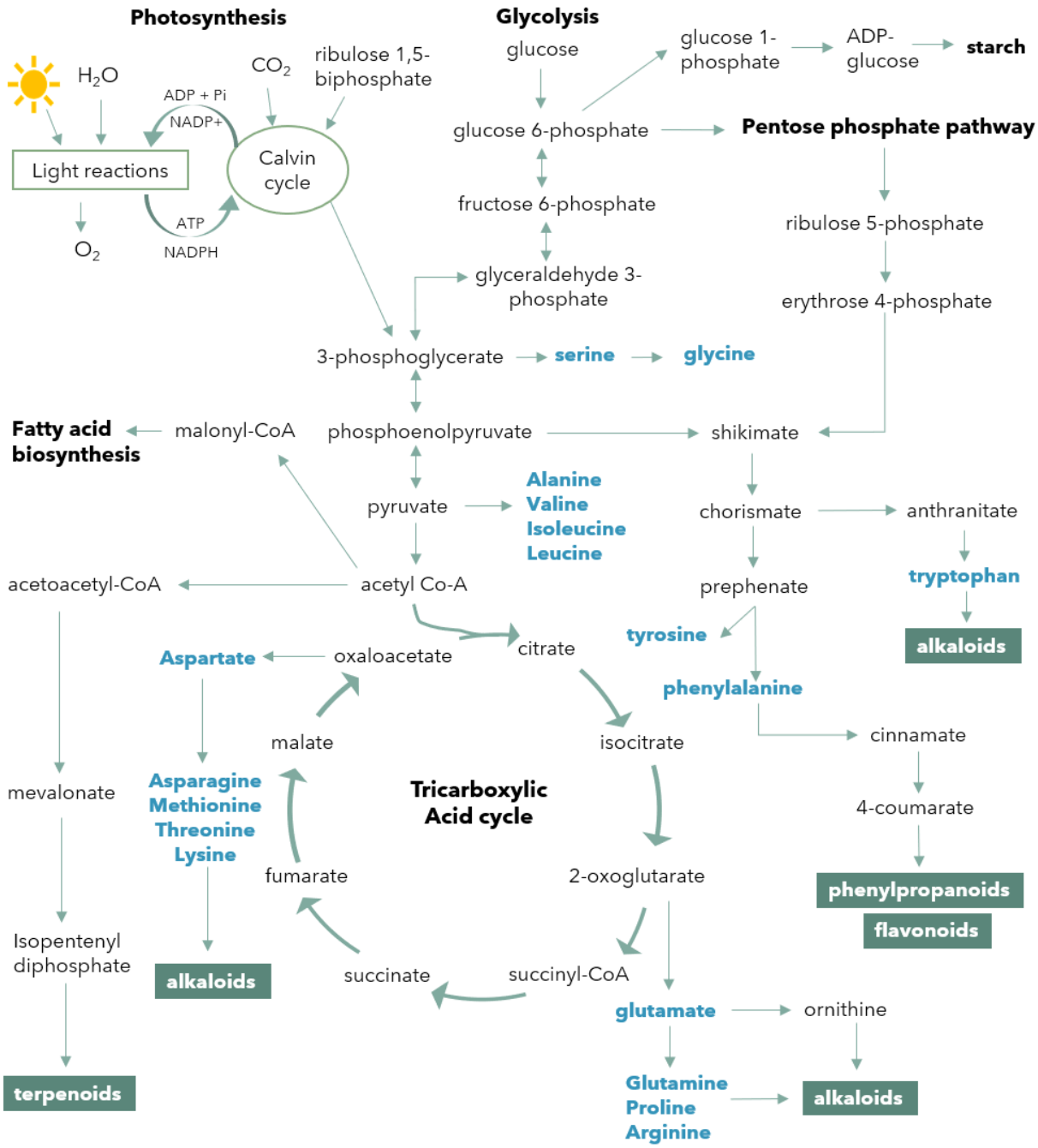


Figure 1: Primary and secondary metabolism in plants. During photosynthesis, light energy is converted into chemical energy and stored in sugar molecules. These molecules are broken down through glycolysis and TCA cycle to produce ATP. TCA cycle connects carbohydrate, fat, and protein metabolism. The pentose phosphate pathway oxidises glucose to generate reducing equivalents and the precursors for the biosynthesis of nucleotides and aromatic amino acids. The latter are produced through the shikimate pathway and used to produce phenylpropanoids, flavonoids, and alkaloids. Acetyl Coenzyme A (acetyl Co-A) can enter the TCA cycle, the fatty acid biosynthesis pathways, or the isoprenoid pathway to produce terpenoids.

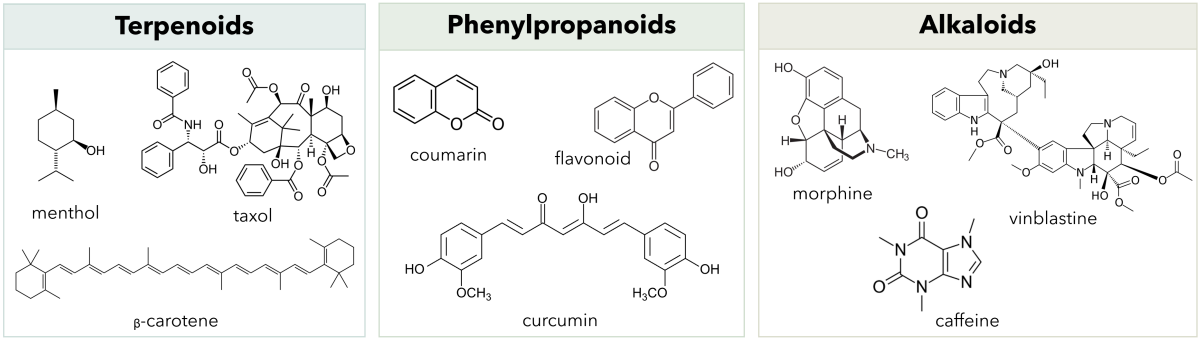


Figure 2: Examples of plant secondary metabolites.

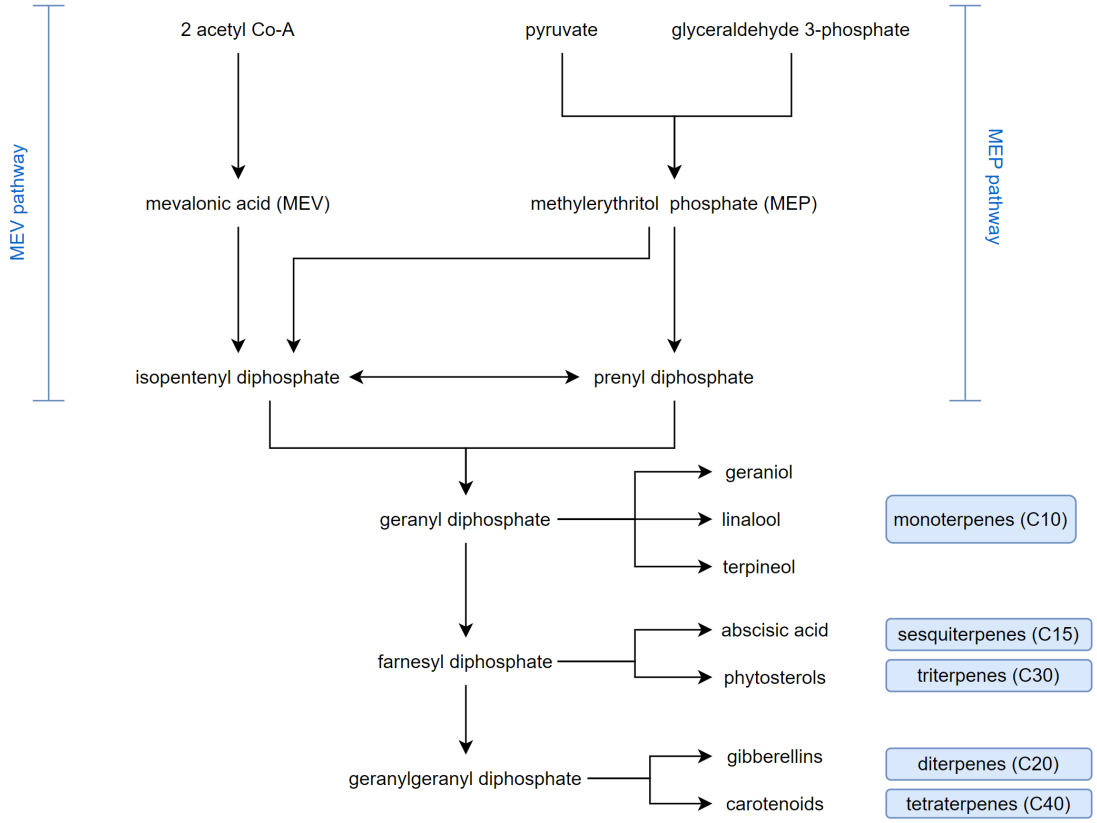


Figure 3: Simplified schema of the terpenoid biosynthesis pathway. IPP and prenyl diphosphate are the precursors for terpenoids and can be produced from mevalonic acid (MEV) or methylerythritol phosphate (MEP) pathways. These originate all terpenoids including monoterpenes, sesquiterpenes, triterpenes, diterpenes, and tetraterpenes.

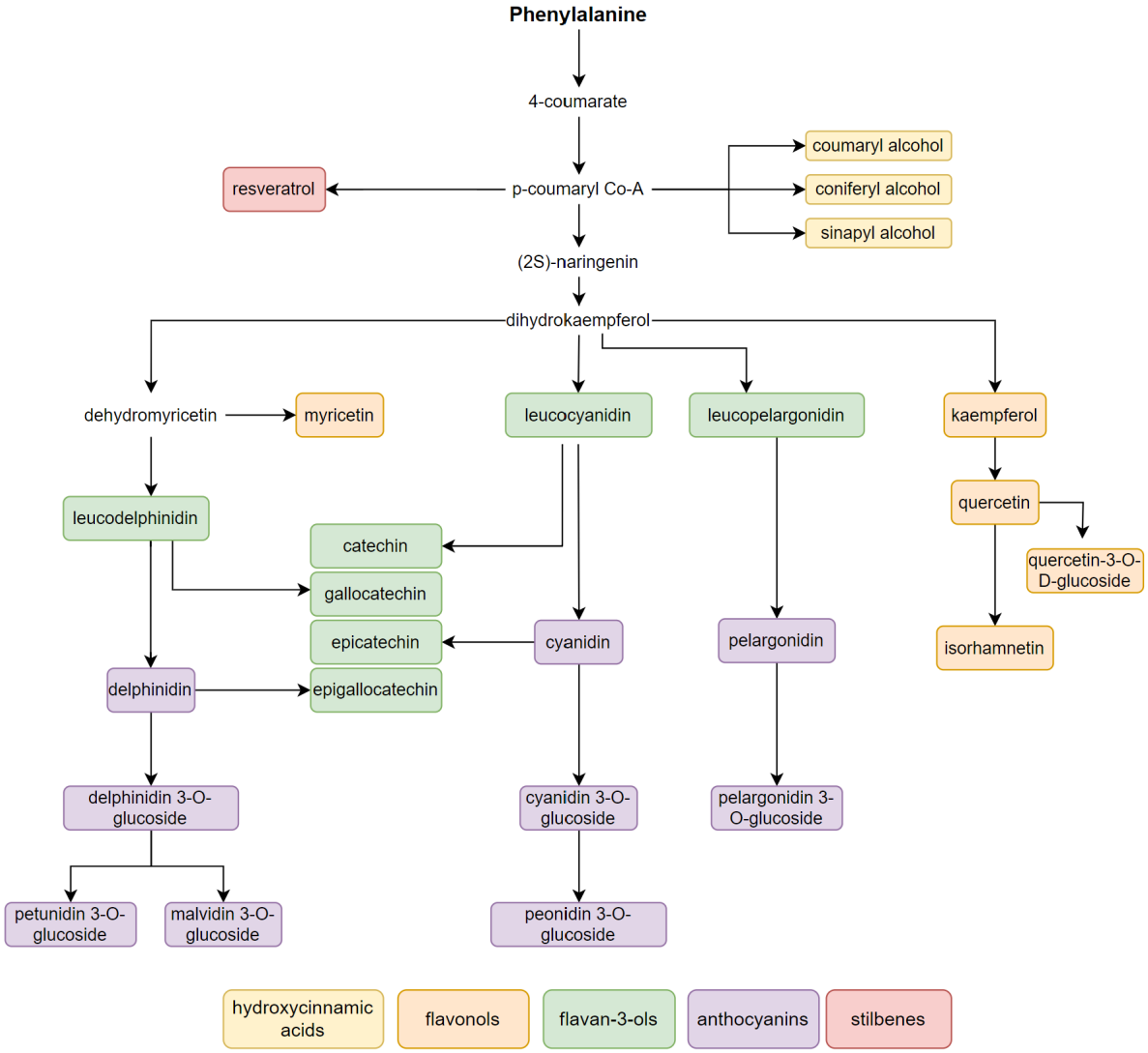


Figure 4: Simplified schema of the production of the main secondary metabolites. This illustrates the phenylpropanoid pathway which starts with the amino acid phenylalanine that is converted to p-Coumaryl Co-A. This metabolite can be used to produce hydroxycinnamic acids and stilbenes, such as resveratrol, or to start the flavonoid biosynthesis pathway to produce different types of flavonoids, such as flavonols, flavan-3-ols, and anthocyanins. Compounds are coloured based on the compound class they belong to.

should be provided in a standard format, such as Systems Biology Markup Language (SBML) [21].

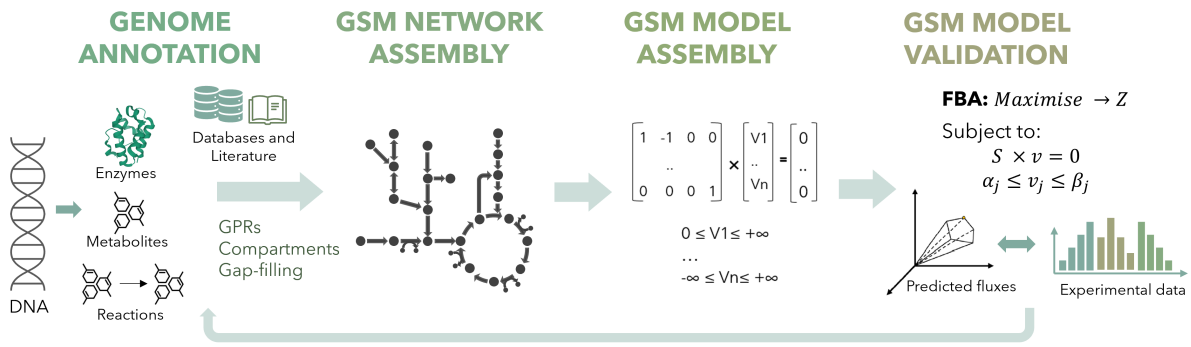


Figure 5: The reconstruction of a GSMM involves four main iterative steps: genome annotation, network assembly, model assembly, and model validation.

As the reconstruction of GSMMs is considered a laborious and challenging process, several software programs have been developed for improving and automating its tasks, although manual curation of the resulting GSMMs is always required. These software tools include, for instance, *Metabolic Models Reconstruction Using Genome-Scale Information (merlin)* [22], *CarveMe* [23], *Constraint-Based Reconstruction and Analysis (COBRA) toolbox* [24] and *ModelSEED* [25] implemented in *The Department of Energy Systems Biology Knowledgebase (KBase)* [26]. *merlin* is a framework with a user-friendly interface developed to annotate a genome and build a compartmentalised draft GSMM automatically, as well as to assist in the manual curation efforts [22].

During the reconstruction of GSMMs, different biochemical databases are accessed to obtain up-to-date information on the organism, such as whole-genome sequences, genome annotations, biochemical data of metabolic reactions, and functional information on enzymes. This information will support the development and refinement of the metabolic network. The main databases used for GSMM reconstruction are described in the following section.

2.2.1.2 Resources of plant metabolic data

Although plant metabolic information is found in different generic metabolic databases, such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [27], some databases are specific to plants or even to a particular species. These databases contain descriptions of metabolic pathways, including valuable details about genes, enzymes, reactions, and metabolites. Table 1 describes the most relevant databases for the study of plant metabolism. KEGG [27] and MetaCyc [28] are the most used generic databases for the analysis of metabolic pathways. NCBI [29], UniProt [30], BRaunschweig Enzyme Database (BRENDA) [31], Transporter Classification Database (TCDB) [32] and PubChem [33] are also generic databases used for extracting detailed information on genomes, proteins, enzymes, transporters, and chemical compounds, respectively. PlantCyc [34], Plant Reactome [35] and MetaCrop [36] are databases with metabolic data for several plants species, whereas SolCyc [37] only includes information for the *Solonaceae* family and The Arabidopsis Information Resource (TAIR) [38] is specific for *Arabidopsis thaliana*. Species-specific

databases for plants have been created from MetaCyc and are available at the Plant Metabolic Network (PMN) [34].

2.2.1.3 Step One: Genome Annotation

The first step in a GSMM reconstruction is the genome functional annotation of the organism. High-quality annotations are essential to obtain an accurate metabolic network [20] and can be retrieved from databases, such as NCBI and KEGG. In cases where genome annotations are not available or not reliable, the functional annotation has to be carried out with sequence similarity alignment tools, such as Basic Local Alignment Search Tool (BLAST) [39], DIAMOND [40], or HMMER [41]. These tools search for annotated genes of other organisms that are homologous to the unannotated gene under analysis and assign a statistical score to each match, representing the confidence level in such similarity. According to the best score, the unannotated gene should be assigned with cellular function(s), gene or Open Reading Frame (ORF) name, and Enzyme Commission (EC) and Transport Classification (TC) numbers, when available. Both EC and TC are useful for automatic network reconstruction.

2.2.1.4 Step Two: Network assembly

The second step in a GSMM reconstruction is the assembly of the network of biochemical reactions and involves several tasks. The first task consists of associating metabolic genes to proteins and reactions through Gene-Protein-Reaction (GPR) associations, which are usually represented by AND or OR Boolean rules. GPR associations must be carefully defined, mainly for isoenzymes, which are sets of proteins catalysing the same reaction, promiscuous enzymes, which are proteins that use different substrates, and for protein complexes and lumped reactions, in which the catalysed reactions should be associated to every gene subunit [20].

Table 1: Description of the most relevant databases of plant metabolic data.

Database	Ref.	Description	Data
KEGG	[27]	Generic resource that comprises genomes, metabolic pathways, chemical compounds, diseases, and drugs.	Metabolic Data
MetaCyc	[28]	Comprehensive database of extensively curated metabolic pathways, containing information on reactions, enzymes, genes, and compounds for several organisms.	Curated metabolic data
BioCyc	[28]	Collection of Pathway/Genome databases (PGDBs), each containing the complete genome and predicted metabolic network of an organism.	Organism-specific predicted metabolic data
NCBI	[29]	Online repository containing several databases for genomics and biomedical information and tools for extracting and analyzing the data.	Reviewed and unreviewed sequence data

Continued on next page

Table 1: Continued from previous page

Database	Ref.	Description	Data
UniProt	[30]	Resource for protein sequence and related information, including manually reviewed data (Swiss-Prot) and automatic, non-reviewed protein annotations (TrEMBL).	Curated and predicted protein data
BRENDA	[31]	Main database of manually annotated enzyme functional data, which uses the EC classification system.	Curated enzyme data
TCDB	[32]	Curated database containing information on transport systems from several organisms, including sequence, structure, and function, and uses the TC system to classify transport proteins.	Curated transport data
PubChem	[33]	The largest database of chemical information, including molecular structure, physical properties, and biological activities of compounds.	Unreviewed chemical data
PlantCyc and PMN	[34]	PlantCyc contains more than 1000 curated metabolic pathways, for at least 350 plant species. This database is the centre of PMN, a resource of plant metabolic databases, and is used as a reference to create plant-specific PGDBs. The current version of PMN (15.0) comprises 126 plant-specific metabolic databases, including curated and predicted databases.	Curated and predicted plant metabolic data
PlantReactome	[35]	Manually curated and comparative pathway database for plants, being part of the Gramene, which is a resource for comparative functional genomics [42]. Plant Reactome used *O. sativa* as a reference species to manually curate metabolic and regulatory pathway data for 97 plant species, also providing a suite of tools for the analysis of large-scale omics datasets.	Curated plant metabolic data
MetaCrop	[36]	Repository of detailed and manually curated metabolic information for six major crop plants with agronomic importance. It allows data to be exported automatically for the creation of metabolic models.	Curated plant metabolic data
SolCyc	[37]	Collection of PGDBs for the Solanaceae species, including databases for *S. lycopersicum* (tomato), *Solanum tuberosum* (potato), *Nicotiana tabacum* (tobacco), *Capsicum annuum* (pepper), and *Petunia x hybrida* (petunia).	Curated metabolic data of Solanaceae species

Continued on next page

Table 1: Continued from previous page

Database	Ref.	Description	Data
TAIR	[38]	Database of genetic and molecular data for *A. thaliana*, including genome sequence and gene structures, products, and expression datasets as well as tools for data visualisation and analysis.	Curated genetic and metabolic data of *A. thaliana*

Information on enzymatic reactions can be obtained from online databases, such BRENDA, through the EC number, and transport reactions can be retrieved from the TCDB database through the TC number. After retrieving available information from the databases, the following information should be compiled: gene names or identifiers, protein names, EC and TC numbers and reaction identifiers, reactants and products, equation and stoichiometry. Then, spontaneous and non-enzymatic reactions should be added to finish the draft network assembly. Organism-specific reactions should be sought in the literature, while non-organism specific reactions can be found in generic databases like KEGG.

The next task is the compartmentalisation of the network, which consists of assigning the reactions to the subcellular organelle where they take place. This task is particularly important for more complex organisms, such as plants, as they present a vast number of compartments. Similar reactions occurring in different compartments have to be specified by assigning distinct identifiers to the reactions and metabolites, according to the corresponding compartment [20, 43, 44]. Bioinformatics tools, such as TargetP [45], WolfPSort [46], and LocTree3 [47] are used to predict the subcellular location of enzymes based on their sequence. In addition, information on protein localisation can be found in literature and databases, such as UniProt for all organisms, and SUBcellular location database for *Arabidopsis* proteins (SUBA) [48] and AraPerox [49] for *Arabidopsis*.

Lastly, manual curation and refinement of the network should be performed through the revision of literature and databases, specific to the organism of interest or phylogenetically close organisms [20]. Several issues should be fixed during manual curation, namely the metabolic potential (presence or absence of reactions), presence of generic terms, substrate and cofactor specificity, stoichiometry and mass balance, reaction reversibility and directionality, and gap filling. Gap filling prevents the accumulation of metabolites, associated with missing reactions in the network, known as gaps.

2.2.1.5 Step Three: Model assembly

After assembling the metabolic network and before converting it to a stoichiometric model, a reaction representing the biomass formation must be created. This reaction describes all known biomass components and their contribution to the cell biomass by including them as reactants. Typically, biomass components include macromolecules; thus, reactions for the biosynthesis of these molecules must be included in the network [20, 50]. Equation 2.1 represents the biomass formulation.

$$\sum_{k=1}^{P} C_k.X_k \rightarrow biomass \tag{2.1}$$

where:

- C_k represents the coefficient of a metabolite;
- X_k represents the metabolite.

The biomass reaction is normalised to one gram of biomass; hence, it expresses how many grams of biomass are produced per gram of existing biomass per hour. Therefore, the flux through this reaction represents the growth rate of the organism (h^{-1}) [50]. The detailed biomass composition should be determined experimentally for the organism of interest or estimated from organism-specific literature. When this information is unavailable, the biomass composition of closely related organisms can be used.

The growth-associated energy is also included in the biomass reaction, represented as the ATP molecules needed per gram of biomass produced [50]. Additionally, non-growth energy requirements need to be included in the model and are usually represented by an irreversible ATP hydrolysis reaction. The energy requirements can also be found in published literature or determined by experimental data [20]. The final step in GSMM reconstruction is the conversion of the metabolic network into a stoichiometric matrix. According to the classic principles of chemical engineering, changes in metabolite concentration over time can be represented by ordinary differential equations, as represented in Equation 2.2 for a given metabolite.

$$\frac{dX_i}{d_t} = \sum_{j=1}^{N} S_{ij} * v_j + \mu X, i = 1, ..., M \tag{2.2}$$

where:

- X_i is the concentration of metabolite i;
- V_j is the metabolic flux of reaction j;
- S_{ij} is the stoichiometric coefficient of the metabolite i in reaction j;
- μX is the system growth rate.

The prediction of concentration profiles over time requires the kinetic rates of all reactions. Currently, there is still a lack of kinetics data, which hinders the development of dynamic models for most organisms. A pseudo-steady-state, in which the metabolite concentration is constant throughout time, is assumed to overcome this limitation [20]. This approach allows converting the dynamic equations into a stoichiometric matrix, as shown in Equation 2.3.

$$S * v = 0 \tag{2.3}$$

where:

- S is the stoichiometric matrix where rows represent metabolites and columns represent reactions.
- v is the flux vector.

As the number of reactions is usually higher than the number of metabolites in a metabolic network, the set of linear equations is an underdetermined system, which results in an infinite number of solutions that may satisfy the imposed constraints. Therefore, the solution space must be narrowed down to a set of feasible solutions, named the flux cone of solutions, by adding new constraints to the system [50]. These new constraints are related to the reversibility of reactions and represent the bounds that define the minimum and maximal flux through each reaction. Mathematically, these constraints are described by inequalities, as shown in the inequality 2.4.

$$vmin_j \leq v_j \leq vmax_j, j = 1, ..., N. \tag{2.4}$$

where:

- v_j is the flux vector;
- $vmin_j$ is the lower bound;
- $vmax_j$ is the upper bound.

Reversible reactions should have bounds set as minus and plus infinity, whereas irreversible reactions should have one bound set to zero, depending on the directionality. Figure 6 represents a metabolic network assembly and its conversion to a stoichiometric model.

Lastly, model simulations are compared with experimental data. When not in accordance, the genome annotation, gap filling, and manual curation steps are revised iteratively until the model achieves accurate results. The final model should be saved in the SBML format [21] to perform *in silico* phenotype predictions using approaches like Flux Balance Analysis (FBA) [51].

2.2.2 Phenotype Simulation

The metabolic phenotypes of the organism under different environmental or genetic conditions can be predicted using the reconstructed GSMM and several constraint-based approaches.

FBA is the most popular constrained-based method to simulate GSMMs [51, 52]. This method uses linear programming to solve the system of equations of the GSMM, resulting in an optimal flux solution. Besides the mass balance and flux constraints of the GSMM, FBA requires the definition of a relevant objective function, which represents the maximisation or minimisation of a metabolic flux during the simulation [51]. The linear programming problem can be formulated as follows:

$$maximize \rightarrow Z \tag{2.5}$$

Subject to:

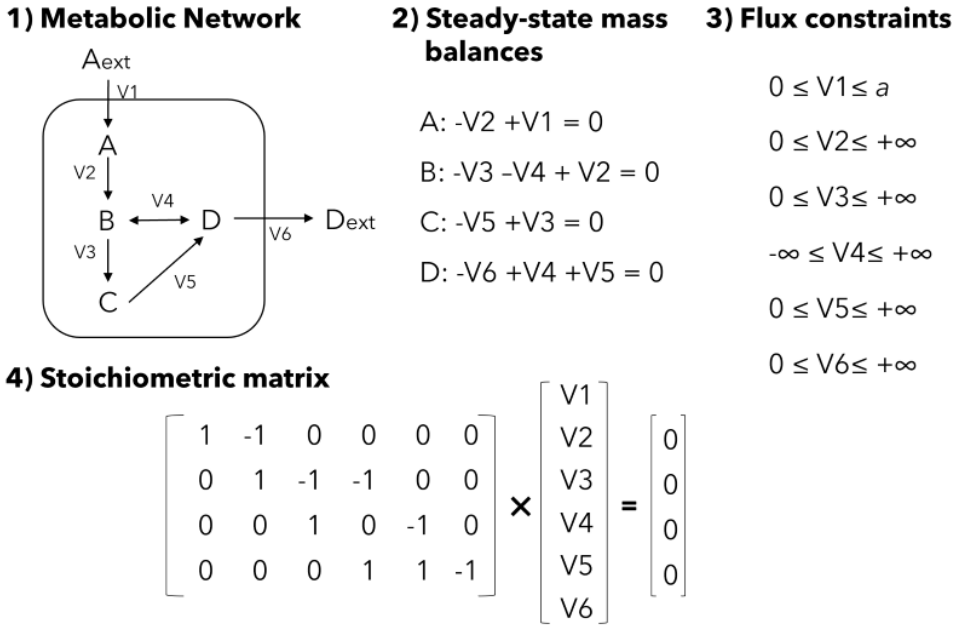


Figure 6: Assembly of a GSMM with four metabolites (A to D) and six fluxes (V1 to v6). The metabolic network and boundaries of a system (e.g. cell) are shown in 1). Fluxes V1 and V6 represent exchange fluxes. Reversible reactions are present with double arrows and irreversible reactions are indicated with a forward arrow. Panel 2) represents the steady-state mass balances and panel 3) shows the constraints for flux values, in which (*"a"*) represents an uptake rate of the metabolite A. Finally, panel 5) shows the representation of the mass balances in the matrix format.

$$\{S * v = 0\}, \tag{2.6}$$

$$\{vmin_j \leq v_j \leq vmax_j, j = 1, ..., N\} \tag{2.7}$$

where:

- Z is the linear objective function;
- v is the flux vector;
- S is the stoichiometric matrix (columns represent reaction fluxes and rows the metabolites mass balances);
- $vmin_j$ and $vmax_j$ are the lower and upper bounds, respectively.

The most commonly used objective function is the maximisation of the biomass growth, as it was demonstrated that most microorganisms have the purpose of growing as fast as possible [50]. However, this objective function may not be appropriate for most multicellular organisms, which are formed by different cell types with specific objectives and functions [15, 44].

FBA can be used to simulate phenotypes with GSMMs using different carbon or nitrogen sources, to identify the best growth medium, to improve the production of a compound of interest, under different gene knockouts, or to prevent the biosynthesis of a toxic metabolite. Likewise, FBA allows the identification

of essential reactions by knocking out a reaction at a time and assessing its impact on biomass production [20, 50].

The genetic manipulations that lead to a metabolic phenotype of interest can be calculated with strain optimisation algorithms, such *OptKnock* [53] and *OptFlux* [54], which use evolutionary algorithms to perform systematic gene knockouts or under/overexpression over the GSMM.

The solution returned by FBA is not unique, meaning that multiple optimal solutions may exist in the solution space for the defined constraints and objective function. This limitation can be overcome with the Parsimonious Flux Balance Analysis (pFBA) method, which selects the flux distribution that minimises the sum of fluxes while maximising growth [55]. Another formulation named Flux Variability Analysis (FVA) is often used to determine the robustness of GSMMs in simulations [56]. FVA calculates the minimum and maximum flux of each reaction for a defined set of constraints.

Two additional optimisation methods, Minimization of Metabolic Adjustment (MOMA) [57] and Regulatory on/off Minimization of Metabolic flux changes (ROOM) [58], have been used to predict metabolic phenotypes of perturbed cells. MOMA uses linear or quadratic optimisation to minimise the effect of the perturbation in the cell by finding the flux distribution closest to the wild-type. ROOM is similar to MOMA, but instead of minimising the total flux value, it minimises the number of altered fluxes.

All these methods have been implemented in several bioinformatics tools such as *COBRA toolbox* [24], *COBRApy* [59] and *ReFramed* [60].

In alternative to FBA, sampling methods, such as Markov chain Monte Carlo (MCMC), can be used to explore the metabolic networks without assuming an objective function by uniformly sampling the feasible flux space. These methods generate a chain of feasible solutions satisfying the network constraints until the flux samples accurately represent the solution space. Flux sampling provides information both on the range and probability distribution of feasible flux solutions [61, 62].

The FBA approach was extended to Dynamic Flux Balance Analysis (dFBA) [63] to allow the modelling of changes in metabolic concentrations over time. dFBA assumes that cells rapidly reach an intracellular steady state after environmental changes. Therefore, intracellular metabolites are still assumed to be at a steady state, but exchange metabolites and total biomass are constrained with dynamic equations representing the rates of uptake or excretion. This framework combines constraint-based modelling with a dynamic simulation of the extracellular environment.

2.2.3 Integration of omics data within genome-scale metabolic models

The development of high-throughput technologies has revolutionised the field of molecular biology by generating large-scale omics datasets, which have allowed the study of the central dogma through the detection and quantification of genes, transcripts, proteins, and metabolites in biological samples. Such massive and heterogeneous data requires bioinformatics tools to analyse and integrate these data, leading to the rise of Systems Biology.

Omics data have been used for a wide range of applications, as they allow the detection and analysis of differential patterns across varied environmental conditions, which can explain metabolic variations

occurring in different phenotypes. Although there are many types of omics data, the studies have mainly focused on genomics, transcriptomics, proteomics, or metabolomics data [64, 65].

In the absence of known regulatory rules of gene expression, omics data, mainly transcriptomics, have been integrated within GSMMs to improve the accuracy of the predictions [66].

2.2.3.1 Resources of plant omics data

Genomics data provide information on genome composition, including the characterisation and quantification of all genes and the identification of single nucleotide polymorphisms, copy number variants, and loss of heterozygosity variants. Transcriptomics data describes the total mRNA in a cell, which reflects the expression of genes at a specific time. These omics data have been mainly generated by microarrays and more recently by next-generation sequencing technologies, such as RNA-Seq [64]. The main databases for genomics and transcriptomics data include Sequence Read Archive (SRA) [67], *GenBank* [68], NCBI's Reference Sequence Database (RefSeq) [69], *Nucleotide* [70] and Gene Expression Omnibus (GEO) [71] from NCBI, DNA DataBank of Japan (DDBJ) [72], European Nucletoide Archive (ENA) [73], *ArrayExpress* [74] and *Expression Atlas* [75], which are described in Table 2.

Regarding plants, Plant Omics Data Center (PODC) [76] was created to integrate large-scale gene expression datasets from mRNA-sequencing. Currently, this resource also provides functional annotations of genes and gene expression networks derived from transcriptomics data of eleven different plants. *PlantExpress* [77] is another plant omics database that builds gene expression networks from microarray data, but only for *O. sativa* and *A. thaliana*.

GRape Expression ATlas (GREAT) is a gene expression atlas that integrates all public RNA-Seq experiments of *V. vinifera* from GEO or Array Express, allowing data analysis and visualisation [78]. In GREAT, the RNA-Seq data is already normalised in transcripts per million (TPM) and the reads were mapped to the new grapevine genome (PN40024.v4) [79].

Proteomics data represent the quantification of expressed proteins in a sample at a specific time. The leading technology used for protein profiling is Mass Spectrometry (MS), though it does not detect all proteins in a sample. The MS output is processed and used for the identification and quantification of proteins and post-translational modifications (PTMs). The final goal is to understand the functional importance of proteins and bridge the gap between the genome and the phenotype [64]. In proteomics data analysis, different databases, such as *ProteomicsDB* [80], PRoteomics IDEntifications (PRIDE) [81], *PeptideAtlas* [82], Global Proteome Machine Database (GBMDB) [83] and Mass Spectrometry Interactive Virtual Environment (MassIVE) [84] (Table 2) are available. Concerning plants, Plant Proteomics Database (PPDB) [85] was created to integrate experimentally identified proteomics data of *Z. mays* and *A. thaliana*.

Metabolomics data reflect the number of small molecules in a sample at a specific condition. MS and Nuclear Magnetic Resonance Spectrometry (NMR) are the leading technologies for quantification of metabolites, but detection cover is lower than for other omics [64]. As the metabolome is the final product of gene expression, its analysis is the closest step to understanding the observed phenotype.

Metabolomics data can be found at *MetaboLights* [86], *MetabolomeExpress* [87], *Metabolomics Workbench* [88] and Golm Metabolome Database (GDM) [89] databases (Table 2).

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

Database	Ref.	Description	Data
SRA	[67]	Archive for next-generation raw sequence data.	Sequences
GenBank	[68]	Comprehensive collection of all publicly available DNA sequences and respective annotations.	Sequences
RefSeq	[69]	A comprehensive, curated, and non-redundant collection of sequences, including genomes, transcripts, and proteins.	Sequences
Nucleotide	[70]	A collection of sequences from different sources including GenBank and RefSeq.	Sequences
GEO	[71]	Repository of functional genomics data, including raw and processed data with descriptive metadata.	Genomics and transcriptomics
DDBJ	[72]	Public database of nucleotide sequences at National Institute of Genetics.	Sequences
ENA	[73]	A comprehensive nucleotide sequence resource, including raw sequencing data, assembly information, and functional annotations.	Sequences
ArrayExpress	[74]	Database of functional genomics data and respective metadata.	Genomics and transcriptomics
Expression Atlas	[75]	A resource for gene and protein expression data for multiple organisms and across different biological conditions.	Transcriptomics
PODC	[76]	Database of mRNA-sequencing expression data for plants.	Transcriptomics
PlantExpress	[77]	Database of gene expression data from microarrays for *O. sativa* and *A. thaliana*.	Transcriptomics
GREAT	[78]	Atlas of RNA-Seq experiments of *V. vinifera* mapped against the PN40024.v4 genome.	Transcriptomics
ProteomicsDB	[80]	Database for quantitative Mass Spectrometry (MS)-based proteomics data. Currently, it also includes RNA-Seq expression datasets, drug-target interactions, and protein turnover data.	Proteomics
PRIDE	[81]	Repository of MS-based proteomics data, including protein identification and quantification, post-translational modifications, analysed mass spectra, and technical metadata.	Proteomics

Continued on next page

Table 2: Continued from previous page

Database	Ref.	Description	Data
Peptide Atlas	[82]	Database of peptides identified in MS proteomics experiments. It provides tools for processing and analysing raw MS output data.	Proteomics
GBMDB	[83]	Database for analysis, validation, and storage of MS proteomics data.	Proteomics
MassIVE	[84]	Community resource for raw MS data, including proteomics datasets.	Proteomics
PPDB	[85]	Database for integrating MS-based proteomics data of *Z. mays* and *A. thaliana*.	Proteomics
MetaboLights	[86]	Repository for metabolomics data and associated metadata, covering metabolite structures, reference spectra, concentrations, and functions.	Metabolomics
MetabolomeExpress	[87]	Online server for processing, interpreting, and storing MS metabolomics data	Metabolomics
Metabolomics Workbench	[88]	Repository for metabolomics data and associated metadata from MS and NMR studies.	Metabolomics
GDM	[89]	Collection of reference mass spectra and retention times for metabolites.	Metabolomics

2.2.3.2 Methods for integrating omics with genome-scale metabolic models

Several methods for integrating omics data within GSMMs have been developed in the last decades with two different designs. The first includes methods that use omics data to improve metabolic flux predictions by constraining the solution space, whereas the second comprises methods that derive context-specific models, which are reaction subsets of the original GSMM, based on the fact that metabolic fluxes change across different environmental conditions, developmental stages, and different cell types or tissues [90, 91].

Most of these methods can be broadly grouped into three main classes, GIMME-, iMAT- and MBA-like algorithms, according to their mathematical formulations. The first class uses omics data to determine the set of active and inactive genes in a specific condition while ensuring a given Required Metabolic Function (RMF). In contrast, the second class classifies reactions as active or inactive based on gene or protein expression data and maximises the agreement between the predicted fluxes and such classification, without establishing any RMF. Finally, the third class divides the reactions into two sets, core, and non-core, and iteratively removes non-core reactions [90].

Furthermore, these methods may be classified as absolute or relative, and as discrete or continuous, according to the survey of Machado *et al.* [91]. Methods are classified as relative if they use differential expression values between conditions to constrain the model or as absolute if they use absolute values

for a single condition. Moreover, methods are classified as discrete or continuous if they constrain the model using discrete or continuous expression levels, respectively. All these methods are detailed in the next subsections and organised by categories in Figure 7.

GIMME-like algorithms

GIMME-like methods define the set of active and inactive genes in a condition, according to omics data, while testing a set of RMF, and use this information to set the bounds on associated reaction fluxes [90]. These methods are further divided into two sub-categories: switch- and valve-based methods. In the switch-based methods, the upper bound of the inactive reactions is set to zero while in the valve-based methods, the upper bound of reactions is set proportional to the normalised expression of the genes encoding the associated enzymes [92]. The switch-based methods include the method of Akesson *et al.* [93], Gene Inactivity Moderated by Metabolism and Expression (GIMME) [94] and its extension, Gene Inactivation Moderated by Metabolism, Metabolomics and Expression (GIM3E) [95].

Akesson *et al.* [93] developed the first switch-based method for improving flux predictions, in which reactions whose associated genes were expressed at low levels in a given condition were deactivated by setting their flux bounds to zero. However, as omics technologies have low sensitivity, some reactions could be incorrectly disabled. Akesson *et al.* manually analysed the inactive gene set and re-enabled specific genes, which they believed to be false negatives to overcome this problem.

In contrast, GIMME [94] is a more advanced algorithm that semi-automatically identifies potential false negatives by minimising the inconsistency between gene expression data and the predicted flux distribution. When the model fails to simulate a phenotype, GIMME re-enables genes, favouring those with a low expression value and associated with low flux reactions. This method generates flux predictions and builds context-specific metabolic models. Although initially developed to integrate transcriptomics data, it has also been used to incorporate proteomics (GIMMEp) [96].

GIM3E [95] is an extension of GIMME and integrates both gene expression and metabolomics data with metabolic models. GIM3E uses metabolomics by adding turnover (production or consumption) metabolites and a respective sink reaction to the initial model. Then, the algorithm forces flux through these sink reactions to ensure that the network in the final models uses the experimentally detected metabolites.

The valve-based methods include the following:

- E-flux [97], which was the first method to use continuous variables to constrain an FBA model. It sets the maximum flux through each reaction as a function of the normalised expression of the associated genes. As a disadvantage, this method can result in multiple optimal solutions, hindering the interpretation of flux predictions. E-flux was extended to E-flux2, by adding the minimisation of the Euclidean norm of the measured fluxes, which generates a unique solution [98] to overcome this problem;

- MEabolic and TRanscriptomics ADaptation Estimator (METRADE) [99] that adds gene expression

and codon usage data to an FBA model by setting the reaction flux bounds as a continuous logarithmic function of the related gene expression levels, and limiting total usage of a codon by the abundance of the corresponding tRNA. It performs multi-objective optimisation to evaluate bacterial metabolic responses to environmental conditions;

- Gene-Expression FBA (GX-FBA) [100] that directly integrates continuous gene expression levels into FBA. However, instead of using absolute values, this method uses relative gene expressions between reference and perturbed conditions. It first predicts the fluxes for the reference state and then the relative expression levels are used to define flux constraints and a new objective function for simulating the perturbed state;

- The method of Fang *et al.* [101] is similar to GX-FBA as it also relies on the differential gene expression between a reference and a perturbed state of interest. This method imposes a set of constraints, reflecting the effects of altered gene expression on the metabolic network, which allows the recalculation of the flux distribution for the perturbed state. It also considers the variability in biomass composition across different growth conditions;

- Probabilistic Regulation Of Metabolism (PROM) [102], which integrates both gene expression data and transcriptional regulatory networks (TRN) within metabolic models. This method calculates the probability of interaction between a transcription factor and its target genes, and this probability is used to restrict the maximum flux for the associated reaction.

iMAT-like algorithms

These methods integrate omics data to improve flux predictions and create context-specific metabolic models without defining a biological objective function [90]. This group comprises the methods Integrated Metabolic Analysis Tool (iMAT) [103], Integrative Network Inference for Tissues (iNIT) [104] and its extension Task-driven INIT (tINIT) [105].

iMAT [103] allows the integration of transcriptomics and proteomics data within GSMMs, based on the method described by Shlomi *et al* [106]. This method divides genes into groups of expression: high, moderate, and low. Then, it maximises the presence of reactions whose predicted fluxes are consistent with this division. The number of reactions associated with highly expressed genes is maximised, while minimising the number of reactions associated with lowly expressed genes.

A similar method named iNIT [104] can handle proteomics or transcriptomics data. This method sets positive weights for reactions whose genes are highly expressed and negative weights for those whose genes have low expression levels. INIT can also incorporate metabolomics data to generate context-specific models. If there is experimental evidence (metabolomics data) that a metabolite is present in a specific cell type, this method ensures that all reactions required for its production are included in the final model.

For multi-cellular organisms, as normal tissues may not have growth as the main objective, it is not easy to define a relevant objective function. Therefore, iMAT and iNIT are suitable for creating tissue-specific models by not imposing a set of RMF [90, 92].

tINIT [105] is an extension of the iNIT algorithm, which enables the definition of metabolic objectives to be satisfied by the context-specific metabolic model.

MBA-like algorithms

The MBA-like methods are also independent of RMF but, unlike the previous group, they only provide a context-specific model reconstruction and not a flux distribution. These methods define two reaction sets, core and non-core, for a specific tissue. The core set is composed of reactions more likely to be active, according to omics data or curated information, while the non-core includes reactions with little or no evidence of being active in a specific tissue. Then, the methods test each non-core reaction to assess which are required to keep all core reactions active and should be included in the final model [90]. This group includes the Model Building Algorithm (MBA) [107], Metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE) [108], Fast Reconstruction of Compact Context-specific Metabolic Network Model (FastCORE) FastCORE [109] and Cost Optimization Reaction Dependency Assessment (CORDA) [110] methods.

In MBA [107], the non-core reactions are randomly removed while testing the consistency of the resulting model. As the order by which the reactions are tested can affect the results, the algorithm runs multiple times, resulting in a set of models. Besides core reactions, the final model must include the reactions present in most candidate models.

The mCADRE algorithm [108] is similar to MBA; however, non-core reactions are ranked according to expression data and connectivity in the metabolic network. Then, reactions are removed sequentially, starting with the one with the lowest ranking. Therefore, mCADRE only reconstructs a single model, significantly reducing computation time.

The FastCORE algorithm [109] operates with two linear programs to search for a consistent subnetwork, containing all reactions from the core and a minimal set of non-core reactions. The problems are solved alternatively at each iteration until the core network is consistent. This algorithm has several applications, as it enables the integration of multiple omics data into a single model. FastCORE was adapted by a workflow named FASTCORMICS [111], which facilitates the integration of transcriptomics data from microarrays, by preprocessing the expression data, generating multiple specific models.

Lastly, CORDA [110] is a recent algorithm that identifies the interdependence of core on non-core reactions, a process named dependency assessment. Each reaction will produce a pseudo-metabolite with a specific cost, assigning a higher cost to non-core reactions. Then, the method performs an FBA, minimising the flux through high-cost reactions.

Other algorithms

Less popular methods for integrating omics data within models did not fit into the categories above [91, 92] and are described in Table 3.

Table 3: Additional methods for the integration of omics data within metabolic models.

Method	Ref.	Description
MADE	[112]	Metabolic Adjustment by Differential Expression (MADE) compares expression data from sequential conditions without the need for subjective thresholds by identifying statistically significant changes in gene expression for each condition. Then, it generates a model reflecting the expression patterns across sequential conditions.
tFBA	[113]	Transcriptionally Controlled FBA (tFBA) is similar to MADE as it also avoids the definition of arbitrary thresholds to identify active and inactive genes. This method adds soft constraints to FBA representing up- and down-regulation cases, according to gene expression data from different conditions. The objective is to find a set of flux distributions across all conditions, which minimises the violation of the up/down constraints.
FCGs	[114]	It uses Flux-Coupled Genes (FCGs) as additional constraints for metabolic simulations. FCGs are defined as genes whose expression patterns upon perturbations are consistent with the respective fluxes. FCGs are used to constrain the fluxes of the corresponding reactions in new conditions, according to changes in their expression levels.
TEAM	[115]	Temporal Expression-based Analysis of Metabolism (TEAM) combines dFBA with the GIMME algorithm and integrates temporal gene expression patterns to predict time course flux distributions.
AdaM	[116]	Adaptation of Metabolism (AdaM) identifies, for each time step, the minimal operating network that is consistent with changes in the expression levels between time steps. Then, instead of calculating flux distributions, it calculates the corresponding set of Elementary Flux Modes (EFMs), which represent the smallest active sub-networks in a steady-state metabolic network. At each time, the reactions are weighted according to the fraction of EFMs that includes them.
Lee *et al.*	[117]	Like the E-flux valve-based method, this also integrates absolute gene expression data into metabolic models but not as flux constraints. Instead, this method offers an alternative objective function that consists of maximising the correlation between the predicted flux distribution and gene expression data.
FALCON	[118]	Flux Assignment with Least absolute deviation Convex Objectives and Normalization (FALCON) is based on the method of Lee *et al.*, but it initially estimates enzyme abundances using the GPR associations in the model and gene expression data to improve model predictions.

Continued on next page

Table 3: Continued from previous page

Method	Ref.	Description
RegrEx	[119]	Regularised Context-specific model Extraction (RegrEx) is a fully automated and unbiased approach to extract tissue-specific models and generate a flux distribution that maximises its correlation to transcriptomics data. Similar to E-flux2, this method is based on context-specific data and regularised least-squares optimisation, not requiring predefined parameters or biological functions nor preclassification of reactions. This approach is advantageous to extract context-specific models for complex or not well-known organisms.
Moxley *et al.*	[120]	It introduces a new parameter defined as metabolite interaction density, which represents, for each pathway, the level of pairwise metabolite-protein interactions. The method predicts flux changes as a function of the variations in gene expression levels while considering metabolite interaction densities. This approach is based on the hypothesis that flux changes in pathways with less interaction density are expected to be more correlated with changes in gene expression.
RELATCH	[121]	RELATive CHange (RELATCH) uses transcriptomics data and predicted fluxes from a reference state to calculate the flux distribution for the perturbed state by minimising their distance to the reference distribution. Therefore, this method does not require any gene expression data from the perturbed state.
EXAMO	[122]	EXploration of Alternative Metabolic Optima (EXAMO) extends the iMAT method by exploring alternative optimal solutions with the same score. This method calculates the frequency of reactions in the solutions to identify reactions that are active in all solutions (high-frequency reactions) and reactions that are always inactive. Then, a context-specific model is built by removing the identified inactive reactions and using MBA to minimise the number of reactions in the model, with the constraint that all high-frequency reactions must have flux.
lbFBA	[123]	Linear-bounds FBA (lbFBA) incorporates transcriptomics or proteomics data into metabolic models, by defining soft constraints on reaction fluxes that can be violated. Reaction flux bounds are defined as linear functions of gene/protein expression data, whose parameters are initially estimated from a training dataset, containing measured fluxomics and transcriptomics or proteomics data. lbFBA tries to minimise the sum of total fluxes, as well as the number of soft constraint violations, to predict flux distributions in new conditions.

Continued on next page

Table 3: Continued from previous page

Method	Ref.	Description
GEMsplice	[124]	Genome-scale metabolic modelling with splice-isoform integration (GEMsplice) is the first to incorporate splice-isoform expression data into GSMMs, and was applied to investigate different metabolic phenotypes of breast cancer. This method uses METRADE to integrate splice-isoform information and to define reaction bounds as a function of the expression level of associated genes.
IOMA	[125]	Integrative Omics Metabolic Analysis (IOMA) incorporates relative proteomics, and absolute metabolomics data, to improve flux predictions following genetic alterations, considering a mechanistic model of reaction rates. This method is formulated as a quadratic programming problem that searches for a steady-state flux distribution, where the flux through reactions with omics data is as consistent as possible with the fluxes kinetically estimated via Michaelis Menten-like equations. The method also requires the estimation of missing kinetic parameters in the optimisation problem.
RIPTiDe	[126]	Reaction Inclusion by Parsimony and Transcript Distribution (RIPTiDe) aims to identify the most efficient usage of metabolism that also represents an organism's costs in transcription, using transcript abundances and pFBA. This method assigns continuous weights to reactions that reflect the normalised transcriptional abundance of the associated genes. A lower weight represents a higher transcript abundance and a higher probability of including the reaction in the final model. Then, a pFBA simulation is performed by limiting the inclusion of reactions with higher weights in the solution, and the reactions without flux are removed from the model. Finally, flux sampling is performed to find pathway utilisation patterns.

Most of the described methods are provided as scripts by the authors but in some cases, their application to new data is very difficult. Some methods for reconstruction of tissue-specific models are included in COBRA toolbox (FastCORE, MBA, GIMME, iNIT and iMAT) of the commercial platform MATLAB, or implemented in open-source Python packages, such as Tissue-specific RecOnstruction and Phenotype Prediction using Omics data (Troppo) [127], which facilitate their use.

Machado *et al.* [91] evaluated the aptitude of seven of these methods to improve metabolic flux predictions, using published data from *E. coli* and *S. cerevisiae*, and comparing the results to pFBA predictions. These methods included GIMME, iMAT, MADE, E-flux, RELATCH, GX-FBA and the method by Lee *et al.*. None of them have consistently outperformed each other nor pFBA under all conditions. lbFBA recently reported improvements over pFBA in specific conditions. These results suggest that gene expression levels do not have a strong correlation with reaction fluxes, and other factors, like post-translational modifications and translation efficiency, may impact flux levels. They have also shown that the use of proteomics rather than transcriptomics data did not improve flux predictions, indicating that enzyme levels have a

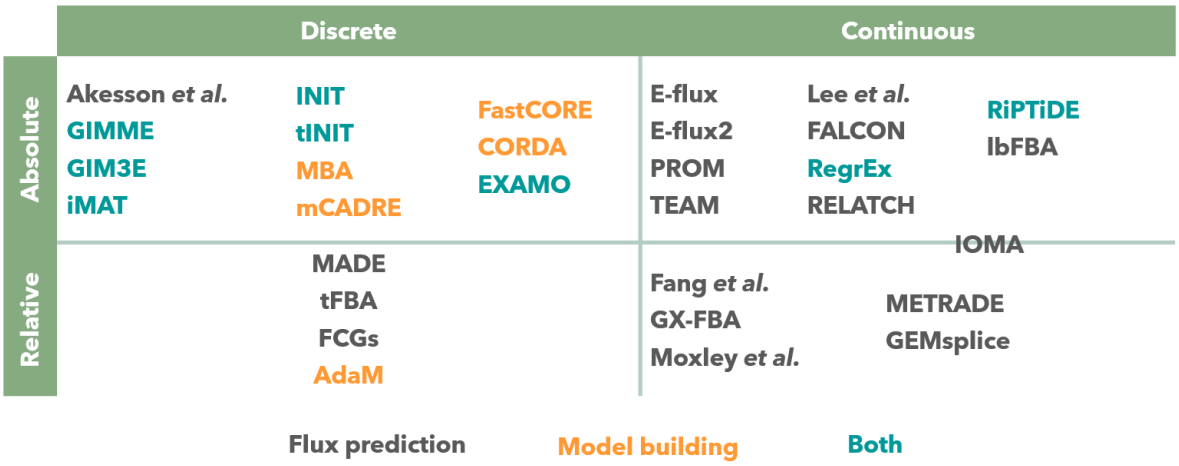


Figure 7: Distribution of methods for integrating omics data within GSMMs. Methods have been classified as absolute or relative whether they use absolute values for a single condition or differential expression values between conditions, respectively. Methods were also divided into discrete or continuous, according to whether they use discrete or continuous expression values to constrain the model. Additionally, they were also divided according to the output they produce: methods in grey only result in a flux prediction, methods in orange only create a specific model, and methods in blue output both a flux prediction and a specific model.

weak correlation with reaction fluxes. Therefore, although there are several methods available, the use of transcriptomics and proteomics data to improve flux predictions is still inaccurate and challenging.

On the other hand, gene expression data have been successfully used to create context-specific metabolic models for multicellular organisms [128]. In this case, gene expression levels are expected to be more correlated with metabolic fluxes, as metabolic pathways can be completely activated or deactivated in specific tissues [91].

2.2.4 Plant genome-scale metabolic models

The first plant GSMM was published in 2009 for *Arabidopsis thaliana* [129]. Several models have been developed since, not just for model plants like *A. thaliana*, but also for more complex and economically important plants [15], such as *Oryza sativa* (rice) [130–133] and *Zea mays* (maize) [134–136].

Figure 8 summarises the plant GSMMs published to date. Generally, these models have proven to be robust and accurately predict specific aspects of central carbon metabolism [15]. The existing plant GSMMs are described below, grouped by organism, and ordered by publication date.

2.2.4.1 *Arabidopsis thaliana*

Poolman et al. [129] reconstructed the first plant GSMM for *A. thaliana* heterotrophic cell suspension culture. This model was mainly derived from the AraCyc database (version 4.5) [137] and produces biomass components in the proportion observed experimentally in heterotrophic suspension cultures. In 2013, Cheung et al. [138] extended this model to include the subcellular localisation of central metabolic reactions across five compartments (cytosol, plastid, mitochondrion, peroxisome, and vacuole) and to

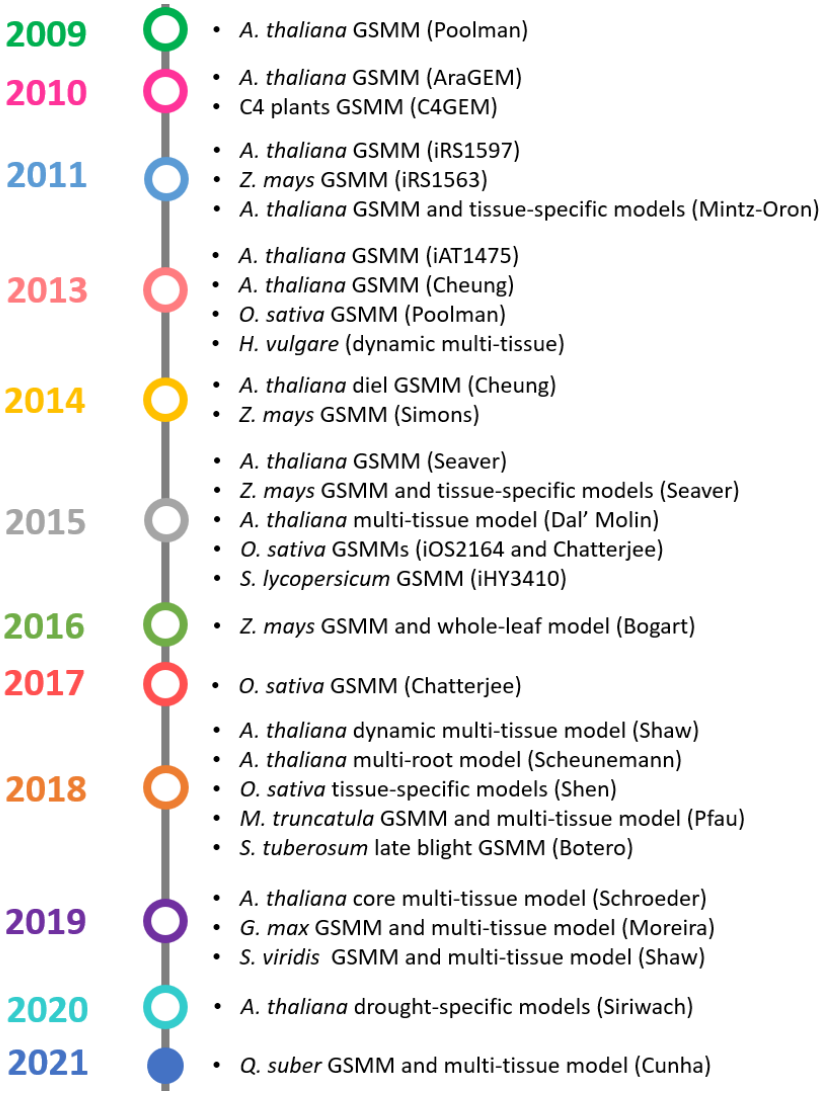


Figure 8: Timeline of the most relevant plant GSMM reconstructions.

account for growth, transport, and cell maintenance energy costs, including ATP and reductive costs. They simulated the model under different environmental conditions and discovered that accounting for energy costs of transport and maintenance substantially improves flux predictions, regardless of the objective function used in the simulation.

AraGEM [139] was the first plant GSMM to represent the metabolism of a compartmentalised photosynthetic cell (same five compartments as in Cheung's model [138]), describing photosynthesis, photorespiration, and respiration while identifying metabolic changes between them. This model was updated by Saha et al. [134] and later by Chung et al. [140] to include terpenoid biosynthesis reactions. Recently, Siriwach et al. [141] have combined the AraGEM model with time-series gene expression data, creating condition-specific models of *A. thaliana* under drought and control conditions to gain insights into the development of tolerant plants.

Mintz-Oron et al. [142] have reconstructed the fully compartmentalised GSMM for *A. thaliana*, which encompasses the subcellular localisation of all reactions, across the five compartments of the AraGEM

model, plus the Golgi Complex and Endoplastmatic Reticulum. They extracted ten tissue-specific models from this generic GSMM by integrating protein expression data of eight tissues and cell cultures in light and dark conditions. The authors then used the seed-specific model to predict the genetic knockouts that result in vitamin E overproduction. Töpfer et al. [143, 144] have combined this generic model with time-resolved transcriptomics data from different temperature and light conditions to understand the metabolic acclimation of *A. thaliana* to stressful environments.

An evidence-based model for *A. thaliana* was reconstructed by Seaver et al. [145] from the generic model available on PlantSEED [146], including the seven compartments mentioned plus the nucleous and the cell wall, and combined with transcriptomics and metabolomics data to extract specific models for eight root tissues at different developmental stages [147]. The tissue-specific models were used to build a multi-tissue model of the root for analysing the flux distribution of hormones indole-3- acetate and trans-zeatin through the root.

A more complex model of *A. thaliana* was developed to represent the leaf metabolism over a day-night cycle [148]. This diel model was reconstructed by duplicating the previous model [138] into two modules, day, and night, and manually adding the transporters between these two phases. The authors simulated both phases in a single optimisation problem by applying constraints specifying that photon influx is allowed during the day (photoautotrophic metabolism) and is set to zero at night (heterotrophic metabolism). This model simulates and clarifies the interactions between the two phases by allowing storage metabolites synthesised during the day to be used at night and vice-versa.

More recently, multi-tissue models for *A. thaliana* were reconstructed to represent different tissues and their interactions. Dal'Molin et al. [149] have developed a framework to create a multi-tissue model comprising the leaf, stem, and root of *A. thaliana* across the diurnal cycle. In this approach, the tissues exchange metabolites through a shared compartment (common pool) rather than directly transported between two tissues, which can reduce redundancy when more than two tissues are interconnected. Additionally, a storage pool manages the storage and retrieval of metabolites. Therefore, in this framework, a multi-tissue model is defined by a stoichiometric matrix representing the internal reactions and three matrices for the exchange reactions with the environment, the transport reactions through the common pool, and the accumulation of metabolites in the storage pool. The multi-tissue model consisted of three replicates of the AraGEM model (representing root, stem, and leaf) and two common pools, one for exchanges between leaf and stem and another for exchanges between stem and root. To simulate the diurnal cycle, the multi-tissue model was duplicated to represent each state (light and dark), and a storage pool was created, with starch being the only stored metabolite. The model was used to study carbon and nitrogen translocation between tissues.

Following this strategy, the diel model was used to build a dynamic multi-tissue diel model (Figure 9) to study metabolic changes across multiple growth stages under different nutrient availability [150]. All reactions of the diel GSMM were replicated to represent the leaf and root model, and the transport between root and leaf was performed through a common pool representing the phloem. This multi-tissue model was simulated by dFBA [63] to explore carbon and nitrogen partitioning between root and leaf over different developmental stages.

A different approach was followed by Schroeder et al. [151] to study the evolution of metabolism across the lifecycle of *A. thaliana*. While previous studies have only considered metabolism at a single point (growth or a single diurnal cycle), this optimisation framework takes a series of "snapshots" of core-carbon metabolism. These snapshots comprise the plant mass, growth rate, and fluxes at one-hour intervals across 61 days of growth, including the stages of seed germination, leaf development, flower production, and silique ripening. In this study, a core multi-tissue metabolic model (referred to as p-ath780) comprising leaf, root, seed, and stem tissues was reconstructed. The core model only includes the central metabolic pathways of *A. thaliana*. The tissue-specific models were built based on the available literature and experimental studies and then merged by OptCom, a framework for modelling microbial communities [152]. The novelty of this method is to simultaneously consider the diurnal cycle, carbohydrate storage, maintenance, and senescence costs, and changes in tissue and whole-plant mass during growth, according to experimental data.

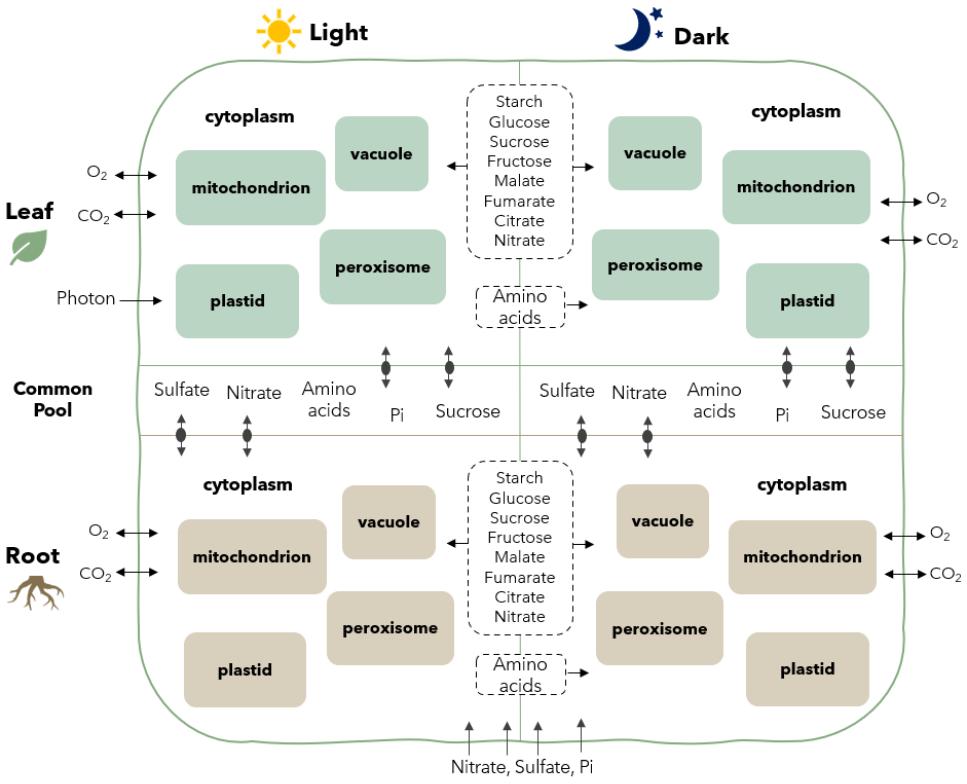


Figure 9: Schematic overview of the diel multi-tissue model of *A. thaliana*, including the leaf and root tissues and the common pool in both light and dark phases [150]. Each tissue module includes five compartments: cytoplasm, mitochondria, vacuole, plastid, and peroxisome. Starch, glucose, sucrose, fructose, malate, fumarate, citrate, and nitrate can accumulate in the light and dark phases of the leaf and root (dashed rectangle between phases). Amino acids can be stored in the light and used in the dark phase. The exchange of amino acids, sucrose, sulfate, nitrate, and phosphate (Pi) was allowed between leaf and root through a common pool using proton pumps. Photon uptake was allowed through the leaf in the light phase while mineral nutrients, such as nitrate, sulfate, and Pi, were allowed through the root in both phases. Exchanges of carbon dioxide and oxygen were allowed through the leaf and root in both phases.

2.2.4.2 *Zea mays*

Dal'Molin et al. made the first efforts towards reconstructing a *Z. mays* GSMM, combining biochemical information of *Sorghum bicolor* (sorghum), *Z. mays*, and *Saccharum officinarum* to build the C4GEM model [153]. This model represents the two leaf tissues, mesophyll (M) and bundle sheath (BS) cells, where photosynthesis of C4 plants takes place, and the interactions between them. It also includes the main five compartments (cytosol, plastid, mitochondrion, peroxisome, and vacuole). Since then, three more GSMMs were reconstructed for *Z. mays* leaf. The first model, referred to as iRS1563 [134], was based on AraGEM (with the same five compartments) and *Z. mays* genome and was used to predict metabolic phenotypes for two natural brown midribs' (bm) mutants with defective lignin biosynthesis.

Another *Z. mays* leaf model was reconstructed by Simons et al. [135], with a significant increase in the number of genes and reactions, representing secondary metabolism. It comprises the two tissues of the leaf, M and BS cells, and gene expression data was used to identify the active reactions in each tissue. Regarding compartments, it includes the five of C4GEM plus the plasmatic, thylakoid, and inner mitochondrial membranes. This model was used to assess the assimilation of nitrogen within the leaf under different nitrogen conditions and later was constrained by incorporating enzyme activity data to detect metabolic differences between nineteen *Z. mays* lines [154].

Seaver et al. [145] have also reconstructed an evidence-based model for *Z. mays*, with the same compartments as the *A. thaliana* model and it was used to extract tissue-specific models for the leaf, embryo, and endosperm of *Z. mays* by integrating gene expression data within the model.

The most recent *Z. mays* leaf GSMM (iRB5204) was developed by Bogart et al. [136], based on the CornCyc database (version 4.0) [155] and previous models. They reconstructed a high-confidence model named iRB2140, including only curated reactions and the same compartments of Simons' model [135], except for the plasmatic membrane. Then, the authors created a two-tissue model to represent M and BS cells of the leaf by duplicating the iRB2140 model and adding transport reactions between the two tissues (Figure 10). They incorporated known nonlinear kinetic constraints and transcriptomics data from more and less differentiated cells to reconstruct a whole-leaf model identified as iEB2140x2x15, representing 15 developmental stages of the maize leaf.

2.2.4.3 *Oryza sativa*

The first GSMM was reconstructed from the RiceCyc database [156] by Poolman et al. [130] and included three compartments: cytosol, chloroplast, and mitochondrion. It was analysed to identify metabolic responses to different light intensities. This model was curated and extended by Chatterjee et al. [131] to encompass the peroxisome compartment and reactions involved in chlorophyll synthesis.

Another *O. sativa* leaf model named iOS2164 was developed by Lakshmanan et al. [133] by adding the vacuole, the endoplasmic reticulum, and the thylakoid as compartments, and all electron-transport reactions. The authors have integrated transcriptomics data within this model to evaluate the metabolic responses to different light conditions. Later, this model was combined with gene expression data of different tissues at different developmental stages to generate tissue-specific models and highlight the

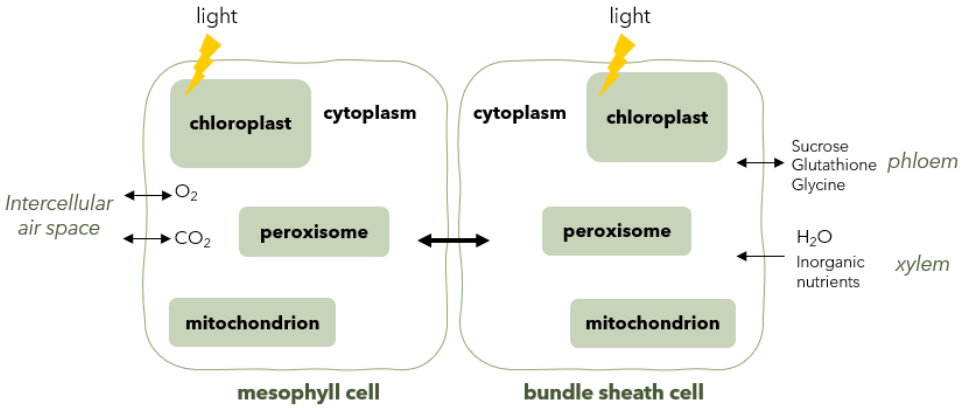


Figure 10: Schematic overview of the two-tissue model of *Z. mays*, representing the M and BS leaf cells of C4 plants [136]. Each cell includes four compartments: cytoplasm, chloroplast, peroxisome, and mitochondrion, and small molecules are directly exchanged between the two tissues by transport reactions. The M cells exchange carbon dioxide and oxygen with the intercellular air space while BS cells exchange sucrose, glutathione, and glycine with phloem and import water and inorganic nutrients from the xylem. This type of model representation is used to understand photosynthesis in C4 plants, mainly the interactions between M and BS cells.

metabolic differences between tissues [157]. All these rice models describing the metabolism of *O. sativa japonica* were reviewed in [158]. Chatterjee et al. [132] reconstructed a GSMM for *O. sativa indica*, which included cytosol, mitochondrion, peroxisome, and chloroplast compartments, and used this model to characterise the metabolic responses to variations in RubisCO activity and light intensity under different enzymatic cost constraints.

2.2.4.4 Other organisms

Other organisms, like *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Medicago truncatula* (barrelclover), *Glycine max* (soybean), *Setaria viridis* (green foxtail) and *Quercus suber* (Cork oak), have only one GSMM available. For *Hordeum vulgare* (barley), a dynamic multi-tissue model was created and stoichiometric multi-tissue models were also created from the GSMMs of *M. truncatula*, *G. max*, *S. viridis*, and *Q. suber*.

- ***Hordeum vulgare***. A framework for analysing metabolic dynamics of *H. vulgare* on a whole-plant scale was developed by integrating a steady-state multi-organ model with dynamic constraints from a functional plant model [159]. Organ-specific models for leaf, stem, and seed were reconstructed by collecting primary metabolism data from literature and databases and combined into one multi-organ model. Next, dFBA was applied and exchange fluxes predicted by the functional plant model were used to constrain FBA at each time interval. This framework allowed studying metabolic interactions between the source and sink organs of *H. vulgare*, accounting for temporal and environmental changes.

- ***Solanum lycopersicum***. The only model of *S. lycopersicum*, referred to as iHY3410, represents the leaf and enables the analysis of metabolic flux distributions on photorespiration pathways under

drought stress [160].

- ***Solanum tuberosum***. Botero et al. [161] reconstructed a GSMM of *S. tuberosum* late blight to study the effect of this disease on leaf metabolism, suggesting the suppression of photosynthesis. This model encompasses the metabolic pathways of the leaf and the interaction between the plant and *Phytophthora infestans* through the integration of gene expression data of infected *S. tuberosum*.

- ***Medicago truncatula***. A fully compartmentalised model for *M. truncatula* was developed by Pfau et al. [162] and allowed the analysis of their rhizobial symbiosis for nitrogen fixation by connecting the plant model to a model of its symbiont and evaluating the effects of the symbiosis in plant growth. Then, a multi-tissue model representing the root and shoot of *M. truncatula* was reconstructed by integrating tissue-specific gene expression data and connecting the resulting root- and shoot-specific model with a combined biomass reaction and inter-tissue transporters derived from literature.

- ***Glycine max***. Moreira et al. [163] reconstructed a GSMM of *G. max* and duplicated this model to create a multi-tissue model representing two tissues of G. max seedlings: the cotyledons and hypocotyl/root axis. The multi-tissue model was constrained with the biomass compositions observed experimentally over four days of seedling growth to simulate the mobilisation of seed reserves during this period and detect metabolic differences between the two tissues, as well as interactions between them.

- ***Setaria viridis***. Similarly, a model of *S. viridis* [164] was reconstructed and used to create a multi-tissue model representing the C4 leaf (including both M and BS cell types) and stem. These models have identified implications of proton balancing on flux distributions during photosynthesis of C4 leaves and reactions involved in the biosynthesis of cellulose, hemicellulose, and lignin in the stem.

- ***Quercus suber***. Recently, a reconciled GSMM for *Q. suber* was semi-automatically reconstructed by Cunha et al. [165] using *merlin* [166] and performing extensive manual curation. This is the first model reconstructed for a woody tree. Transcriptomics data was integrated with the model to obtain tissue-specific models for the leaf, inner bark, and phellogen, which were merged into a diel multi-tissue model to predict interactions among tissues at light and dark phases and study the synthesis of suberin monomers. In the future, this model can be extended and used to explore the metabolic patterns associated with high-quality cork, which is economically relevant for Portugal.

Overall, the most significant advances in plant metabolic modelling were made first for *A. thaliana*. AraGEM [139] is the most used plant model and was the first to allow the simulation of photosynthesis and photorespiration metabolic processes, including compartmentalization. Next, a relevant advance was the diel model of Cheung et al. [148], which allows the simulation of the leaf metabolism over the diurnal cycle in a single problem. Then, the creation of the multi-tissue model by Dal'Molin et al. [149] represented a significant improvement of these models. This model allows the analysis of the metabolism

across different tissues (leaf, stem, and root) and also the metabolic interactions between them. Although the multi-tissue model of Shaw et al. [150] only comprises two tissues, leaf, and root, its novelty was to include dynamic constraints for the exchange metabolites.

Finally, it is important to highlight the integration of omics into models to create tissue- or condition-specific models to originate more realistic flux predictions. Although transcriptomics data are usually used to reconstruct specific models [141, 143, 144, 147], Mintz-Oron et al. have constrained the models with proteomics [142]. Regarding *Z. mays*, the models helped study the photosynthesis of C4 plants, which occurs between M and BS cells. Of these models, the one from Simons et al. [135] stands out as it includes more secondary metabolism reactions, more compartments, and constraints based on gene expression. Likewise, iOS2164 [133] is the most complete model for *O. sativa*, as it contains more compartments and transcriptomics-based constraints. Meanwhile, the advances made for *A. thaliana* were applied in the reconstruction of complex multi-tissue models for other organisms, such as *M. truncatula* [162], *G. max* [163], *S. viridis* [164], and a woody tree, *Q. suber* [165].

2.2.5 Major challenges and limitations in plant metabolic modelling

Although several plant GSMMs and studies that successfully use them to understand plant metabolic processes are available, the existing approaches still have limitations as plant metabolic modelling is very challenging [15].

Annotation of plant genomes is incomplete, and database information on plant enzymatic reactions and metabolites is limited, especially for secondary metabolism, resulting in an inaccurate model with network gaps, requiring extensive and time-consuming validation. Most plant metabolic models have been validated to predict changes in plant central carbon metabolism, though generally neglecting secondary metabolism. Therefore, these models cannot correctly predict plant adaptation to the environment and interactions with pathogens [44]. Exceptions are the models from *Z. mays* [135] and *G. max*, which present extensive coverage of the secondary metabolism.

Another challenge in plant modelling is to place reactions in the correct compartment. Plant cells are composed of multiple compartments, and little is known about the subcellular localisation of reactions and metabolites. Most enzymes of the central metabolism are known to be present in more than one compartment, which makes the modelling process even more difficult. Adding compartments to models raises other problems, such as the lack of information about transport reactions, substrate specificity, and energetic costs [43]. The assignment of reactions to compartments in plant models is typically performed by searching databases and using subcellular localisation prediction tools.

As plants are exposed and adapted to several environmental stresses, their cellular objectives are surely much more complex than maximising cell growth. For instance, during environmental changes or pathogen interactions, the fluxes are redirected from the primary metabolism to secondary metabolic pathways to produce the metabolites for the plant's adaptation and defence [15, 44]. The most used objective functions in plant models are minimising the total flux, minimising the photon uptake, and maximising biomass. Although these objective functions have been successfully applied for simulating the

metabolism of plant tissues at specific developmental stages or under certain environmental conditions, they do not apply to all possible scenarios [15]. Therefore, defining an appropriate objective function in plant models remains exceptionally challenging.

Another challenge in plant modelling is the definition of constraints affecting plants. Most plant models use biomass synthesis at a defined growth rate as a constraint when the objective function is the minimisation of total reaction fluxes [129, 132, 148, 160, 162, 163]. However, this may not be enough to correctly predict the fluxes, as net biomass synthesis uses only a fraction of the cell's total energy [15]. Indeed, Cheung et al. [138] proved that accounting for transport and maintenance energy costs increases phenotype predictions' accuracy, showing that the definition of appropriate constraints is essential for obtaining realistic metabolic predictions.

Moreover, plants' photosynthesis, photorespiration, and respiration add complexity to their metabolic networks and complicate the modelling process [44]. Other factors affecting plant metabolism include complex interactions with symbionts and pathogens, competition mechanisms, and changes in available nutrients.

Finally, one of the main problems of plant modelling is that most models are generic and include all reactions known to take place in that plant, regardless of cell type or environmental conditions. As plants contain a wide variety of cell types, each with its specific active metabolism, generic models may lead to wrong interpretations as certain reactions or pathways are inactive in a specific cell type, even though being strongly active in others. Similarly, environmental conditions may influence the expression of metabolic genes; thus, enzymes may be active in specific conditions while inactive in others. Therefore, the reconstruction of context-specific models is crucial to obtain more realistic metabolic flux predictions. As seen before, several studies have integrated transcriptomics data with generic plant GSMMs to improve flux predictions and create plant tissue- and condition-specific models [133, 142–145, 154, 157]. However, as gene expression levels generally do not strongly correlate with reaction fluxes [91], the use of omics to improve the GSMMs is still challenging and inaccurate.

Tissue-specific models can be merged to form a multi-tissue model that simulates metabolic interactions between tissues [150, 167]. However, reconstructing multi-tissue models raises the challenge of defining the metabolites transported between tissues, highlighting the lack of information regarding this topic. As depicted above, there are already several multi-tissue models of plants. In most models, the different tissues replicate the original model, with few metabolic differences, and are connected by inter-tissue reactions or a common compartment. Light availability is a common constraint to differentiate context-specific models, for instance, leaves and roots, or diurnal and nocturnal leaves. The huge advantage of these multi-tissue models is that they simulate metabolic interactions between different tissues and organs, providing insights into complex resource allocation processes occurring in plants [167].

Despite the several challenges of plant metabolic modelling, significant advances have been made in the last years, which have allowed the reconstruction of more accurate plant generic, context-specific, and multi-tissue metabolic models. These have been successfully applied for simulating phototrophic and heterotrophic metabolism, improving the production of metabolites of interest, and understanding metabolic phenotypes under different environmental conditions or at different developmental stages. Therefore,

metabolic modelling approaches have proven to be a relevant tool for understanding plant metabolism and, through the integration of omics data, the fluxes predicted by these models became more accurate and adjusted to environmental conditions. An improvement to the current studies with plant GSMMs would be integrating more than one type of omics data to increase the accuracy of the simulations and contribute to a better understanding of complex biological processes across the whole plant. However, the challenge remains on how to integrate multiple heterogeneous data into predictive multi-scale models [168].

2.3 Multi-view learning for multi-omics data integration

As seen before, the recent development of high-throughput technologies has led to the generation of large amounts of omics data that have revolutionised molecular biology research. Omics data provide direct insights into genetic and molecular profiles of cells or organisms, elucidating the underlying mechanisms leading to different phenotypes. Omics have been extensively applied to the study of diseases, identification of biomarkers, and characterisation of complex biochemical systems. However, as omics datasets are large, complex, and heterogeneous, extraction of knowledge from data is a very challenging task. Hence, the processing and interpretation of omics data require the use of appropriate tools, such as ML algorithms, which have been widely applied for this task. ML tools can identify patterns, select relevant features from large datasets, and make inferences from the observed data, without defining biological assumptions [7, 8]. ML algorithms have been applied in interpreting large metabolic datasets and developing tools to study cellular metabolism [9, 169, 170].

2.3.1 Machine Learning

ML has been defined as the study of algorithms that can automatically learn and improve by experience and adapt to new data input without being explicitly programmed [171].

An important distinction in ML is between “supervised” and “unsupervised” learning methods. In supervised learning, the model learns from a training dataset with both inputs and desired outputs, so it can later make predictions on the output of unseen observations. Supervised learning problems can be further divided into classification, in which the output has discrete values (such as sample labels), and regression, in which the output is a continuous value. In contrast, unsupervised models are trained with unlabelled data to identify the underlying structure, patterns, or data distribution [171]. Principal Component Analysis (PCA), clustering and T-distributed Stochastic Neighbor Embedding (t-SNE) are the most used unsupervised methods for dimensionality reduction and identifying sub-groups in the datasets. Some of the most used supervised and unsupervised algorithms are described in Table 4.

Regardless of the application, the development of supervised ML models involves several steps:

- data collection;
- pre-processing and feature selection;

Table 4: Description of the most used supervised and unsupervised ML algorithms.

ML algorithm	Type	Description
Principal Component Analysis (PCA)	Unsupervised	Linearly transforms the variable space into uncorrelated variables, named principal components, which capture most data variability.
Clustering	Unsupervised	Analyses the underlying data structure and groups data observations with similar features into clusters.
T-distributed Stochastic Neighbor Embedding (t-SNE)	Unsupervised	Dimensionality reduction algorithm that allows for nonlinear data separation.
Autoencoder	Unsupervised	Unsupervised artificial neural network that compresses and encodes the input data and then learns how to reconstruct the compressed data by minimising the differences with the original data.
Support Vector Machine (SVM)	Supervised	Prediction algorithm that aims to find a hyperplane that separates data observations into two classes, while maximising the distance between data points of both classes.
Artificial Neural Network (ANN) and Deep Learning	Supervised	Inspired by biological neural networks, an ANN comprises a collection of connected units named neurons that receive a set of weighted inputs and perform a weighted sum of these inputs, which is filtered by an activation function to generate the neural output signal. Deep Learning networks are complex ANNs, with more layers and neurons capable of reaching higher accuracy.
Regression algorithms	Supervised	Estimate the functional relationship between the output and the input features. Linear regression is used when the output variable is continuous, while logistic regression predicts the discrete output. Regressions are often combined with regularisation algorithms, such as the least absolute shrinkage and selection operator (LASSO) and elastic nets.
k-Nearest Neighbors (KNN)	Supervised	Instance-based method that compares new observations with the previously trained examples that have been stored in memory.
Decision Trees	Supervised	Build a tree-like model of decisions, wherein nodes denote the attributes, branches represent attribute values, and leaf nodes hold the class labels. The paths from the root to the leaf represent classification rules.
Random Forests (RF)	Supervised	An ensemble of decision trees trained individually with a different subset of features that are selected randomly.

- construction of training and test sets;
- selection and optimisation of learning models;
- evaluation of model performance.

The performance of ML models is usually evaluated through cross-validation methods, which allows the use of all instances in the dataset by splitting the complete set of observations into k subsets, and using each subset at a time for testing, while the remaining $k-1$ subsets are used to train the model [172]. Different error metrics can be calculated to evaluate the models. For classification problems, a confusion matrix can be created, which compares the true and the predicted classes for a set of observations (Figure 11). From the confusion matrix, various metrics can be calculated to assess the model's performance, such as accuracy, which represents the overall correctness of the model (Equation 2.8), recall (Equation 2.9), which measures the model's ability to identify all positive instances, precision (Equation 2.10), which measures the model's ability to avoid false positives, specificity (Equation 2.11), which measures the model's ability to avoid false positives in the negative class, and F1 score (Equation 2.13) that represents a balance between precision and recall. Balanced accuracy (Equation 2.12) is often used for assessing models trained with imbalanced datasets, taking into account the difference in the number of instances between each output class. In addition, Receiver Operating Characteristic (ROC) curves are useful for evaluating classification models as they relate sensitivity with the false positive rate (1-specificity), showing the ability of a model to separate two classes (Figure 12). Curves closer to the upper left corner have an area closer to 1, representing models with better performance, while diagonal curves with an area of 0.5 represent random predictions. The model accuracy can be measured by the Area under the ROC Curve (AUC).

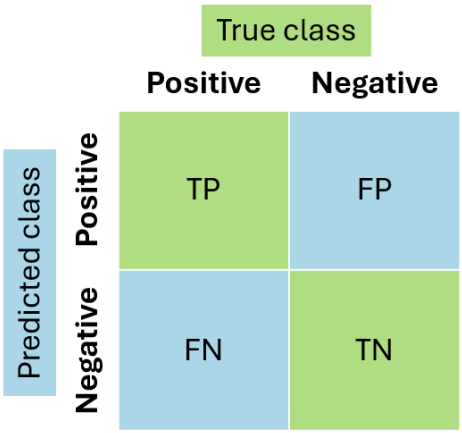


Figure 11: Confusion Matrix which compares the true and predicted examples. The correctly classified examples are the true positives (TP) and true negatives (TN). The incorrect ones are the false positives (FP) and false negatives (FN).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

$$recall = \frac{TP}{TP + FN} \tag{2.9}$$

$$precision = \frac{TP}{TP + FP} \tag{2.10}$$

$$specificity = \frac{TN}{TN + FP} \tag{2.11}$$

$$balanced\ accuracy = 0.5(recall + specificity) \tag{2.12}$$

$$F1\ score = \frac{2TP}{TP + 0.5(FP + FN)} \tag{2.13}$$

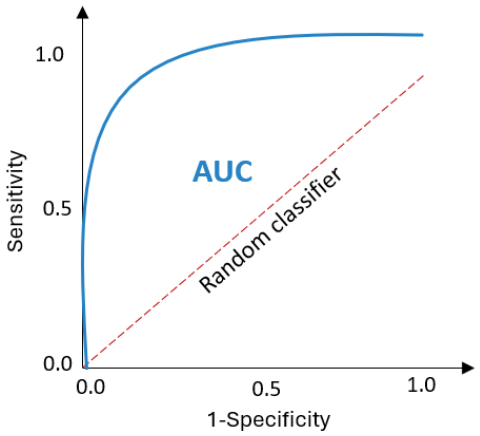


Figure 12: Example of a ROC curve.

The performance of the models can be improved by tuning the model's hyperparameters, which are the model settings that need to be defined before the training process and can significantly impact the results. The goal is to identify the combination of hyperparameters that leads to the best performance of the model in the validation set. Hyperparameter optimization can be performed using different methods, such as Grid search, which for given values exhaustively considers all parameter combinations, or randomized search over parameters, which samples parameter settings from a distribution over possible parameter values [173].

A common problem in the development of ML models is overfitting, which happens when the model learns the training observations too well, including their intrinsic noise and randomness. These complex models are unable to generalise to new data. Overfitting can be prevented by feature selection, a process where redundant data features are removed from the dataset, keeping the most relevant features for the models, and by introducing regularisation penalties to decrease the complexity of the model. However, these strategies can create too simple models that do not fit the training data well enough, resulting in poor predictive performance (high bias) and low generalisation as well. This is known as underfitting. As the purpose of ML algorithms is to generalise well, a trade-off between overfitting and underfitting is crucial to obtain good predictive models [174].

ML has been applied for analysing, detecting and extracting information in a wide range of molecular biology studies. It has become a crucial tool for transforming the massive amounts of data generated by high-throughput technologies into valuable knowledge, which has been one of the most significant current challenges in computational biology. Regarding plant molecular biology, ML has been applied to a wide variety of studies involving the analysis and interpretation of omics and imaging data [175]. Some illustrative examples include functional classification of plant ribosomal proteins [176], classification of genes and genera of a virus family that infects plants [177], and image processing to assess salt stress tolerance in wheat [178], classify grapevine varieties [179] and detect bacterial infections in melon plants [180].

Although most studies with plant omics use straightforward statistical approaches to process, interpret, and correlate the data, ML has been used in the analysis of genomics and transcriptomics data, mainly for the annotation of sequence elements, detection of differential expression patterns and identification of genes of resistance [175]. A list of the main ML studies applied to plant omics is available in Supplementary File 1. Despite the wide range of applications of ML to plant molecular biology, it is still underexploited by plant scientists due to the requirement of programming skills. Therefore, the development of user-friendly tools for analysing plant omics data is essential to fully explore ML's potential for extracting information from plant omics data.

2.3.2 Multi-omics integration

Separate analyses of different omics types limit our understanding of biological systems. The integration of multiple omics aims to identify complex biological relationships that may become evident only through the combination of multiple omics data, providing a comprehensive understanding of the relation between genotype and phenotype. Several tools have been developed to process and analyse omics data, but challenges remain in integrating different omics data types. These challenges arise from the different sizes, formats, and scales of the data being integrated, as well as the different complexities, noisiness, contents, and levels of agreement [181].

ML has been widely used in data integration methods, as per its ability to integrate diverse biological data types and perform data reduction, clustering, and phenotype associations. ML algorithms trained with multiple heterogeneous sources are classified as "multi-view" learning algorithms [182].

Data integration methods are broadly divided into two types of analysis: multi-staged or meta-dimensional [183]. In multi-staged analysis, also referred to as sequential or hierarchical, the multiple omics layers are analysed sequentially, assuming a linear causal effect of one layer on another. In meta-dimensional analysis, multiple data types are combined and simultaneously analysed, which requires a previous step of data fusion. Meta-dimensional methods are the most successful for large-scale data integration and are roughly categorised into concatenation-, transformation- and model-based methods (Figure 13). Alternatively, these categories are also called early-, intermediate-, and late-stage integration methods, respectively [7].

Concatenation-based integration combines multiple data matrices into a single comprehensive matrix,

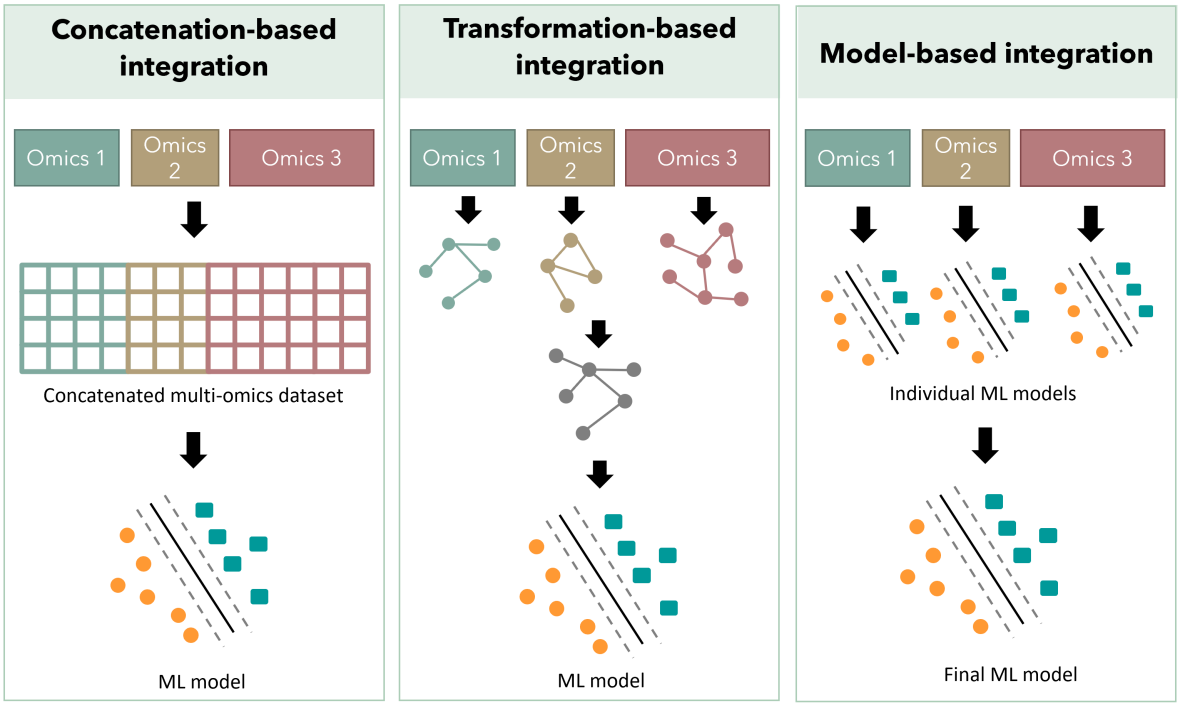


Figure 13: Strategies for omics data integration with machine learning methods. In concatenation-based (early-stage) integration, all omics data are combined in a single matrix before applying ML. Transformation-based (intermediate-stage) integration converts different omics types into intermediate representations, which are then integrated. Model-based (late-stage) integration involves the creation of individual ML models for each omics type and then combining their outcomes into a final ML model.

to which ML models are applied. This strategy has the advantage of considering interactions between different types of omics data and allowing the easy application of any statistical method to the final matrix. However, combining multiple matrices is very challenging due to the different scales and inherent noise, which can affect the final results. Hence, these methods will require a preliminary normalisation step, and may still lack reliability. Additionally, data reduction may be required as these methods can originate a high-dimensional matrix with more variables than samples [7, 183].

Transformation-based integration converts each data type into an intermediate form, such as a graph or a kernel matrix (a matrix describing similarity among observations). Then, the multiple graphs or kernel matrices are integrated and used to train ML models. This approach preserves the specific properties of each data type, and can easily combine any data structure by transforming it into an appropriate intermediate representation. However, the identification of interactions between data types is more complicated, hampering the interpretation of the results [7, 183].

Finally, model-based integration creates ML models for each data type, preserving data-specific properties, and then integrates them in a meaningful way to generate the final model. This strategy offers more flexibility than the previous ones, but can also miss interactions among different data types. It is particularly suitable for extremely heterogeneous omics data [7, 183].

The increase in the number of multi-omics studies has led to the development of a high number of approaches to integrate and analyse multiple omics data types, which have been extensively reviewed

[181, 182, 184–187]. These approaches are based on different statistical and ML methods, which are adapted and extended to the set of omics under analysis. Most of the available studies are dedicated to the integration of genomics and transcriptomics data and are based on dimension-reduction methods, like Canonical Correlation Analysis (CCA), and Partial Least Squares (PLS), network-based methods, and kernel-based methods.

2.3.2.1 Dimension-reduction methods

CCA is a versatile multivariate statistical method, developed to investigate the associations between two sets of variables [188]. It decomposes each set into linear combinations of variables (loading factors) that best explain the variability within each set while maximising the correlation between sets. Traditional CCA is not valid for analysing omics datasets due to their high dimensionality [185]. Thereby, this approach has been adapted for omics data analysis through the introduction of sparsity regularisation, resulting in sparse CCA (sCCA) methods, which select smaller loading factors that maximise the correlation between sets.

PLS is another multivariate statistical technique for the identification of the fundamental relationships between two sets [189]. This approach tries to find the loading factors that maximise the covariance between predictors and the response. PLS is widely used to reduce the set of predictive variables; however, for high-dimensional omics data, sparse solutions are usually required to facilitate data interpretation [185]. For this reason, extensions of PLS were developed, such as sparse PLS (sPLS) [190] that includes regularisation, and orthogonal PLS (oPLS) [191], that removes the systematic variation of predictors not correlated to the response, facilitating the interpretation. Additionally, oPLS was extended to *O2PLS* [192], which also models the uncorrelated variation of predictors. This approach has been explored for the integration of plant omics data [193].

These methods and their extensions are incorporated in the *mixOmics* R package for integration of multi-omics datasets with a focus on variable selection. Another popular method available in the *mixOmics* R package is Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO). DIABLO [194] is the first multivariate integrative classification method for the identification of biomarkers. It is based on PLS and uses sparse generalized CCA to maximise the correlation between variables of different datasets and project the data into linear combinations of variables (latent components) while discriminating between the output groups. The output class of new data is predicted for each omics dataset based on distance scores. As different omics datasets can predict different classes, the consensus class is determined using majority vote or weighted vote functions. Three distance metrics are calculated: maximum, centroid, and Mahalanobis distances. The maximum distance predicts the output class with the highest score for a new observation. For the other two metrics, a centroid is calculated for each output class based on the training examples. The predicted class of a new observation is the one where the distance between its centroid and the latent components is minimal based on the Euclidean (Centroids) or Mahalanobis distances [195]. The analysis of variable loading weights on each component allows identifying the most important variables for the classification that can represent biomarkers of an output group. In addition,

the calculated correlations allow for defining relationships between variables from different datasets.

2.3.2.2 Network-based methods

Network-based methods have also been used to integrate multi-omics data and provide an intuitive way to model the interactions between biological molecules. Networks are usually used to represent biological systems and modelled as graphs, wherein nodes are biological entities, such as genes or proteins, and edges are links between two entities, representing their interactions. Multi-omics data can be integrated as attributes of nodes and edges, and graph measures, like degree and connectivity, can be useful for the identification of valuable biological information. There are several biological networks with different features, such as protein-protein interaction networks, representing undirected physical interactions between proteins, and gene-co expression networks, which represent the correlations between gene expression levels across different conditions. Additionally, multiplex and heterogeneous multilayered networks have been developed to connect multiple biological domains. Both types are composed of different layers and interactions between them, represented by inter-layer links, as well as interactions between nodes within the same layer. However, while in multiplex networks, each layer represents a different characterisation of the same nodes, multilayered networks are composed of layers of different node types [196] (Figure 14). These networks represent a flexible framework to integrate multi-omics data, in which each layer represents a type of omics data, and inter-layer links represent correlations between different omics types. Network diffusion and other ML algorithms have been used for integrative analysis of multi-omics networks and were recently reviewed in [196, 197].

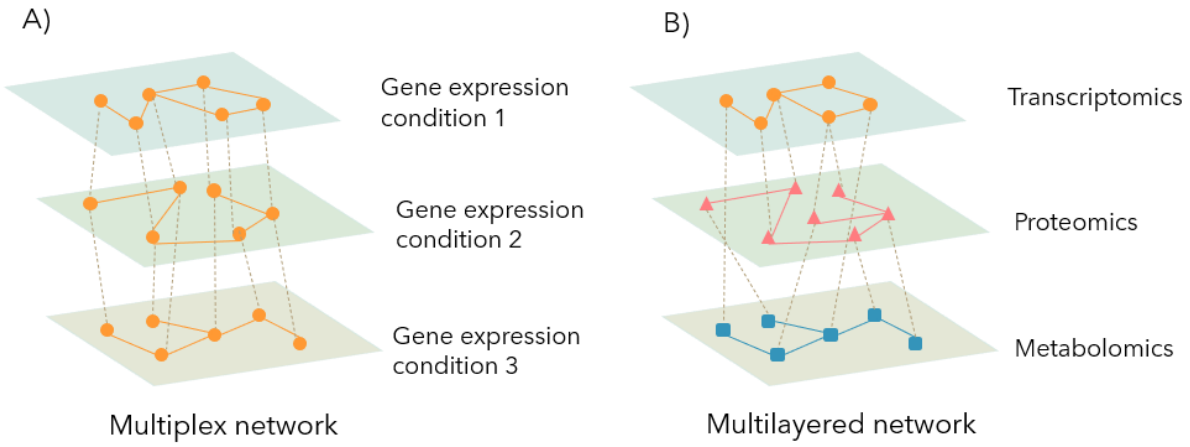


Figure 14: Multiplex and heterogeneous multilayered networks. A) A multiplex network is composed of layers with different measurements for the same set of nodes (e.g. genes). B) A heterogeneous multilayered network consists of layers of different sets of nodes (e.g. genes, proteins, and metabolites). Both types of networks include connections within the same layer and between layers.

2.3.2.3 Kernel-based methods

Kernel-based methods are a class of pattern analysis algorithms that have been widely used in classification, regression, clustering, and feature extraction approaches [182]. These methods map the original data to a higher-dimensional space, called the feature space, using kernel functions, which measure the similarity between data observations as the inner product between representations in the feature space [181]. The main advantages of kernel-based methods are that they only require defining the kernel function, can integrate several data types, and are dimension-free, meaning that optimisations do not depend on the number of data features [181, 182]. The most popular kernel-based algorithm is Support Vector Machine (SVM) [198] described above, which has been widely applied to prediction problems in biology [199].

In the context of multi-omics data, as the definition of the optimal kernel function can be difficult, Multiple Kernel learning (MKL) approaches [200] have been used. MKL technique first computes a kernel matrix for each omics type and combines these matrices to generate a single kernel matrix of integrated omics data. All kernel matrices must be projected on the feature space to be correctly combined. Therefore, as this method performs a transformation-based integration, the detection of interactions between different omics types can be more difficult, which represents the main limitation of kernel-based methods [181].

2.3.2.4 Multi-omics studies in plants

As plant responses to biotic and abiotic stresses are complex mechanisms that require strong regulation and multiple interactions between different cellular components, studies integrating different levels of omics are extremely important to provide a comprehensive view of plant acclimation. Most multi-omics studies of plants perform parallel or post-analysis data integration, in which the different omics datasets are analysed individually, and then key features of each dataset are combined into a new dataset or network 5. Furthermore, simple correlation analyses, such as Pearson or Spearman methods, have also been put forward for integrating and comparing two different omics datasets [201]. For example, the work of Ghan et al. [202] has integrated transcriptomics, proteomics, and metabolomics data to analyse and differentiate biochemical features from different grapevine cultivars. Dimensionality reduction of each dataset was achieved by PCA and the relationship between transcript and protein abundances was assessed through linear regression models and pairwise computation of Pearson's correlations. Another example is the study of Zaini *et al.*, which used transcriptomics, proteomics, and metabolomics data to analyse the mechanisms underlying the response of *Vitis vinifera* to *Xylella fastidiosa* infection. In this case, the authors calculated Pearson's correlation between all pairs of samples of all datasets, and samples of the same omics dataset showed a higher correlation than the samples of the same group of grapevines (infected vs non-infected grapevine) [203]. Thus, although these studies comprise more than one omics dataset, they are still limited as only pairwise correlations between samples were performed, not considering interactions between different samples or omics types.

However, more sophisticated correlation methods and regression models, such as CCA, PLS and

O2PLS, and multi-omics networks have been applied to study plants 5. Rajasundaram *et al.* [204] have studied the relation between cell wall polysaccharide composition and the characteristics of cotton fibres by integrating phenomics measurements and glycomics data. CCA was applied globally and simultaneously to model the association between glycan measurements and multiple fibres' characteristics. Then, sPLS was used to predict the cell wall polysaccharides that play a crucial role in the development of fibre characteristics.

Bylesjo *et al.* [193] has investigated the effects of different light conditions on wild-type *Populus tremula x Populus tremuloides* (hybrid aspen) by applying the O2PLS method to integrate transcriptomics and metabolomics measurements under long and short-day conditions, identifying the key transcripts and metabolites that defined most of the systematic variation across the two datasets.

A previous study has integrated transcriptomics, proteomics, and metabolomics data to identify stage-specific biomarkers for berry development and withering. First, PCA and pairwise O2PLS discriminant analysis were applied to each dataset to group observations and identify variables and putative biomarkers associated with those groups. The predictive latent variables identified for each dataset were integrated pairwisely by O2PLS to identify correlated variables across different datasets [205].

Later, the O2PLS method was extended to the OnPLS method [206], which generalises to more than two datasets. This extension was used to integrate transcriptomics, proteomics, and metabolomics data from wild-type and mutant hybrid aspen plants, to study multi-level oxidative responses in plants [207].

Anesi *et al.* [208] have studied the effect of environmental factors (*terroir*) on berry composition in grapevines of seven different vineyards over three years, using transcriptomics and metabolomics data. Exploratory and correlation analysis were performed with PCA and PLS-based discriminant analyses to investigate differences in the berry metabolome composition for each vineyard and identify the metabolites that best differentiate between *terroirs*.

Another multi-omics study has analysed gene expression data, metabolomics, and proteomics to predict phenotypic traits of *S. tuberosum* (potato) [209]. Random Forests were applied to each dataset to identify the features that significantly explained the variation in each quality trait. These features were used to construct a partial correlation network for each trait, with genes, metabolites, proteins, and traits as nodes and correlation values as edges.

Network analysis was also used to investigate the phenylpropanoid composition of berries, integrating gene co-expression data and interactions between micro RNA (miRNA) and genes and long non-coding RNA and genes. The integrated network includes genes, transcription factors, and both RNA types as nodes and the interactions and correlation values as edges [210].

Another example is the work of Jiang *et al.* [211], who have combined genomics, transcriptomics, proteomics, and phenomics data from *Z. mays* and built a multi-omics network to obtain insights into its development. First, weighted networks were created, one for each data type, and then these were fused into the final weighted network by iteratively updating each network with data from other networks, making them more similar at each step.

Finally, time-series gene expression datasets from the *Chlamydomonas reinhardtii* microalgae under

light and dark conditions were analysed to discover functional links between conditions [212]. Each condition was considered a dataset, and co-expression networks were built for each dataset, which were then aligned into a single network and clustered to identify the genes linking functions from different datasets [212].

Table 5: Examples of plant multi-omics studies.

Author	Ref.	Application	Method	Omics
Ghan 2015	[202]	Assess biochemical differences in *V. vinifera* cultivars	PCA, Pearson Correlations	transcriptomics, proteomics, metabolomics
Zaini 2018	[203]	Analyse *V. vinifera* metabolic response to *Xylella fastidiosa* infection	Pearson Correlations	transcriptomics, proteomics, metabolomics
Rajasundaram 2014	[204]	Study the relation between cell wall polysaccharide composition and cotton fiber characteristics	CCA, sPLS	glycomics and phenomics
Bylesjo 2007	[193]	Analyse the effects of different light conditions on *P. tremula x tremuloides*	O2PLS	transcriptomics and metabolomics
Zamboni 2010	[205]	Predict biomarkers for berry development in *V. vinifera*	O2PLS	transcriptomics, proteomics and metabolomics
Srivastava 2013	[207]	Study oxidative responses in *P. tremula x tremuloides*	OnPLS	transcriptomics, proteomics and metabolomics
Anesi 2015	[208]	Analyse the terroir effect on berry composition of *V. vinifera*	PCA, PLS, O2PLS	transcriptomics and metabolomics
Acharjee 2016	[209]	Predict phenotypes of *S. tuberosum*	Random Forest, Network analysis	transcriptomics, proteomics and metabolomics
Jiang 2019	[211]	Study *Z. mays* development	Network analysis	genomics, transcriptomics, proteomics and phenomics
Nguyen 2019	[212]	Study the relation between day and night metabolism in *C. reinhardtii*	Network analysis	transcriptomics light and dark conditions

2.3.3 Combining Constraint-based Modelling and Machine Learning

So far, ML and Constraint-Based Modelling (CBM) approaches have been mostly applied independently in biological and biomedical research. However, their combination holds great potential due to their common formulations and complementary characteristics. Recently, the first studies combining ML and CBM approaches have emerged and were previously reviewed in [8–10, 168, 169, 213–215]

The integration of ML and CBM comprises three main cases: fluxomics analysis, multi-omics analysis, and generation of constraint-based models and fluxomics (Figure 15). In fluxomics analysis, the flux

distribution predicted by CBM methods is analysed by ML methods. In multi-omics analysis, omics data can be included as GSMMs' constraints to create context-specific models and generate more accurate flux predictions. The predicted fluxes can be integrated with experimental omics to be jointly analysed by ML methods. In the third case, ML is trained with experimental omics data to predict metabolic models and fluxomics data. All three cases can apply supervised or unsupervised ML methods. Thus far, these studies were mainly applied to bacteria, yeast, and human cells, but not to plants. In the following sections, we review the representative studies of the field and summarise them in Table 6.

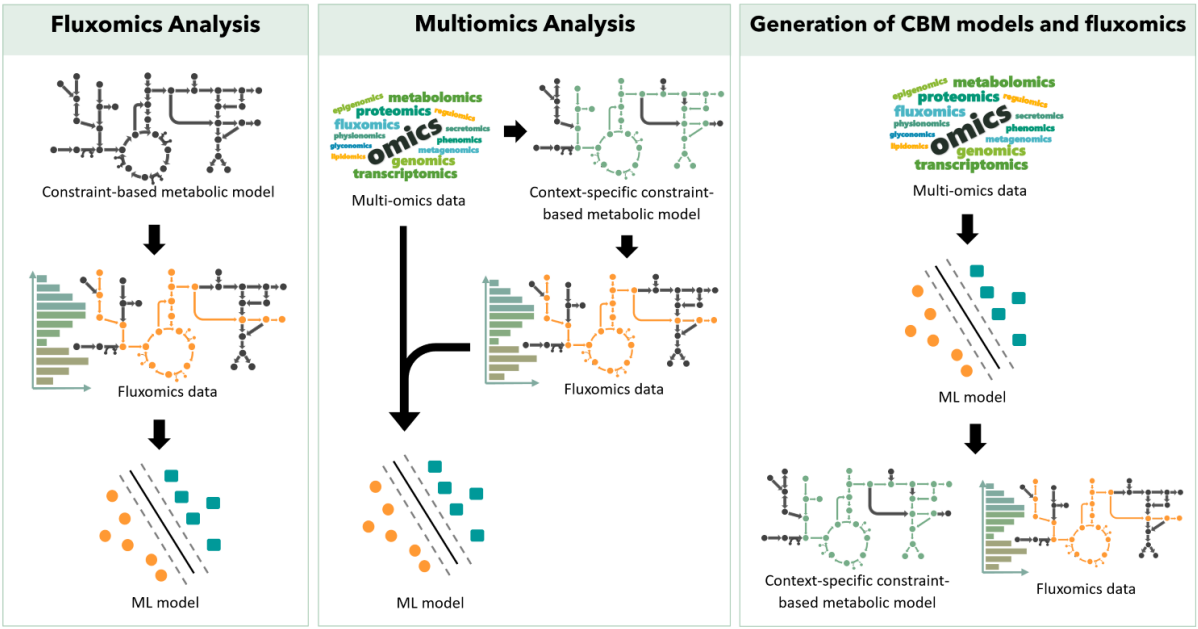


Figure 15: Type of analyses combining constraint-based modelling and machine learning. Fluxomics analysis consists of applying ML to the fluxomics data predicted by metabolic models' simulations. In multi-omics analysis, omics data can be integrated within metabolic models to generate context-specific fluxomics data, which can be analysed by ML in combination with omics data from high-throughput technologies. Alternatively, ML can be trained with omics datasets to produce or improve metabolic models or fluxomics data.

2.3.3.1 Fluxomics analysis

In the following examples, fluxomics data is generated by CBM methods and analysed by ML. Although PCA has been widely applied to simplify and identify patterns in metabolic data, its results are complex and difficult to interpret biologically. Therefore, two approaches, named Principal Elementary mode Analysis (PEMA) [216] and Principal Metabolic Flux Mode Analysis (PFMA) [217], have combined PCA and CBM to identify the flux modes that explained most flux variance and have minimum deviations from a steady-state condition.

The PEMA approach was extended to dynamic conditions using supervised ML (dynEMR-DA) in a subsequent work [218]. Here, dynamic EFMs were defined as partially activated EFMs at each time point of the simulation. A small kinetic model of *S. cerevisiae* was simulated under different conditions. The

non-steady-state flux distributions were decomposed into a set of dynamic EFMs, which were examined by discriminant analysis to identify the pathways that best differentiated between conditions.

Hierarchical clustering was used by Magnusdottir et al. [219] to predict ecological interactions between human gut bacteria. The models were simulated alone and paired with every other model to represent co-growth under different fibre diets and oxygen conditions. The relative fitness was calculated for each pair of organisms and used to define the type of interaction. Lastly, the ratio of pairwise interaction types was clustered per condition and per taxonomy, which has resulted in three main clusters comprising microbes with different carbohydrate fermentation capabilities, suggesting that this capability may define the types of interactions between microbes.

Interactions of human gut bacteria were also predicted by DiMucci et al. [220] but using supervised ML and dFBA. Single and pairwise simulations were performed using dFBA, and the relative final biomass was used to classify the interactions as negative or non-negative. A random forest (RF) classifier was trained with vectors representing the presence or absence of exchange reactions in each organism to predict potential interactions between two microbes and identify relevant fluxes for the prediction.

Another example is the study of Shaked et al. [221] that has used ensemble learning to predict drug side effects. A human GSMM was used to calculate the flux bounds of the reactions through FVA after knocking out the genes that represent drug targets. These reaction bounds were used as features to an ensemble of Support Vector Machines (SVMs), where each model represented a side effect, resulting in a list of all potential side effects of the metabolically acting drug.

Other studies have trained ML models with fluxomics data obtained by CBM methods to predict microbial production rates under different bioprocess settings [222, 223]. Recently, this strategy was used to predict amino acid concentrations in a fed-batch Chinese hamster ovary (CHO) cell culture. The metabolic model was constrained with experimental measurements and predicted the initial amino acid consumption rates. Then, the flux predictions were refined and extended by linear regressions to a time-course profile [224].

Another study tried to clarify the glycosylation process by training Artificial Neural Networks (ANNs) with the fluxes of the reactions involved in nucleotide sugar donor synthesis, which were calculated by a stoichiometric model of CHO cells, to predict the glycan distribution of the antibodies produced [225].

2.3.3.2 Multi-omics analysis

Multi-omics analysis involves the integration of predicted fluxomics with experimental data using ML. For instance, this approach was used to predict essential genes [226] and reactions [227] and yielded better results than using only CBM methods.

Li et al. [228] have created condition-specific models by integrating gene expression data of cancer cell lines under different environments within a human GSMM. These models were simulated through FBA, and the K-nearest neighbors (KNN) model used the resulting fluxes to predict new targets for cancer drugs.

Following a two-stage data integration strategy, Kim et al. [229] have developed a normalised, well-annotated multi-omics database for *E. coli* to provide high-quality data for data-driven predictive analysis. Hence, ML models were trained with experimental data, including transcriptomics, proteomics, metabolomics, and growth rates, and with fluxomics data obtained by condition-specific models, which resulted from the integration of proteomics and transcriptomics with a GSMM. This work was the first to integrate a comprehensive set of omics to train an ML model.

Recently, Culley et al. [230] have integrated transcriptomics and fluxomics data of different *S. cerevisiae* mutants to predict growth, using three different strategies: early, intermediate, and late integration. The strain-specific fluxomics data were obtained by simulating the models constrained with the transcriptomics data. Gene expression and fluxomics were analysed separately and as a single dataset by ML models and feature selection was applied to reduce dimensionality for the early integration. In the intermediate integration, two multi-view methods were applied: Bayesian Efficient Multiple-Kernel Learning (BEMKL), which creates and combines different kernel matrices for each dataset, and a multimodal artificial neural network (MMANN), that contains a layer for each dataset fused via additional layers. Finally, RFs trained with each dataset independently were combined in an ensemble model for the late integration. The authors concluded that adding fluxomics to gene expression makes the results more accurate and biologically interpretable. This study was extended by Magazzù et al. [231] who showed that regularised linear models could outperformed MMANN in multi-omics analyses and highlighted the relevance of using fluxomics for better understanding the interactions among genes and phenotypes.

Lewis et al. [232] have integrated predicted fluxomics with experimental omics to overcome the lack of metabolomics in cancer datasets and created an ML classifier to identify biomarkers for radiation resistance. Context-specific models were generated by integrating transcriptomics and mutation data from radiation-resistant and non-resistant tumours. The FBA-predicted fluxomics were integrated with experimental omics using the late integration approach: multiple classifiers were trained on an individual dataset and combined in a meta classifier that calculates the final probability of radiation resistance.

Regarding drug side effects, Guebila et al. [233] have integrated drug-induced gene expression data within a metabolic model of the small intestine epithelial cells to generate drug-specific fluxomics data. Gene expression and the predicted fluxomics were trained by a multilabel SVM to predict gastrointestinal drug effects and the drugs were clustered according to their metabolic and transcriptomic profiles which give new insights into the adverse reactions in the gut.

In another study, Vijayakumar et al. [213] proposed a pipeline that applies FBA and ML to improve phenotypic prediction in cyanobacteria. Condition-specific models were constrained by transcriptomics data and simulated using multi-objective FBA with three goals: biomass, photosystems I and II, and ATP maintenance reactions. Then, ML was trained with transcriptomics and predicted fluxomics which allowed the identification of key genes and reactions related to each condition. This pipeline clarified the mechanisms used by cyanobacteria to deal with variations in light intensity and salinity that could not be detected using transcriptomics alone.

Kavvas et al. [234] presented the Metabolic Allele Classifier (MAC), a metabolic model-based ML classifier that uses FBA to predict allele-specific antimicrobial resistance. MAC was formulated within the

FBA structure but using allele-specific flux capacity constraints and antibiotic-specific objective functions. This method takes the genome sequence of a *Mycobacterium tuberculosis* strain as input and classifies the strain as either resistant or susceptible to a specific antibiotic by optimising the antibiotic-specific objective function. Thus, MAC uses linear programming that acts as an ML classifier, in which the predicted flux state corresponds to each class, elucidating the biochemical processes leading to antibiotic resistance.

Lastly, a deep-learning approach, entitled DeepMetabolism, was developed to predict *E. coli* phenotypes, using CBM, biological knowledge, and gene expression data to define the neural network structure [235]. The first step is unsupervised pre-training and consists of an autoencoder with five layers: the first three represent gene expressions, protein abundances, and phenotypes, respectively, while the fourth and fifth layers are decoders representing reconstructed protein and gene layers, respectively. The layers were connected using biological knowledge: GPR rules from the GSMM connected the first and second layers, and GSMM's simulations identified essential reactions for each phenotype to connect the second and third layers. The second step consists of supervised training, using the same autoencoder model, though only with the first three layers trained to predict phenotypes.

2.3.3.3 Generation of constraint-based models and fluxomics

Instead of using ML to analyse fluxomics data, ML models can be trained with experimental data, and the results can be used to create models or improve flux predictions [96, 236–238]

For instance, a web-based platform called Mflux was developed to predict the central bacterial metabolism fluxes using ML models [236]. The models, namely SVM, KNN, and decision trees, were trained with experimental fluxomics data under different conditions to predict the flux of the central metabolism reactions and associate metabolic fluxes with the conditions. As the fluxes predicted by ML may not follow the stoichiometry of metabolic networks, quadratic programming was applied to adjust the predicted fluxes to satisfy the stoichiometric constraints.

ML can also be used to preprocess omics data and define additional constraints for CBM. For instance, Brunk et al. [237] presented a workflow that combines ML, metabolomics, and a GSMM to characterise *E. coli* strain variation. PCA is applied to reduce the dimensionality of metabolomics data and identify key metabolites driving strain variation. Then, the preprocessed metabolomics data were used to adjust the flux bounds of the GSMM of *E. coli* to achieve a better characterisation of the flux network of each strain.

Another example is the unsteady-state flux balance analysis (uFBA) workflow for integrating time-course metabolomics to predict metabolic fluxes in dynamic conditions [96]. The first step is to discretise nonlinear metabolomics data into time intervals of linear metabolic states using PCA. Then, the authors perform a linear regression to estimate the change rate for each metabolite for a specific state, using a 95% confidence interval of the rate as reaction flux bounds in a constraint-based model. As metabolomics data can be incomplete due to experimental errors, uFBA also implements a relaxation algorithm to determine the minimum number of unmeasured metabolites whose concentration needs to vary for the model to be feasible.

Finally, another study has trained ANN models with experimental enzyme concentrations to predict the

fluxes for the NADH consumption by glycerol-3-phosphate dehydrogenase in the upper part of glycolysis. In this case, no kinetic parameters or stoichiometric constraints were considered but a large and diverse dataset of enzyme concentrations is needed to obtain accurate flux predictions [238].

2.3.3.4 Other applications

ML has also been indirectly applied to other steps of GSMM reconstruction, namely genome annotation and gap-filling [8]. For instance, an approach used several multiclassification ML models to classify enzymatic reactions using a dataset of hydrolysis and redox reactions. The ML models predicted whether oxidoreductases or hydrolases catalysed specific reactions and the subclasses for each type [239].

Another method named DeepAnnotator applies deep learning models trained with DNA embeddings to identify genes and annotate prokaryotic genome sequences [240].

Regarding gap-filling, a set of ML models predicted the pathways present in an organism [241]. The models were trained with curated information on which pathways are present and absent in six organisms, achieving performance similar to other pathway prediction algorithms. Another approach uses association rule mining trained with UniProt entries to predict metabolic pathways in prokaryotes [242].

As mentioned above, using datasets with a large number of samples is crucial to obtain good ML models. The studies cited in Table 6 show that the number of samples used varies greatly depending on the type of ML and the application, ranging from a few dozen to thousands of samples.

Generally, unsupervised studies use datasets with fewer samples, except for the study of Magnusdottir et al. [219], which includes over 200000 samples. Furthermore, the studies for bacteria and yeast usually have more samples than those using human data, which is expected as data acquisition is cheaper and more accessible for smaller organisms. The study of Lewis et al. [232] was the human study that used the most extensive dataset, including data for 915 patient tumours.

The problem of having datasets with few samples is even more evident for plants due to the lack of efforts to collect, integrate and standardise plant omics datasets and experimental conditions. Usually, some procedures are adopted to deal with small datasets. First, it is common to choose simpler models with few parameters, such as logistic regression, to avoid overfitting. In addition, the use of regularization techniques and the creation of ensemble models enhances the power of generalisation. Second, it is also important to remove outlier observations and to select the relevant features to decrease the bias in the dataset. Another strategy that can improve the results is to extend the dataset by creating synthetic observations or integrating data from other sources, which is still very challenging. If the dataset is unbalanced, one solution is to perform an oversampling, which consists of increasing the number of observations of the minority class [243]. Finally, choosing the method for model validation is also important to get realistic performance metrics. For instance, the Nested Cross Validation approach was proven to make an unbiased performance evaluation [244]. Furthermore, other strategies have been developed to overcome the low-quality data problem in diverse applications, such as decomposition methods to generate extra data samples and impute missing features [245].

Table 6: Hybrid studies combining ML and CBM approaches, including the CBM and ML components and the application.

First Author	CBM	ML	Task
		Fluxomics analysis	
Folch-Fortuny 2016 [216]	EFMs	PCA (21-26 samples)	Identify metabolic patterns
Bhadra 2018 [217]	EFMs	PCA (12-28 samples)	Identify responsive pathways
Folch-Fortuny 2018 [218]	Dynamic EFMs	Discriminant analysis (64 samples)	Identify distinguishing metabolic patterns between conditions
Magnusdottir 2017 [219]	FBA	Hierarchical clustering (298378 samples)	Explore ecological interactions
DiMucci 2018 [220]	dFBA	Random Forests (RF) (9900 samples)	Predict microbial interactions
Shaked 2016 [221]	FVA, gene knockouts	Ensemble of SVMs (190-426 samples)	Predict drug side effects
Oyetunde 2019 [222]	FBA	PCA, SVM, elastic net, RF, k-Nearest Neighbors (KNN), Artificial Neural Network (ANN), ensemble (1200 samples)	Estimate titer, production rate and yield of microbial factories
Czajka 2021 [223]	FBA, gene knockouts, gene overexpression	RF, elastic net, kNN, gaussian process regression, support vector regressors (2915 samples)	Predict Yarrowia lipolytica bioproduction
Schinn 2021 [224]	Flux sampling	Linear regressions (80 samples)	Predict amino acid concentrations in CHO cell cultures
		Multi-omics analysis	
Plaimas 2008 [227]	RF, gene KO	SVM (1356 samples)	Predict essential reactions
Nandi 2017 [226]	Flux Coupled Analysis	SVM-RFE (768 samples)	Predict essential genes
Li 2010 [228]	Condition-specific models	Kernel kNN (260 samples)	Predict new drug targets

Continued on next page

Table 6: Continued from previous page

First Author	CBM	ML	Task
Kim 2016 [229]	Condition-specific models	RNN, LASSO regression, ensemble (649 samples)	Predict cross-omics states in *E. coli*
Culley 2020 [230]	Strain-specific models	Support Vector Regressor, RF, ANNs, BEMKL, MMANN, ensemble (1143 samples)	Estimate yeast growth rate
Magazzù 2021 [231]	Strain-specific models	Regularised linear models, ANNs, MMANN (1143 samples)	Estimate yeast growth rate
Lewis 2021 [232]	Patient-specific models	Ensemble of gradient boosting machines (915 samples)	Identify biomarkers of radiation resistance
Guebila 2019 [233]	Drug-specific models	SVMs, clustering (605 samples)	Predict gastrointestinal drug effects
Vijayakumar 2020 [246]	Condition-specific models	PCA, k-means clustering, least absolute shrinkage and selection operator (LASSO) regularization (24 samples)	Improve phenotypic prediction in cyanobacteria
Kavvas 2020 [234]	MAC	MAC (375 samples)	Predict allele-specific antimicrobial resistance in *M. tuberculosis*.
Guo 2017 [235]	FBA, gene KO	Deep ANN (30000 samples)	Predict phenotypes (Deep Metabolism)
		Generation of CBM models and fluxomics	
Wu 2016 [236]	Stoichiometry	SVM, kNN, decision trees (450 samples)	Estimate metabolic fluxes
Brunk 2016 [237]	FBA	PCA (126 samples)	Characterise strain variation
Bordbar 2017 [96]	Random sampling	PCA, linear regression (22 samples)	Estimate metabolic fluxes in dynamic conditions
Nagaraja 2019 [238]		ANNs (121 samples)	Predict fluxes for the upper part of glycolysis

2.3.4 Perspective: Machine learning and plant metabolic modelling

Although several CBM-ML hybrid studies have emerged in the last few years, these are still limited, and more work is required to understand the best ways to combine CBM and ML effectively [168]. Nevertheless, combining these two techniques for studying complex biological processes and interactions occurring in plants seems auspicious. On the one hand, ML is crucial for condensing and interpreting large and heterogeneous omics datasets to extract biological knowledge. On the other hand, CBM allows the analysis of metabolic fluxes associated with specific states, conditions, or tissues, which may involve integrating omics data within models. As integrating regulatory networks in GSMMs is still very challenging, ML models trained with transcriptomics data allow detecting and rectifying systematic errors associated with the GSMMs' flux predictions [8].

Multi-omics analyses seem to be the most promising application of combining ML and CBM, as these involve integrating traditional omics with fluxomics data predicted by CBM methods, which can provide meaningful insights about complex biological processes. Rana et al. [8] proposed an iterative scheme to combine these approaches. In such an approach, ML is initially used to analyse the data that will define the input constraints in a GSMM and later to analyse the predicted fluxes combined with experimental omics data. This process iteratively refines a GSMM until reaching consistency between CBM simulations, ML predictions, and experimental data.

The integration of omics from high-throughput technologies with fluxomics data provided by GSMMs is advantageous, as it seeks to overcome the specific limitations of each data type [9, 168]. Firstly, experimental omics data covers several areas, such as genomics, transcriptomics, proteomics, or metabolomics, while CBM is usually limited to fluxomics. Secondly, generating omics data does not require prior knowledge of the underlying networks, whereas GSMMs are based on extensive prior knowledge of metabolic networks, making their reconstruction time-consuming. Thirdly, although omics can be obtained promptly, these may contain intrinsic noise and experimental errors, requiring pre-processing and potentially leading to ambiguous interpretations. In contrast, GSMMs are curated and have straightforward interpretation, though relying on strong assumptions and the accuracy of flux predictions limited by the model quality and available knowledge [168]. Therefore, integrating omics data with metabolic models or the predicted fluxomics data can reduce ambiguity, generate accurate predictions, and provide more comprehensive analyses. Challenges remain in combining heterogeneous omics datasets with GSMMs. In the future, the increase in the number of omics layers will lead to the development of new multi-view algorithms. Hence, their combination with CBM is expected to grow as well [9, 168].

As plants are very complex, the studies of plant metabolism and physiology will significantly benefit from multi-omics analysis and the combination of ML and CBM approaches. Plant growth, responses to biotic and abiotic stresses, fruit composition, and emerging phenotypes involve complex mechanisms and multiple interactions between system components; thus, they must be studied as a system, comprising information of all levels. Currently, no work combining ML and CBM methods to study plant metabolism is available. However, several studies have integrated omics data with plant metabolic models and built context-specific and multi-tissue models. In addition, ML models have already been used to analyse

plants' omics data, but most plant multi-omics' studies analyse the different omics types separately [247]. Given the large amount of plant omics data generated, ML models can combine plant multi-omics and integrate them into CBM models. The resulting fluxomics data and experimental omics can then be jointly interpreted with ML models to identify, for instance, key genes or reactions associated with specific phenotypes. If enough data is available, part of the hybrid studies described in the previous section can be applied to analyse plant metabolism. For instance, the study of Vijayakumar et al. [213] can be used to identify key genes or reactions that best differentiate between conditions, elucidating the mechanisms of plants to adapt to different environmental conditions, such as variations in water and salt levels, and light intensities. Also, using multi-objective FBA for simulating the condition-specific models can be suitable for plants as their cellular objectives are complex and might differ between conditions. Similarly, the work of Lewis et al. [232] can be applied to identify biomarkers for plant tolerance to adverse environmental conditions, such as drought- or salt-tolerance, or diseases. Both examples can be applied to analyse the metabolic differences between plant varieties.

Another possible application of ML to CBM is predicting interactions between plants and microbes, pathogens, or symbionts. In the first case, the goal is to understand the mechanisms leading to disease and disease resistance and identify new drug targets. The latter aims to predict the plant-symbiont interaction network and analyse the effect of symbionts on metabolites or fruit production. For instance, the previous hybrid studies that predict interactions between human gut bacteria [219, 220] could be adapted to predict interactions between plants and microbes. This will rely on phenotype predictions from metabolic models of both plants and microbes, creating models encompassing both organisms and their metabolic interactions. The plant-pathogen models can also be used to predict drug side effects on plants, using approaches similar to the studies [221, 233]. In addition, as plant GSMMs present extensive metabolic gaps, ML models can be used for gap-filling and for predicting interactions between different tissues to generate better multi-tissue models, which will allow studying complex mechanisms related to plant responses to the environment and fruit quality.

As depicted above, one of the challenges in plant metabolic modelling is the definition of constraints. Based on the outputs of ML models trained with experimental omics, approaches like uFBA [96] and the work of Nagaraja et al. [238] can be adapted to define the appropriate constraints for specific metabolic processes, such as photosynthesis and photorespiration. Also, the best objective functions for describing a particular condition could be inferred from the context-specific omics data available.

Hence, we believe that most CBM-ML hybrid approaches can be applied to plants, including supervised and unsupervised methods. One main challenge will be collecting suitable plant omics data for the analysis of interest. Although large amounts of plant omics datasets have been generated, most of these are scattered and non-standardised, which hampers their analysis. Choosing the best ML method to use will depend on the available data and the purpose of each analysis.

For unsupervised learning, studies like the ones of Folch-Fortuny et al. [216], Bhadra et al. [217], Magnusdottir et al. [219], and Brunk et al. [237], were developed to explore data variation, identify metabolic groups and characterise metabolic patterns, using PCA or clustering methods and unlabelled

data.

The other studies have used supervised learning models, such as SVMs, ANNs, LASSO regressions, and RFs, and created predictors that can be applied to new data. The applications of supervised models included the prediction of drug effects and essential genes, the identification of biomarkers and novel drug targets, and the estimation of microbial growth rate. The use of deep learning models with CBM is very limited, as omics datasets usually contain few samples, which is even more evident in plant omics. The number of samples in plant omics datasets is still low for traditional ML methods, hampering the development of good predictive models. Therefore, ML complements CBM methods by defining the input constraints to metabolic models and improving the interpretation of the results. Given the complexity of plant metabolic networks, CBM-ML hybrid studies will give a more comprehensive and accurate view of the metabolic processes and variations in plants.

3

iplants repository

The work of this chapter was included in the publication submitted to the *18th International Conference on Practical Applications of Computational Biology and Bioinformatics* and is waiting for revision:

Sampaio, M., Rocha, M., and Dias, O. iplants: a repository of plant metabolic data. Submitted.

3.1 Introduction

The development of high-throughput technologies has revolutionised the field of molecular biology by generating huge amounts of data that can give insights regarding the gene, protein, or metabolite content of biological samples, which may lead to new knowledge. This rise in data and knowledge has led to an increase in the number of available biological databases that store, organise and facilitate access to all biological information. As described before, plant metabolic data can be found in different generic metabolic databases, such as KEGG [27], MetaCyc [28], and PlantCyc [34], or in species-specific databases created from MetaCyc, such as AraCyc for *A. thaliana*, available at PMN [34].

These databases are very important to assist in the study of metabolism, but there are still some limitations that hamper the integration of information from different sources, which is crucial for the reconstruction of GSMMs. First, a database Application Programming Interface (API) is missing or limited to a set of functions that do not allow a flexible extraction of all the information needed for reconstructing a GSMM. Also, the same metabolite, reaction, protein, or pathway has different names and identifiers in the different sources, and cross-references between databases are not always available, which makes the automatic integration of data almost impossible. In addition, after database updates, some identifiers are removed or updated to new identifiers without keeping track of the different versions. Hence, the old identifiers in the GSMMs cannot be found in the new versions of the database and available cross-references in other databases would be outdated too. This leads to the existence of incomplete and inconsistent information in the different databases that cannot be integrated. Maintaining the data updated and clean in all these databases requires a combined and time-consuming human effort, which is not always possible. Therefore, the use of data management systems to store large volumes of biological data is very relevant and can help to overcome the current limitations [248].

With this in mind and intending to reconstruct plant GSMMs, a repository named *iplants* was developed in this work to integrate metabolic data from PlantCyc and MetaCyc databases, enzyme data from UniProt, as well as the content of plant GSMMs. The repository includes tools to extract and transform data from different sources and unify the same entities into common objects. Also, a pipeline was created to allow for automatic database updates and to keep track of older versions. In addition, an API was created with important functions that facilitate automatic access to the repository and the collection of the data needed for metabolic model reconstruction.

Usually, after reconstructing a generic plant GSMM, omics data, mainly transcriptomics, are integrated with the model to create specific models for the different tissues of the plant. Therefore, another pipeline was developed to extract the metadata of the omics datasets available on GEO and *ArrayExpress* and save the information in *iplants* to facilitate the analysis of available omics datasets and the selection of the most suitable dataset for the integration with the GSMM. This pipeline can be run every time new datasets become available in the source databases to keep the *iplants* information up to date.

3.2 Data warehousing

The creation of a repository that integrates data from multiple sources requires extensive planning, schematisation, and standardisation to deal with different levels of data heterogeneity. As these tasks are usually not straightforward and extremely laborious, several tools and procedures were created to ease the process of data integration.

Data warehousing is a system that processes and centralises large amounts of dispersed data from different sources into a unified repository, easing the extraction of valuable insights from data [249]. This methodology resorts to Extract, Transform and Load (ETL) tools, which are responsible for extracting data from multiple sources, transforming the data into predefined formats and structures, and finally loading the transformed data into a unified and standardised repository [250]. Usually, these steps are performed sequentially in a pipeline, as represented in Figure 16.

Figure 16: A simplified ETL-based pipeline to extract, transform, and load heterogeneous data from different sources into a data warehouse.

The first step consists of extracting data from different external sources which can be flat files or databases that can be accessed through APIs. The unfiltered, untransformed data is usually stored in a staging area, which can be a non-integrated database or a simple file folder. This data is then transformed into a specific format that matches the schema of the target system. As different data sources often have distinct representations and formats, several procedures, such as filtering and joining, are adopted to standardise data attributes. Finally, the transformed data is loaded into the target database after verifying data integrity, which guarantees that key fields are not missing and that previously loaded data is not affected.

Data maintenance greatly benefits from the implementation of ETL tools, as these allow periodic automatic updates of the database as well as the integration of new data sources without affecting the existing system [250]. However, ETL only offers the generic guidelines for creating a data warehouse. Specialised knowledge of the data and the purpose of the repository is required to define the adequate database schema and data integration rules.

A data warehouse is based on a DataBase Management System (DBMS) which is responsible for defining the database schema and managing and querying the stored data. Traditionally, the relational DBMS is the most used. It is based on the relational model [251], which organises data into tables that can

be linked based on a common attribute between tables. This model offers great data integrity, consistency, and low redundancy, and is suitable for well-structured data. Data can be stored, manipulated, and extracted using Structured Query Language (SQL) [252]. In recent years, NoSQL databases (standing for "not only SQL") [253] have become more popular as these are non-tabular databases that allow for more flexible schemes and horizontal scalability, being more suitable for large volumes of unstructured data, which is the case of biological data. The most popular NoSQL DBMSs are described in Table 7.

Table 7: Description of the most popular NoSQL database models.

NoSQL database models	**Description**	**Example**
Key-value store	Data is stored as key-value pairs, mapping unique keys to associated records	Redis [254]
Document store	Data is stored in flexible documents with formats like JavaScript Object Notation (JSON) or Extensible Markup Language (XML)	MongoDB [255]
Column store	Data is stored by columns instead of rows	Cassandra [256]
Graph store	Data is stored in graphs with nodes and edges representing relationships in data	Neo4j [257]

3.3 Overview of the iplants repository

The data warehouse proposed in this work includes ETL tools to extract, transform, and integrate data to build a repository of plant metabolic data relevant to the reconstruction of GSMMs, as well as to collect and organise the metadata of plant omics datasets. This repository is also composed of two NoSQL databases, a graph store model using Neo4j [257] and a document store model using MongoDB [255]. The subsystems of this repository are schematised in Figure 17 and include data extraction, data transformation, data integration, data loading, data lake, and the Central Data Storage (CDS). In addition, an API was created to allow access to the CDS.

First, data is automatically extracted from relevant sources of metabolic or omics data, including database flat files and metabolic models, which are saved in the data lake. Then, transformers process the collected data by cleaning, filtering, and converting it into the predefined schema. This module includes parsers for reading the files in the DAT and Simple Omnibus Format in Text (SOFT) formats and uses CobraPy [258] to read the metabolic models. The transformed data is stored in JSON files in the data lake subsystem. Before the loading process, integration rules are applied to check if the new object identifiers already exist in the database or not, and decide which objects should be created or updated. Finally, the loading subsystems perform CRUD (Create, Read, Update, and Delete) operations with Django and using Neomodel, an Object Graph Mapper (OGM) for the Neo4j database, and Mongoengine, an Object Document Mapper (ODM) for MongoDB.

These ETL actions are organised into pipelines that are managed by Luigi, which is a Python package that manages data processing workflows, handling task dependencies and failure resolution.

In addition, an API was created to allow access to the CDS. This API is based on Django and Django REST framework and includes useful functions to retrieve the data needed for model reconstruction.

The tools and databases in this repository were containerised using Docker. In detail, the two databases are running in two containers, and the ETL tools and API are available in another container running in parallel. These tools are available at the https://github.com/martaS95/iplantsdb and https://github.com/martaS95/iplantsdb_omics GitHub repositories.

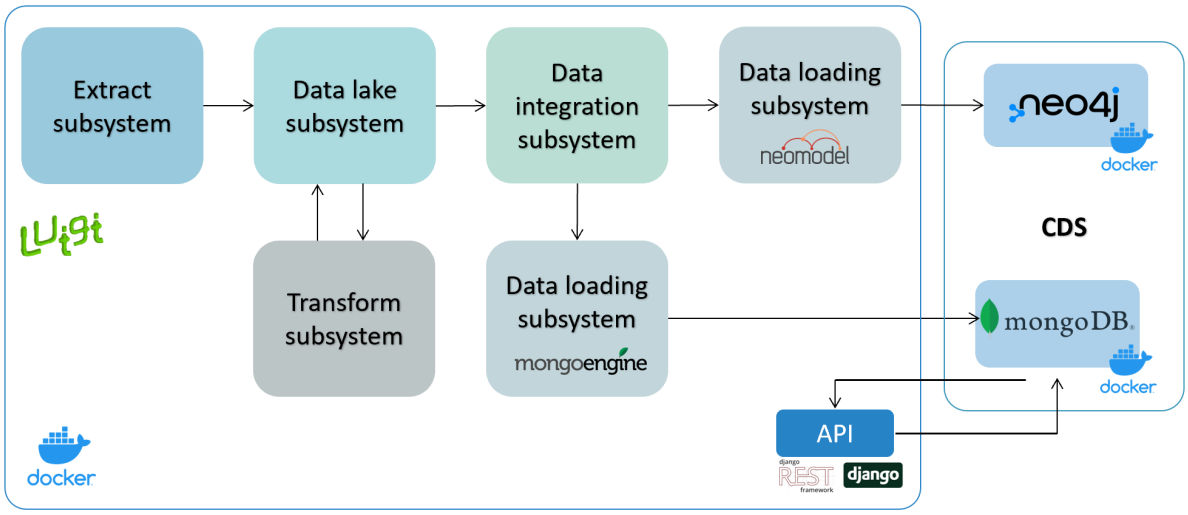


Figure 17: Overview of the *iplants* repository. The extraction subsystem downloads the relevant data sources, which are saved in the data lake subsystem. The transformation subsystem reads and transforms the collected data and saves the transformed data in the data lake as JSON files. The integration subsystem integrates data through a set of rules that ensure data uniqueness. Lastly, the integrated data is loaded to the CDS by the loading subsystems using Neomodel for Neo4j and Mongoengine for MongoDB. In addition, an API is available to allow user access to the data. The tools and the databases are deployed in Docker containers and the pipelines are managed by Luigi.

3.3.1 Central Data Storage Subsystem

The CDS subsystem comprises two databases, the Neo4j graph store and the MongoDB document store. Both databases save information about the same objects, but the Neo4j database focuses on the relationships between the different data objects, such as reactions and enzymes, while MongoDB saves the metadata of each object. The idea of using two different databases for storing the data was to speed up the extraction of the relationship information between objects from Neo4j.

3.3.1.1 Neo4j

Neo4j is the most popular DBMS of the graph store model. In a graph database, data is organised as nodes and edges of a graph. Usually, nodes represent the entities while edges are the relationships between node entities. Both nodes and relationships can have properties, which are attributes describing

the objects. Cypher is the query language used to perform CRUD operations in a Neo4j database. Neo4j databases can be easily extended due to their simple and flexible schemes that prioritise the representation of connections in the data [257].

The Neo4j database structure follows generic rules adapted from graph theory. A database object is an instance of an entity, such as a metabolite or reaction, with a unique identifier associated with a *namespace* related to the entity. An object can have many properties (attributes) with several data types. A relationship is a directed edge between two different objects. It is also associated with a *namespace* according to the type of interaction between the two objects and can have many properties with different types as well. Two objects can have several relationships in the same *namespace*. Also, two objects are considered different only if they have different identifiers and *namespaces*. Hence, the Neo4j database of *iplants* comprises a single universal graph containing all objects and relationships, including 7 entities and 13 relationship types. Each entity represents a *namespace* in the graph and all objects are associated with a single *namespace*. In detail, the Neo4j database is composed of the following entities:

- **Metabolite:** A metabolite object represents a chemical compound transformed by a given reaction.
- **Reaction:** A reaction object represents a biochemical reaction occurring in an organism.
- **Enzyme:** An enzyme object represents a protein that catalyses a given biochemical reaction.
- **Pathway:** A pathway object represents a biochemical pathway composed of a set of reactions.
- **Gene:** A gene object represents a segment of DNA responsible for expressing a given protein.
- **Organism:** An organism object represents a biological living system.
- **Metabolic Model**: A metabolic model object represents a plant GSMM from which metabolite and reaction data were extracted.

These entities, except for the metabolic model, are characterised by the following properties: *entry_id*, defining the object identifier, *name*, defining the name of the entity, *database_version*, defining the source from which the object was collected and *timestamp* defining the date and time the object was created or updated. It should be noted that not all objects have the property *name* as this information could be missing from the original data source. The metabolic model entity is defined by a model identifier (*model_id*), the *organism*, the *author* of the model, the *year* of reconstruction, and the *timestamp*.

All entities and relationships between them are schematised in Figure 18. The reaction object is connected to the metabolite by two relationship types, "REAC_USES_MET" and "REAC_PRODUCES_MET" to distinguish the reactants and products of a reaction, respectively. Both relationships have the property *stoichiometry* that saves the quantity of a metabolite used or produced in a reaction. An enzyme object is linked to the reactions it catalyses through the relationship "ENZ_CATALYSES_REACT". Also, as an enzyme can be a protein complex composed of several monomers, an enzyme object can be linked to another enzyme object by the relationship "COMPOSED_BY", which stores the number of monomers in the property *number*. A pathway object is also connected to the reactions that belong to it, through the relationship "PATH_HAS_REAC". A gene object is linked to the enzymes it expresses, using the relationship "GENE_EXPRESSES_ENZ". The organism entity is linked to enzymes, genes, and pathways by the relationships "ORGANISM_HAS_ENZ", "ORGANISM_HAS_GENE" and "ORGANISM_HAS_PATH",

respectively. Similarly, the metabolic model object is connected to an organism through the relationship "BELONGS_TO" and to the metabolites, reactions, and enzymes it contains by the relationships "MODEL_HAS_MET", "MODEL_HAS_REAC", and "MODEL_HAS_ENZ", respectively.

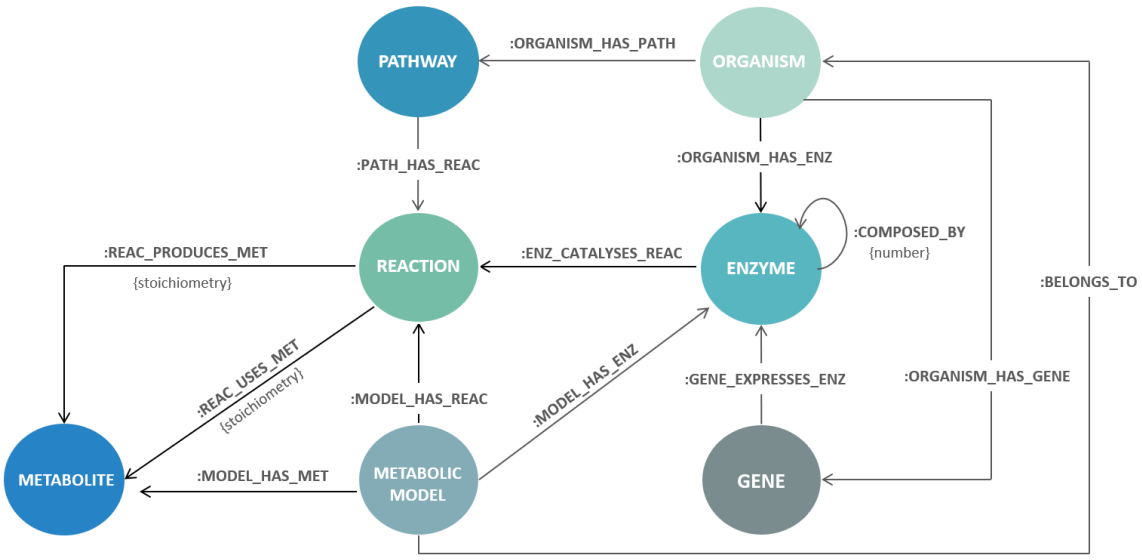


Figure 18: Neo4j database schema representing the relationships between the data. This includes the nodes for metabolites, reactions, enzymes, pathways, genes, organisms, and metabolic models. The nodes are defined by four attributes: database identifier, name, timestamp, and database version, except for the metabolic model node, which includes a model identifier, organism, author, year, and timestamp. 13 relationship types are represented in the database: reactions are linked to metabolites by two different types of relationships, defining their reactants and products, to enzymes, and pathways. These two nodes and genes are also linked to the organism node. Enzymes are linked to genes and can be linked to other enzymes, in the case of protein complexes that are composed of protein monomers. Finally, the metabolic model node is linked to an organism and to metabolites, reactions, and enzymes, which define its content.

The Neo4j database was built with Django and using Neomodel, which is an OGM for the Neo4j graph database built on top of Python's Neo4j driver. Neomodel is based on object-oriented programming and allows performing CRUD operations using simple Python objects instead of using Neo4j's query language, Cypher.

3.3.1.2 MongoDB

MongoDB is the most popular DBMS of the document store model. Document databases store data in flexible documents instead of using tables with fixed rows and columns. A document is a record in the database that stores data and related metadata of an object. Here, data is organised in field-value pairs, where field names are strings and values can have several data types, such as strings, numbers, arrays, or dictionaries. Documents can be written in several formats, such as JSON, XML, or YAML (Ain't Markup Language), but JSON is the most popular choice for representing data. A collection is a group of documents that have similar schemes but can have different fields. For instance, an object of the metabolite collection can be added without the *name* attribute. Therefore, document databases offer an intuitive and fast way of storing unstructured data and can be easily extended with new data due to

their flexibility [253]. MongoDB saves documents as binary JSON (BSON) files as it allows for more data types. It has its own query language and API methods for performing CRUD operations but has support for several programming languages, such as Python [255].

In MongoDB, each document requires a unique identifier field (*_id*) that acts as a primary key. If the *_id* is not defined by the user, an automatic value is generated by MongoDB. In *iplants*, the object identifiers match the identifiers of the source database. In total, MongoDB includes eight collections, matching the different entities in Neo4j, and a collection for saving the metadata of transcriptomics datasets. The saved attributes differ depending on the collection and are described below. Not all documents of a collection have these fields. The value data type for each field is also indicated: *str* stands for *string*, which is a sequence of characters, *int* stands for *integer*, which is a whole number (not a fraction), *dict* is a dictionary, and *boolean* only takes two possible value: "true" or "false".

Metabolite

- **entry_id** (*str*): identifier of the metabolite in the database
- **common_name** (*str*): name of the metabolite in the database
- **synonyms** (*list*): other names for the metabolite
- **crossrefs** (*dict*): cross-references of the metabolite in other databases
- **met_type** (*list*): compound classes
- **formula** (*str*): chemical formula of the metabolite
- **inchi** (*str*): inchi representation of the metabolite
- **inchikey** (*str*): inchikey representation of the metabolite
- **smiles** (*str*): smiles representation of the metabolite
- **models** (*dict*): metabolic models that contain the metabolite
- **state** (*str*): *"deprecated"* if the metabolite is not in the new database version
- **database_version** (*dict*): database version from which the metabolite data was retrieved
- **timestamp** (*datetime*): date and time of the document creation or update

Reaction

- **entry_id** (*str*): identifier of the reaction in the database
- **common_name** (*str*): name of the reaction in the database
- **synonyms** (*list*): other names for the reaction
- **crossrefs** (*dict*): cross-references of the reaction in other databases
- **direction** (*str*): reaction directionality (reversible or irreversible)
- **ecnumber** (*str*): EC number associated with the reaction
- **in_pathway** (*dict*): pathways in which the reaction participates
- **reactants** (*dict*): reactants of the reaction and respective stoichiometry
- **products** (*dict*): products of the reaction and respective stoichiometry
- **genes** (*dict*): genes associated with the reaction

- **enzymes** (*dict*): enzymes that catalyse the reaction
- **reaction_type** (*str*): type of the reaction (enzymatic or transport)
- **by_complex** (*boolean*): *"True"* if the reaction is catalysed by a complex protein, otherwise *"False"*.
- **sub_reactions** (*dict*): lumped reactions that combined form this reaction
- **lower_bound** (*int*): lower bound for the reaction in the model
- **upper_bound** (*int*): upper bound for the reaction in the model
- **compartment** (*dict*): compartments of the reaction
- **models** (*dict*): metabolic models that contain the reaction
- **state** (*str*): *"deprecated"* if the reaction is not in the new database version
- **database_version** (*dict*): database version from which the reaction data was retrieved
- **timestamp** (*datetime*): date and time of the document creation or update

Enzyme

- **entry_id** (*str*): identifier of the enzyme in the database
- **common_name** (*str*): name of the enzyme in the database
- **synonyms** (*list*): other names for the enzyme
- **crossrefs** (*dict*): cross-references of the enzyme in other databases
- **genes** (*dict*): genes that encode the enzyme
- **reactions** (*dict*): reactions catalysed by the enzyme
- **organisms** (*dict*): organisms known to express the enzyme
- **protein_type** (*str*): type of the enzyme (complex or peptide)
- **component_of** (*dict*): complex proteins of which the enzyme is a subunit
- **components** (*dict*): subunits of the enzyme
- **uniprot_status** (*str*): status of the enzyme entry in the UniProt database
- **uniprot_product** (*str*): name of the enzyme in the UniProt database
- **uniprot_function** (*list*): functions of the enzyme in the UniProt database
- **uniprot_location** (*list*): subcellular location of the enzyme in the UniProt database
- **sequence** (*str*): amino acid sequence of the enzyme
- **state** (*str*): *"deprecated"* if the enzyme is not in the new database version
- **database_version** (*dict*): database version from which the enzyme data was retrieved
- **timestamp** (*datetime*): date and time of the document creation or update

Gene

- **entry_id** (*str*): identifier of the gene in the database
- **common_name** (*str*): name of the gene in the database
- **synonyms** (*list*): other names for the gene
- **crossrefs** (*dict*): cross-references of the gene in other databases
- **reactions** (*dict*): reactions catalysed by enzymes encoded by the gene

- **enzymes** (*dict*): enzymes encoded by the gene
- **sequence** (*str*): nucleotide sequence of the gene
- **state** (*str*): *"deprecated"* if the gene is not in the new database version
- **database_version** (*dict*): database version from which the gene data was retrieved
- **timestamp** (*datetime*): date and time of the document creation or update

Pathway

- **entry_id** (*str*): identifier of the pathway in the database
- **common_name** (*str*): name of the pathway in the database
- **synonyms** (*list*): other names for the pathway
- **organisms** (*dict*): organisms that contain the pathway
- **reactions** (*dict*): reactions included in the pathway
- **super_pathways** (*list*): pathway class
- **pathway_links** (*dict*): pathways that this pathway is linked to
- **state** (*str*): *"deprecated"* if the pathway is not in the new database version
- **database_version** (*dict*): database version from which the pathway data was retrieved
- **timestamp** (*datetime*): date and time of the document creation or update

Organism

- **entry_id** (*str*): identifier of the organism in the database
- **common_name** (*str*): name of the organism in the database
- **scientific_name** (*str*): scientific name of the organism
- **taxid** (*int*): taxonomy identifier of the organism
- **species** (*str*): species of the organism
- **genus** (*str*): genus of the organism
- **family** (*str*): family of the organism
- **order** (*str*): order of the organism
- **org_class** (*str*): class of the organism
- **phylum** (*str*): phylum of the organism
- **kingdom** (*str*): kingdom of the organism
- **superkingdom** (*str*): superkingdom of the organism
- **genes** (*dict*): genes of the organism
- **pathways** (*dict*): pathways of the organism
- **enzymes** (*dict*): enzymes of the organism
- **database_version** (*dict*): database version from which the organism data was retrieved
- **timestamp** (*datetime*): date and time of the document creation or update

Metabolic Model

- **model_id** (*str*): identifier of the metabolic model in the database
- **organism** (*str*): name of the species
- **taxid** (*int*): taxonomy identifier of the organism
- **year** (*int*): year of model reconstruction
- **author** (*str*): the author of model reconstruction
- **annotation** (*dict*): results of DIAMOND annotation
- **genes** (*list*): genes in the model
- **enzymes** (*list*): enzymes in the model
- **reactions** (*list*): reactions in the model
- **metabolites** (*list*): metabolites in the model
- **pathways** (*list*): pathways in the model
- **gprs** (*list*): GPR rules in the model
- **timestamp** (*datetime*): date and time of the document creation or update

Transcriptomics

- **accession_number** (*str*): database identifier for the omics dataset
- **database** (*str*): database where the dataset is stored: GEO or *ArrayExpress*
- **title** (*str*): title of the omics dataset
- **data_type** (*str*): type of transcriptomics data
- **organism** (*str*): species from which data was collected
- **platform_id** (*list*): database identifier of the platform used
- **contributors** (*list*): names of the authors of the omics study
- **overall_design** (*str*): small description of how the dataset was obtained
- **samples** (*dict*): samples in the dataset and their description
- **last_update_date** (*datetime*): date and time of the document creation or update

The MongoDB database was also built with Django but using Mongoengine, which is an ODM for the MongoDB document database and Python. Mongoengine is also based on object-oriented programming, allowing to perform CRUD operations using simple Python objects.

3.3.2 Data lake subsystem

A data lake is a storage repository that saves large amounts of structured and unstructured data in its raw format as extracted from different sources. As data lakes store the original data, there is no need for predefined schemes or transformation functions. Then, the collected data, or a part of it, can be used for further analysis.

The data lake in this repository consists of folders that save all the source raw files, which include several formats, such as DAT, Text File (TXT), Comma Separated Values (CSV), and FASTA, as well as the SBML or XML format of the metabolic models, and SOFT and Sample and Data Relationship Format

(SDRF) formats of the omics datasets metadata. In addition, it also contains the transformed data created by the transform subsystem as JSON files. Hence, the data lake is different from the CDS, as it stores the collected diverse data in a schemaless and unstructured way, allowing different exploratory data analyses.

3.3.3 Data extraction subsystem

The data extraction subsystem has functions that allow the direct download of the source files for the two metabolic databases, PlantCyc and MetaCyc. The first idea was to use the PythonCyc package, a Python API for using the Pathway Tools software. This is a software implemented using the Common Lisp (CL) programming language that allows for extracting data from the Cyc databases and performing metabolic reconstructions, comparative genome analysis, and analysis of biochemical pathways. However, this API had some limitations that prevented the flexible extraction of all the data needed for reconstructing the metabolic models. For instance, no function was available to get the stoichiometry of a reaction to check whether it was mass-balanced or not. In addition, it also required the installation of Pathway Tools and required that, when updating the database, the new database files be compiled by the same version of the installed pathway tools. Thus, due to these limitations, data extraction was done by directly downloading the database DAT files from the source. However, no raw files were available regarding the organisms in the database. Therefore, BioCyc web services were used to extract the taxonomy identifiers of the organisms, and the Python package Taxoniq was used to access the NCBI taxonomy database and extract the organism lineage.

At first, only data from PlantCyc was going to be extracted. However, when integrating the data from plant GSMMs, many model identifiers were missing from the database but were available in MetaCyc. Thus, MetaCyc data was also extracted. At the moment, the database includes PlantCyc version 14.0 and MetaCyc version 26.1. Since January 2024, the access and extraction of MetaCyc information require a paid subscription. Therefore, this hampers the update of MetaCyc information on the *iplants* repository.

Regarding omics datasets, GEO's API was accessed using functions from *Biopython* and *GEOparse* Python packages that allowed filtering the records by experiment type and organism and extracting the SOFT files for the filtered records, in this case for *V. vinifera* RNA-Seq and microarray studies, which were saved in the data lake. SOFT is a line-based plain text format that contains the description of the platform and samples of each GEO dataset.

Since June 2022, *ArrayExpress* datasets have been stored in the BioStudies database [259]. This database was accessed through its API and the SDRF files of microarray and RNA-Seq experiments of *V. vinifera* were extracted and saved in the data lake. These files contain the sample description of *ArrayExpress* datasets. Additional metadata about the experiment, including title, description and authors, was directly retrieved from the API in JSON format. As *ArrayExpress* also imports experiments from GEO, the accession numbers extracted from GEO and *ArrayExpress* were compared and for those in common between the two sources, the metadata was only extracted from GEO.

The plant GSMMs were manually downloaded from their original publications and those with BioCyc identifiers were chosen to facilitate the integration. These models include *Arabidosis thaliana* (Poolman

et al. [129] and Cheung et al. [138]), *Oryza sativa* (Poolman et al. [130] and Chatterjee et al. [132]), *Solanum lycopersicum* (Yuan et al. [160]), *Zea mays* (Bogart et al. [136]), *Medicago truncatula* (Pfau et al. [162]) , *Setaria viridis* (Shaw et al. [164]), and *Glycine max* (Moreira et al. [163]). For simplicity, from now on these models will be referred to by their identifier in the database as follows:

- *Arabidosis thaliana* (Poolman et al. [129]): *athaliana2009*
- *Arabidosis thaliana* (Cheung et al. [138]): *athaliana2013*
- *Oryza sativa* (Poolman et al. [130]): *rice2013*
- *Oryza sativa* (Chatterjee et al. [132]): *rice2017*
- *Solanum lycopersicum* (Yuan et al. [160]): *tomato2015*
- *Zea mays* (Bogart et al. [136]): *maize2016*)
- *Medicago truncatula* (Pfau et al. [162]): *medicago2018*
- *Glycine max* (Moreira et al. [163]): *soybean2019*
- *Seraria viridis* (Shaw et al. [164]): *setaria2019*

In addition, protein sequences from *A. thaliana* were also manually downloaded in FASTA files from the TAIR 10 database to complete enzyme annotation, in cases where the UniProt identifier is missing from the enzyme cross-references.

3.3.4 Data transformation subsystem

As the database files of PlantCyc and MetaCyc have the same structure, the same transformer functions were applied for both databases. In general, this tool reads the extracted data, retrieves the relevant information, and transforms it into the desired structure. Firstly, a parser function is used to read the raw database DAT files and convert the data into a Python dictionary, where each key-value pair represents a database record. Then, it uses the Pydantic Python package to define the structure of the data and the validation rules, according to the desired schema of the database. Hence, the necessary data is filtered and retrieved from the created dictionaries and processed to fit the Pydantic object structure. In this process, the relationships between objects are also identified and saved in the appropriate attributes. The transform tool does many data processing operations, including:

- filtering relevant data using keywords or regular expressions;
- handling missing data;
- removing HTML tags, white spaces, and unwanted special characters from the text;
- joining data of different file fields in the same attribute;
- standardising data attributes. For instance, the direction of a reaction has different values in the database files that were converted to ”reversible” or ”irreversible” for simplicity.

In this repository, there is a different transformer for each entity, Metabolite, Reaction, Enzyme, Gene, Pathway, and Organism, as these are composed of different attributes. The transformed data of each entity is saved into a JSON file in the data lake subsystem.

Different transformers were also applied to read and process data from GEO and *ArrayExpress*, applying similar data processing operations. *GEOparse* was used to extract the information from GEO SOFT files and the data was saved into JSON files in a structure that matched the structure of the collection in the database. Some information from *ArrayExpress* was directly retrieved from the API in a JSON format and the SDRF files were parsed to get the metadata about the samples of each dataset. All the data was organised into a unique JSON file that was saved in the data lake.

In addition, different transformers were applied for reading and processing data from plant GSMMs. CobraPy was used to read and filter the relevant data from the SBML model files, which was saved into CSV files with the desired fields. After, these files needed to be manually analysed to check which attribute has the BioCyc identifier of the object. Usually, it is on the object identifier or name in the model, but additional characters can be attached to it, defining, for instance, the compartment of the object. Hence, this attribute must be cleaned to allow an exact match between model and database identifiers, which is crucial for data integration. In addition, as some BioCyc identifiers in the model can be deprecated, these need to be mapped and converted to the new database identifiers. Therefore, after manual curation, a new column is created in the CSV file that defines the *entry_id* of the object in the databases. Then, JSON files are created from the CSV files to be integrated and loaded into the database. The pipeline for integrating plant GSMMs into iplants cannot be automated with Luigi because of the required manual step.

3.3.5 Data integration subsystem

Different databases can have the same object with different identifiers, names, and annotations, which hampers data integration. Although PlantCyc and Metacyc generally have the same identifiers and names for the same object, a data integration tool is still needed to merge duplicated entries and normalise the data. Furthermore, the objects from the metabolic models can have identifiers and names different from those in the database but have already been manually converted to the desired identifiers in the transformation step. However, in the transform subsystem, no rules are defined to avoid the existence of duplicate objects in the transformed data. This is only performed in the data integration task.

In detail, the existence of duplicate objects is assessed by comparing the attribute *entry_id* that corresponds to the object identifier in the database. If two objects of the same type (e.g. Metabolite) have the same *entry_id*, these are considered to be the same. As the main goal of this repository is to use its data to reconstruct plant GSMMs, PlantCyc data is prioritised and loaded first to the database without integration as no duplicates are expected. Then, when adding MetaCyc data, PlantCyc data is not affected and some rules have been defined to complement this data with data from MetaCyc. Different integration tools were created for each database (Neo4j and MongoDB) and each entity.

Regarding MongoDB, when integrating an object already in the database from PlantCyc, the objects are compared and MetaCyc data is added to the attributes of type *list* and *dict* as well as to the object empty attributes, without affecting the data already in the database. For instance, all cross-references from PlantCyc are kept and the ones from MetaCyc are added to the database to have the data as complete

as possible. In case some attributes have contradictory data, for instance, in the *reactants* and *products* attributes of a reaction, PlantCyc data is kept (Figure 19). This is a limitation because in some cases MetaCyc data is the one that was corrected. However, there is no way of predicting this.

Regarding Neo4j, data integration follows the same rules mentioned above, which translates into creating new relationships between objects. For instance, if an Enzyme object is associated with a new reaction according to MetaCyc data, a new relationship between these two objects is created, without affecting the old ones.

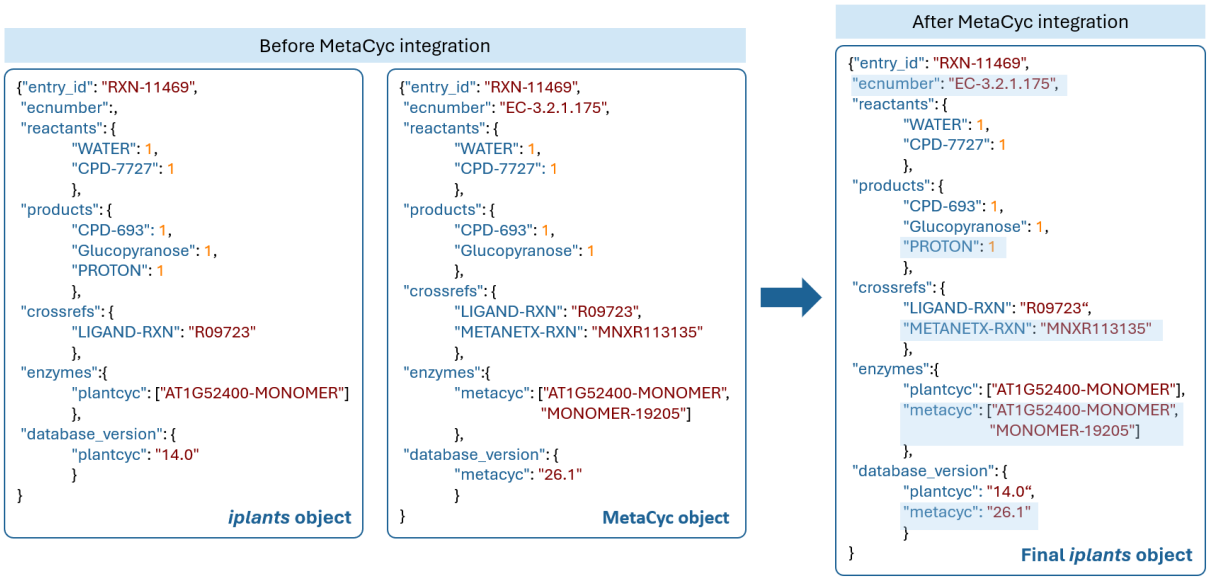


Figure 19: Example of the integration of a MetaCyc object in *iplants* database. Only a few attributes are displayed for simplicity. As the *entry_id* of the new object is already in the database, some attributes are compared. As the *ecnumber* field is empty, it is updated with the new information from MetaCyc. The same happens with the *crossrefs*, *enzymes*, and *database_version* attributes. As observed, the enzyme and database version information is kept separately from the Plantcyc data. The products are not updated as data from PlantCyc is preferred.

When integrating information from a new database version, the data from the older version of that database is replaced by the new one (Figure 20). In addition, the *database_version* attribute of the objects in the new version is updated to that version. On the contrary, the objects not in the new version are kept in *iplants*, but the attribute *state* is updated to *"deprecated"*.

When integrating new enzyme or gene objects that are not in the database yet, additional attributes are annotated using external sources, namely UniProt for enzymes and NCBI for genes. It is important to have sequence data on the repository to allow genome or protein annotation, which is the first step in the GSMM reconstruction. As no sequence data is available on PlantCyc and MetaCyc, this data needs to be extracted from other sources, and cross-references to these databases are crucial for executing this step.

Regarding enzymes, if the new object has the cross-reference for the UniProt database, it is used to connect to the UniProt's API and extract the amino acid sequence for the enzyme as well as its status, name, location, and functions described in UniProt. This data is saved in *iplants* to complete enzyme annotation. In case this cross-reference is missing and the enzyme record is from PlantCyc, the *entry_id* is searched on a FASTA file with protein sequences of *A. thaliana* from TAIR 10 database. In case of a

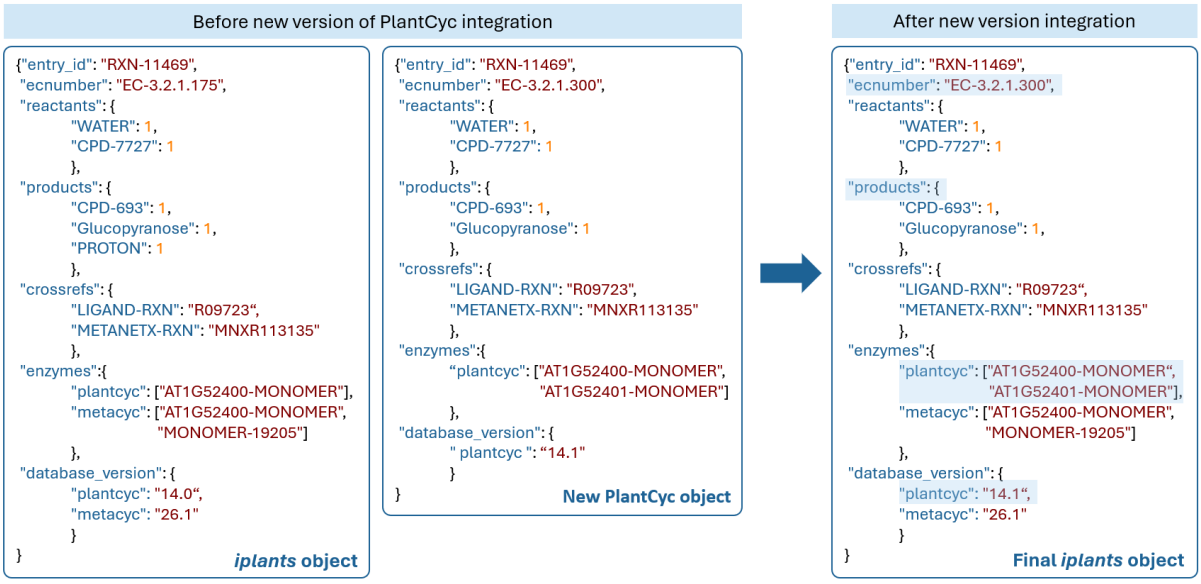


Figure 20: Example of an object update from a new PlantCyc version. Only a few attributes are displayed for simplicity and some field values are not real data. As the object in the database has data from MetaCyc, it can not be completely replaced by the new one. Instead, the attributes are compared and their values are updated with the new data. In the case of attributes that keep the information from both sources, like *enzymes*, only the data from PlantCyc is updated while the MetaCyc value remains the same.

match, the sequence is saved in the database. In the case of genes, cross-references for the NCBI are used to get the nucleotide sequence through NCBI's API using the BioPython package. This is performed here instead of in the extraction subsystem to only be executed for new objects not yet in the database and thus considerably reduce update time.

The content of plant metabolic models is also integrated based on the *entry_id* of metabolite, reaction, and enzyme objects. In Neo4j, if the *entry_id* is already in the database, a relationship is created between the metabolic model and the matched object (Figure 21 A). In MongoDB, the attribute *models* of the matched object is updated to include the *model_id* as key (Figure 21 B). The value of the key can be another dictionary or a tuple with the identifier of the object in the model and the respective compartment. The integration of metabolic model content into the database is more straightforward as the most difficult task is done before in the transformation step, in which the identifiers in the model were manually mapped to *iplants* identifiers.

The metadata of the omics datasets is saved in a separate collection of MongoDB, not affecting the metabolic data already in the database. No duplicate entries are expected in the data integration subsystem as the accession numbers of datasets from GEO and *ArrayExpress* are compared in the extraction step. Hence, the objects are directly loaded into the database without integration rules.

3.3.6 Data loading subsystem

The loading subsystem is responsible for loading the transformed, integrated data into the databases. This repository has two loading tools, one for each database. Here, JSON files from the data lake are

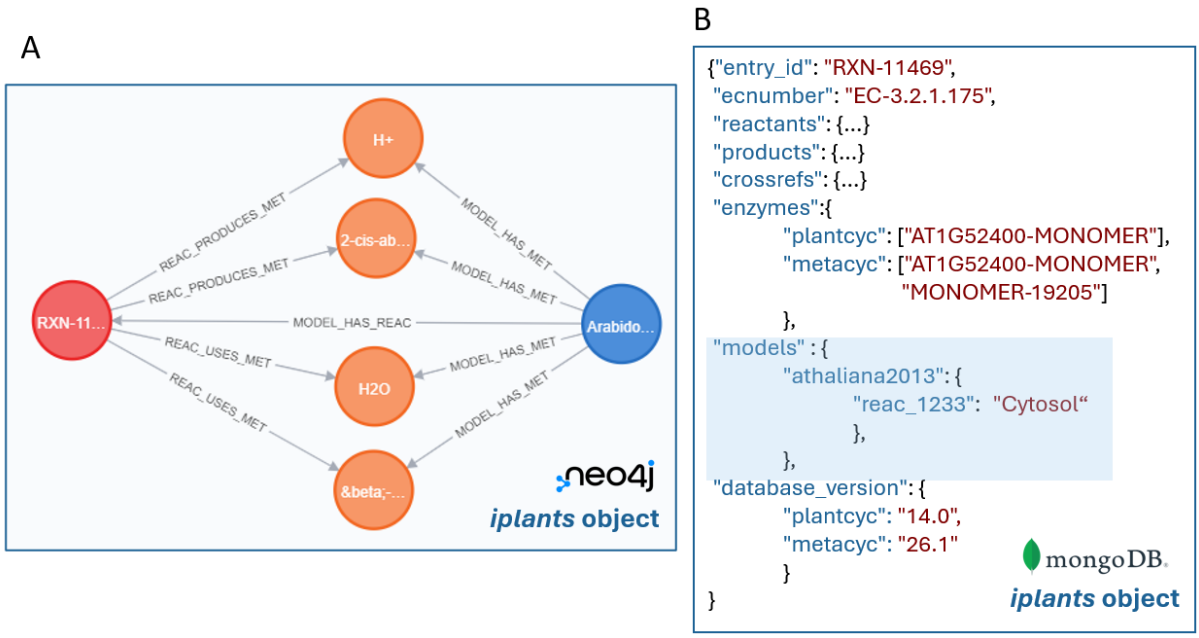


Figure 21: Example of the integration of a plant GSMM content in Neo4j (A) and MongoDB (B). In Neo4j, a node is created for the metabolic model, and relationships are created with the reactions and metabolites in the model. In MongoDB, a new attribute named *models* is created in the metabolite and reaction objects, defining the models that contain the object, and the name and compartment of the object in the model.

loaded and the data is converted and saved in database objects, using Neomodel and Mongoengine, as previously mentioned.

Data integration and loading subsystems are interconnected and work simultaneously. Data objects go through the integration rules and then are loaded into the database, one at a time. Hence, the loading tool does not perform data integration or normalisation. Instead, it loads or updates the objects according to the instructions given by data integration rules. The *timestamp* attribute is created during object loading and stands for the time of object creation or update. The information for Neo4j relationships is also stored in the JSON files and the relationships are created during the loading process, according to the integration rules. As different entities have different types of relationships, a loading function is defined for each entity. Therefore, the entities are loaded in the following order: Metabolite, Reaction, Enzyme, Gene, Pathway, and Organism. This facilitates the creation of relationships between different entities. For instance, when a reaction instance is loaded into the database, the associated metabolites are already there and the relationships between the reaction and metabolites can be easily created. As MongoDB only saves the entity metadata, only a generic function is available for loading the data. Thus, data objects are transformed, integrated, and loaded without affecting the CDS.

3.3.7 Pipelines

The ETL actions described so far are organised into three pipelines. One pipeline is used to extract, save, and update the metabolic data from PlantCyc and MetaCyc (Figure 22); the other is used to extract, save, and update the metadata of the omics datasets from GEO and *ArrayExpress* (Figure 23); and the last

extracts and integrates the metabolic content of plant GSMMs into the repository (Figure 24). The first two pipelines are managed by Luigi, which executes all the tasks automatically and consecutively in the correct order, handling dependencies and errors. The last pipeline can not be run automatically because a manual curation step is required for accurate data integration.

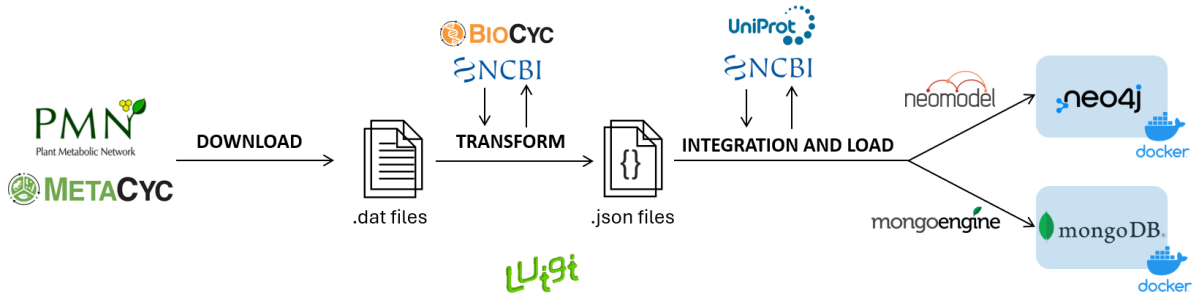


Figure 22: Pipeline for metabolic data update in *iplants*. The DAT files are downloaded from PlantCyc or MetaCyc databases and stored in the data lake folder. These files are parsed and the relevant information is filtered and transformed into objects with the same structure as database objects, which are saved in JSON files. Details on the organisms are not available on the DAT files. This information is retrieved from BioCyc through its API and NCBI Taxonomy database using the *taxoniq* Python package. The JSON files with the transformed data are parsed, and data is integrated and loaded according to the set of rules that ensure data uniqueness and consistency. In the case of a new Enzyme object, additional information is retrieved from UniProt when the cross-reference is available, mainly the amino acid sequence, function, protein name, and subcellular location. For new Gene objects, the nucleotide sequence is retrieved from NCBI databases. Finally, data is loaded into the databases using Neomodel for Neo4j and Mongoengine for MongoDB. This pipeline is managed by Luigi which allows the automatic execution of these tasks in the correct order, managing the dependencies and handling the errors.

3.3.8 Application Programming Interface

The repository was built to serve as a metabolic data source for reconstructing plant GSMMs. Hence, this repository should allow other applications to access the data. Thus, an API was built to facilitate access to the repository data through programmed requests. In practice, two APIs were created, one for MongoDB and one for Neo4j, using Mongoengine and Neomodel, respectively, to perform CRUD operations in the databases. These were implemented using Django and Django REST framework that allows serializing database objects into several formats, like JSON, and provides a browsable interface for the API. This API is available at https://iplantsdb.bio.di.uminho.pt/ and includes a documentation schema created with the *drf-spectacular* package.

Different views were created for each database: Neo4j was used to get the relationships between objects while MongoDB was used to get the metadata of an object. For instance, the reactions associated with a certain enzyme are retrieved from Neo4j while enzyme sequences are retrieved from MongoDB. The APIs include views to get all objects in the repository, details of an object, and create a new metabolic model object and link it to the related reactions, metabolites, and enzymes in the database. In addition, a view is available for updating the database. However, as MetaCyc requires a paid subscription, details of a subscribed account are required as input to update the MetaCyc data in the repository.

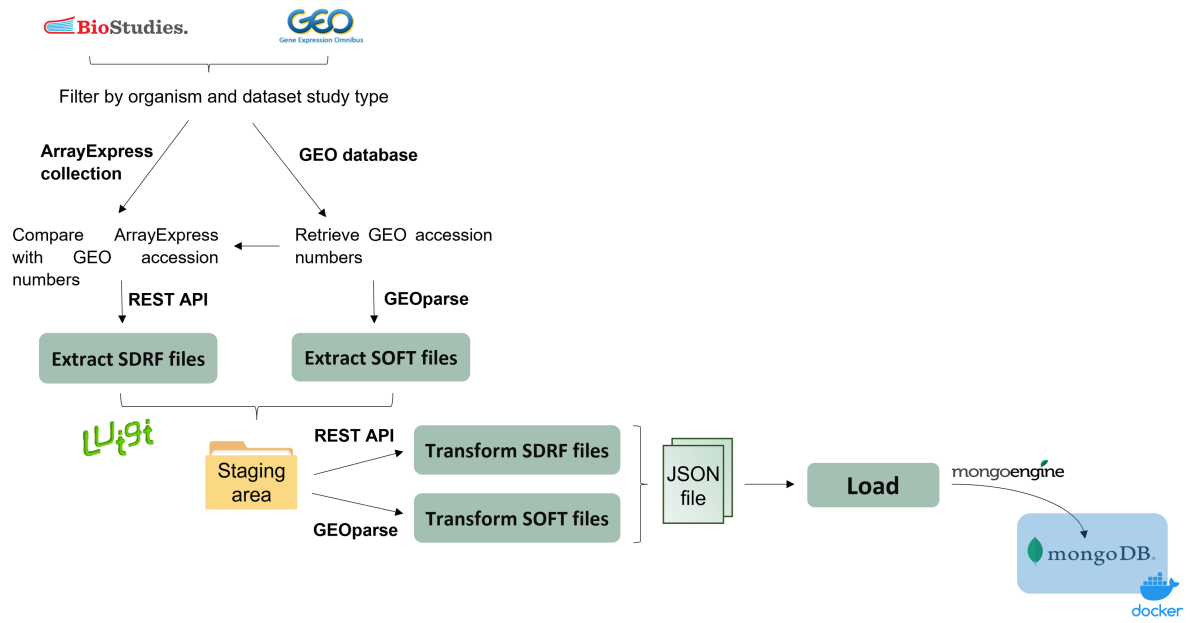


Figure 23: Pipeline for omics metadata update in *iplants*. The Biopython Python package is used to remotely access the GEO database and get the accession numbers of the studies of interest, filtering by organism and dataset type. Then, the GEOparse package is used to extract the SOFT files from the retrieved accession numbers. BioStudies database is accessed through its API to get the metadata of *ArrayExpress* studies and again, the accession numbers are filtered by organism and study type. These accession numbers are compared to the previous ones of GEO and the ones in common are ignored in this step as their metadata has already been extracted. For the unique accession numbers, metadata is extracted in JSON and SDRF formats. The files are then parsed and filtered in the transformation step and saved into JSON files with the desired structure to be loaded to MongoDB using Mongoengine. No integration rules are defined as only the objects with new accession numbers are loaded. This pipeline is managed by Luigi which executes these steps automatically and handles the dependencies and errors.

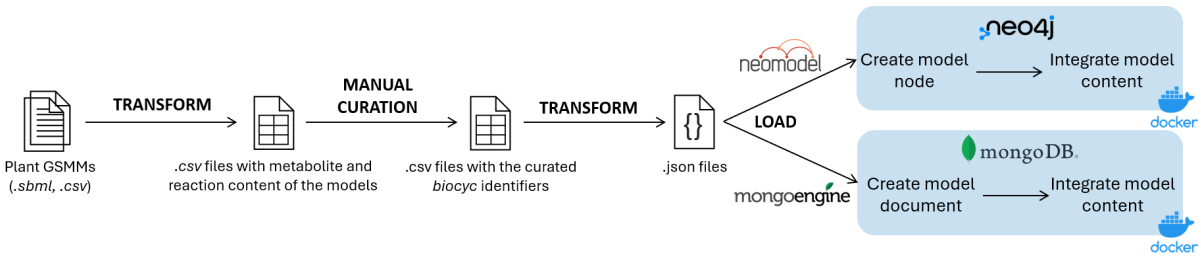


Figure 24: Pipeline for integrating the content of plant GSMMs in *iplants*. Plant GSMMs are manually downloaded and parsed using CobraPy or pandas packages. The relevant information is filtered from the models and saved into CSV files with the desired fields for integration. A manual curation step is mandatory for correct data integration in the repository. A new column named "biocyc" must be added to the CSV files to define the object identifier in MetaCyc or PlantCyc databases. After manual curation, the CSV files are then transformed into JSON files with the desired structure. Before integrating the data, a new object representing the metabolic model needs to be created in both databases. Then, the objects are loaded according to the integration rules and are linked to the metabolic model. This pipeline is not performed automatically because of the manual curation step that is crucial for correct data integration.

3.4 Data integration results

3.4.1 Central Data Storage statistics

The CDS is mainly composed of metabolites, reactions, and enzymes as well as the relationships between these entities. The number of instances of each entity is present in Figure 25 for both MongoDB and Neo4j. According to the figure, the metabolite, reaction, and enzyme instances in MongoDB account for nearly 80% of all objects in the database. Genes represent 17% of the database objects while pathways and organisms represent only 4% each. These results also reflect the integration of nine plant GSMMs into the databases as well as the omics metadata in MongoDB. In total, 221 objects were created to save the metadata of *V. vinifera* omics datasets. Of these, 193 were from GEO and only 28 were retrieved from *ArrayExpress*. Most of these objects were related to experiments using microarray data (67%).

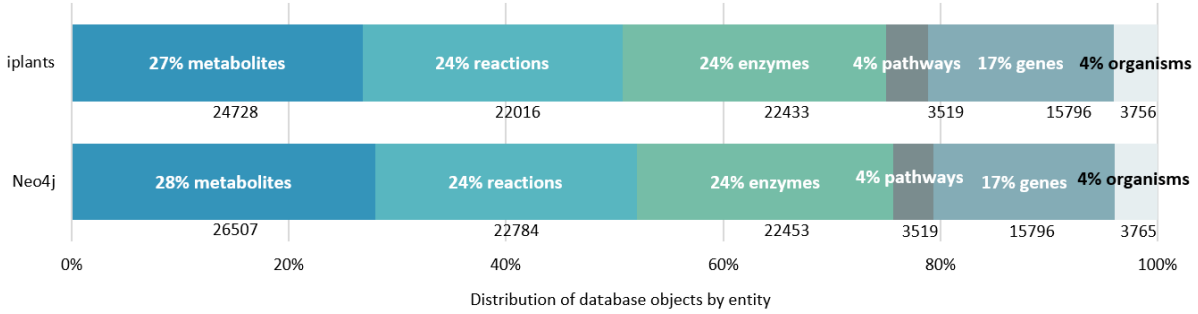


Figure 25: Distribution of database objects by entity for MongoDB and Neo4j. The stacked bars show the relative frequency of the objects for each entity, namely metabolite, reaction, enzyme, pathway, gene, and organism. Metabolic model and omics objects are not shown as these represent less than 1% of database objects.

Regarding Neo4j, similar results are observed at the node level. The number of metabolite, reaction, enzyme and organism instances is different in MongoDB and Neo4j because in Neo4j nodes are also created for entries that are not listed in the list of objects but are associated with another object. For instance, a reaction can have a product that is not included in the metabolite list. Hence, a new node is created for this metabolite and linked to the reaction node. In MongoDB, the product is listed in the *products* attribute of the reaction but no metabolite instance is created as no metadata is available for this metabolite. In total, Neo4j has more 1779 metabolites, 768 reactions, 20 enzymes, and 9 organisms than MongoDB.

In addition, Neo4j includes 247332 relationships between nodes which represent 72% of all database objects (nodes and edges) (Figure 26). From these relationships, more than 50% are between reactions and metabolites (42%) and reactions and enzymes (13%). The interactions between pathways and reactions account for 7% of the edges in the database while 14% represent the relationships of organisms with enzymes and pathways. The links of metabolic models with metabolites and reactions represent 13% of all relationships in the database.

As the final goal is to use the repository for reconstructing plant GSMMs, the sequence content in the database is very important to allow for a genome annotation as accurate and complete as possible. In

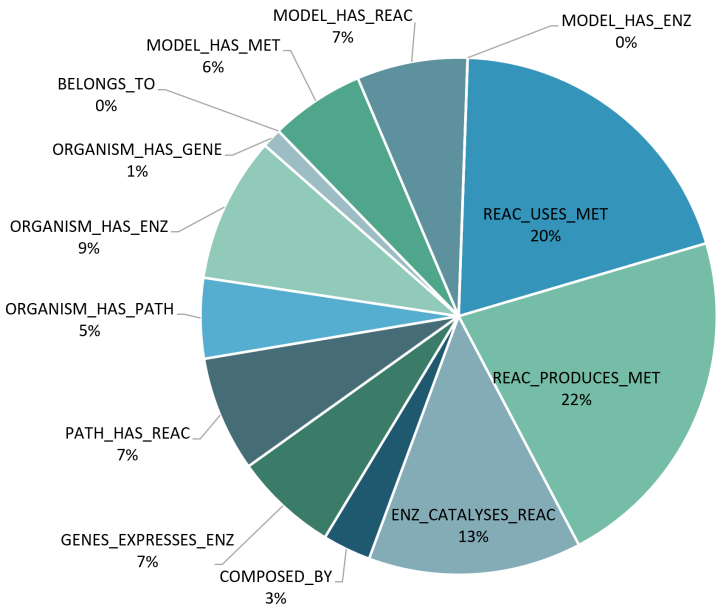


Figure 26: Distribution of relationships in Neo4j. The pie chart represents the relative frequency of each relationship type in the database.

MongoDB, 72% of enzymes have amino acid sequences and 40% of genes have nucleotide sequences. Improving these results is a very challenging task as it requires mapping the *iplants* identifiers with the identifiers from the genome annotation and this mapping is not always available.

Usually, during model reconstruction, there is the need to search for additional information on a reaction or enzyme in other databases. Hence, it is important to map the identifiers of this repository with the identifiers of other databases. This mapping was saved in the *cross-references* attribute for different databases, such as KEGG, MetaNetX, ChEBI, PubChem, and UniProt. In addition to the sequence, location and function data of enzymes was retrieved from UniProt to improve object annotation. Figure 27 shows the relative frequency of metabolites, reactions, and enzymes that have some attributes that are specific to the entity.

All objects in the database have the *entry_id* that is the identifier in the created repository. Almost all metabolites have a name, and more than 60% have cross-references for PubChem and MetaNetX databases. This percentage is lower for ChEBI (50%) and even lower for KEGG (only 34%). Regarding reactions, fewer objects have the attributes under analysis: 73% have an EC number, 65% a MetaNetX identifier, but only 30% have a KEGG identifier and only 14% have a name.

Enzymes in the database were further annotated using data from UniProt when a UniProt identifier was available, which was the case for 63% of the enzymes. From these, 62% had an associated function, 48% had a protein name (product) and 27% had a subcellular location defined. In addition, 80% of all enzymes in the database have a name, and, as stated before, 72% have a protein sequence that was extracted from UniProt or TAIR, or databases. Hence, these results show that extracting UniProt data has significantly improved the functional annotation of the enzymes and it was crucial to get protein sequence

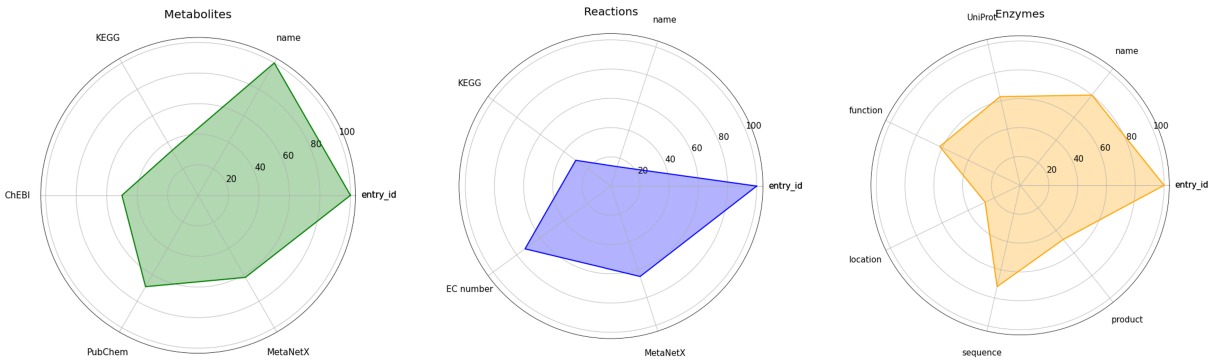


Figure 27: Relative frequency of objects having each attribute. The analysed attributes depend on the entity. For metabolites, the attributes include *entry_id*, name, and the identifiers for KEGG, ChEBI, PubChem, and MetaNetX databases. For reactions, the same attributes *entry_id*, name, KEGG, and MetaNetX identifiers were analysed as well as the EC number. Finally, the enzyme attributes include *entry_id*, name, UniProt identifier, function, location, product, and sequence.

data.

Regarding organisms (Figure 28), the species belonging to the Viridiplantae kingdom represent 18% of all organisms in the database. 60% of the organisms are bacterial species while 10% are fungi. Archaea and Metazoa represent 5% and 4% of the organism objects. The remaining organisms are other eukaryotes (3%) and viruses (0.4%). More than 50% of the enzymes in the database are from bacteria while only 21% are from plants. This reflects that the knowledge about bacterial metabolism is way larger than that of plant metabolism.

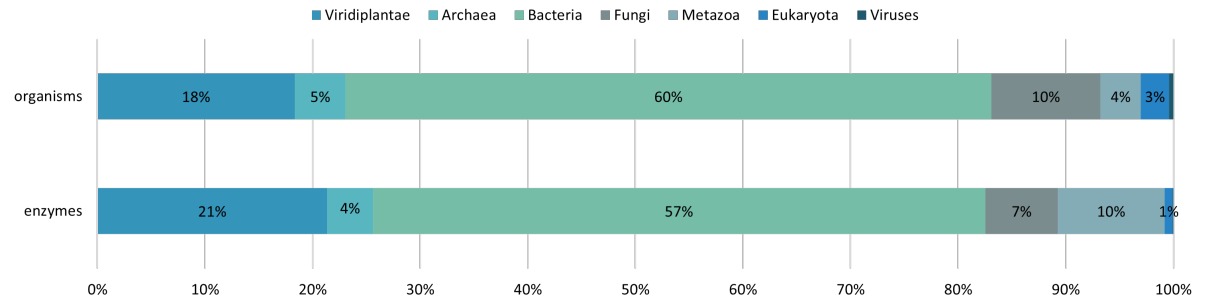


Figure 28: Distribution of organism and enzyme nodes by kingdom/ superkingdom. The stacked bars show the relative frequency of the objects distributed by different kingdoms/superkingdoms of life, mainly Viridiplantae, Archaea, Bacteria, Fungi, Metazoa, Eukaryota, and Viruses.

3.4.2 Integration results

The integration results are shown in the following tables for both sources (Table 8) and the nine plant GSMMs (Table 9). As the results were similar, the statistics were retrieved from MongoDB. The only difference between databases is the extra number of nodes created in Neo4j, as explained before, which did not affect the integration results. All 19096 objects from PlantCyc were loaded into the database. The number of objects from MetaCyc was significantly higher, surpassing the 88000 objects. Of these, 16745

objects were also on PlantCyc, so their data from both sources was integrated according to the integration rules. Then, 71259 new objects were created when loading MetaCyc.

Therefore, MetaCyc objects represent a large part of the database. Most objects from PlantCyc were also in MetaCyc. For metabolites, reactions, pathways, and organisms of PlantCyc, more than 90% were also on MetaCyc while for genes and enzymes, this percentage is lower but still higher than 70%.

In total, 24333 metabolites were loaded, 19% of which are shared between the two sources. Regarding the other entities, 24% of reactions, 31% of pathways, 15% of enzymes, 14% of genes, and 17% of organisms were present in both sources.

Table 8: Statistics for the integration of PlantCyc and MetaCyc databases in MongoDB. The table describes the number of objects of each source that were loaded into the database and shared by both sources. These numbers are distributed by entity, namely metabolites, reactions, enzymes, genes, pathways, and organisms.

	number of objects						
	metabolites	reactions	enzymes	pathways	genes	organisms	total
PlantCyc objects after loading	4807	5234	4216	1163	2974	702	19096
MetaCyc objects before integration	24185	20183	21506	3445	14976	3709	88004
unique to MetaCyc	19526	15284	18217	2356	12822	3054	71259
unique to PlantCyc	148	335	927	74	820	47	2351
objects shared by both sources	4659 (19%)	4899 (24%)	3289 (15%)	1089 (31%)	2154 (14%)	655 (18%)	16745 (19%)
total objects loaded to *iplants*	24333	20518	22433	3519	15796	3756	90335

Regarding model integration, the results are available in Table 9. These results only include the integration of the objects in the metabolic models with the objects in *iplants* and not with objects from other models. Only the metabolites and reactions in the models were integrated with the database as protein or gene identifiers in models were not standardised and the mapping was not possible.

In the transform step, the number of metabolites and reactions in the model decreases, as these objects are grouped by the BioCyc identifier. For instance, the same metabolite can be in more than one compartment in the model but it corresponds to only one entry in the database. In fact, *medicago2018* model had the highest difference between the number of objects in the model and the database as it includes more metabolites and reactions in different compartments.

In total, the plant models included 4210 metabolites and 5695 reactions. From these, 91% of the metabolites and 74% of the reactions had a match in the database, meaning that these objects were also on PlantCyc or MetaCyc sources. This percentage is not higher for reactions due to the lack of standardisation in defining identifiers for transport reactions. In Neo4j, 200 extra model metabolites matched the extra metabolites in the database.

The model with more objects integrated into the database (around 95%) was the *athaliana2009* model while the *medicago2018* model was the one with a lower percentage of integrated objects, around 82%. This was expected as the *medicago2018* model has more compartments and transport reactions than the other models. These two models were also the ones with the highest and lowest number of reactions integrated, respectively. Regarding metabolites, the *tomato2015* model has the highest integration percentage while the *rice2017* has the lowest.

Table 9: Statistics for the integration of metabolite and reaction contents of the nine plant GSMMs. The table describes the number of metabolite and reaction objects of each model integrated into *iplants*.

	number of objects					
	before transforming		**before integration**		**integrated**	
	metabolites	reactions	metabolites	reactions	metabolites	reactions
A. thaliana 2009 [129]	1297	1406	1234	1359	1211 (98%)	1264 (93%)
A. thaliana 2013 [138]	2739	2769	2275	2581	2155 (95%)	2168 (84%)
O. sativa 2013 [130]	1535	1735	1433	1693	1382 (96%)	1477 (87%)
O. sativa 2017 [132]	1380	1136	1224	1092	1014 (83%)	971 (89%)
S. lycopersicum [160]	2078	2143	1656	1964	1631 (98%)	1669 (85%)
Z. mays [136]	1125	1535	943	1383	876 (93%)	1157 (84%)
M. truncatula [162]	2884	2909	1374	2014	1235 (90%)	1554 (77%)
S. viridis [164]	2429	2473	2055	2308	1958 (95%)	2037 (88%)
G. max [163]	2814	3001	2402	2779	2255 (94%)	2377 (86%)

Then, the metabolite and reaction contents of the models were compared pairwisely between models, and the results are depicted in the heatmaps of Figure 29. Regarding metabolites (Figure 29 A), the *soybean2019* and *setaria2019* models share the highest number of metabolites, around 1790, which corresponds to 74% of all *soybean2019* metabolites and 87% of *setaria2019* metabolites. The *athaliana2009* model was the one with the highest percentage of shared metabolites, around 93%, with the other model of *A. thaliana* (*athaliana2013*). The remaining models also share more metabolites with the *athaliana2013* than with the others. Finally, the *athaliana2013* model shares more metabolites with the *soybean2019* model. All of the metabolites of *setaria2019* are shared with at least one other model. The *rice2017* model has the highest number of metabolites not shared with any model, around 124 (10%).

Regarding reactions (Figure 29 B), *soybean2019* and *setaria2019* models are once again the ones that share the highest number of reactions, around 1919, which corresponds to 69% of all reactions from *soybean2019* and to 83% of all reactions from *setaria2019*. Again, the *athaliana2009* model shares the highest percentage of reactions (88%) with the *athaliana2013* model. The remaining models also share more reactions with the *athaliana2013* than with the others, except for the *rice2013* model that shares more reactions (67%) with the *tomato2015* model. Finally, the *athaliana2013* model shares more reactions with the *soybean2019* model. The *athaliana2009* model shares almost all of its reactions with at least one other model (around 98%) while *medicago2018* has the highest number of reactions not shared with any model that corresponds to 16% of its reactions. Once again, this can be explained by the fact that this model has a higher number of transport reactions whose identifiers do not match the identifiers in the other models.

Although two models of *A. thaliana* and *O. sativa* are being compared, the *athaliana2013* model shares more content with the *soybean2019* model than with the other *A. thaliana* model, and the models of *O. sativa* share more content with those of *A. thaliana* and *S. lycopersicum* than between each other. Therefore, although these models were reconstructed from the genome of different species, they are based on the same metabolic information available in the databases, mainly of *A. thaliana*.

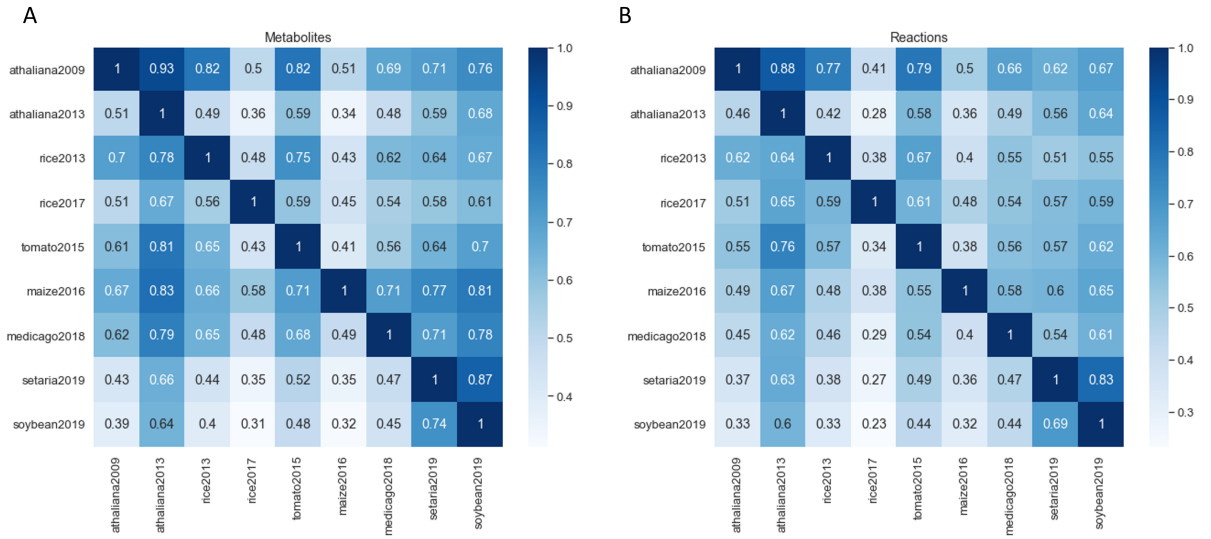


Figure 29: Heatmap representation of the comparisons between the integrated plant models at metabolite (A) and reaction (B) levels. The values of each row were divided by the total number of metabolites/ reactions in each model, which corresponds to the value in the diagonal of the matrix. Hence, the shown values represent the percentage of metabolites/reactions in a model that are in common with each of the other models. Different rows correspond to different models so the percentages are not comparable between rows.

Additionally, the out-degree of metabolites and reactions to a metabolic model was determined (Figure 30). The out-degree of an object is represented by the number of outgoing relationships with other objects. In this case, the out-degree was determined for metabolite and reaction nodes by the relationships associated with metabolic models. This way, it is possible to obtain the number of metabolic models that most metabolites and reactions are linked to. In this analysis, only the metabolites and reactions associated with at least one model were considered.

As expected, most metabolites and reactions are only present in one metabolic model and the number of objects decreases as the out-degrees increase until reaching the value 7. Then, there's an increase in the number of objects present in eight and nine models, which is more evident for the metabolites. This means that more metabolites are shared between the nine models than between only seven models.

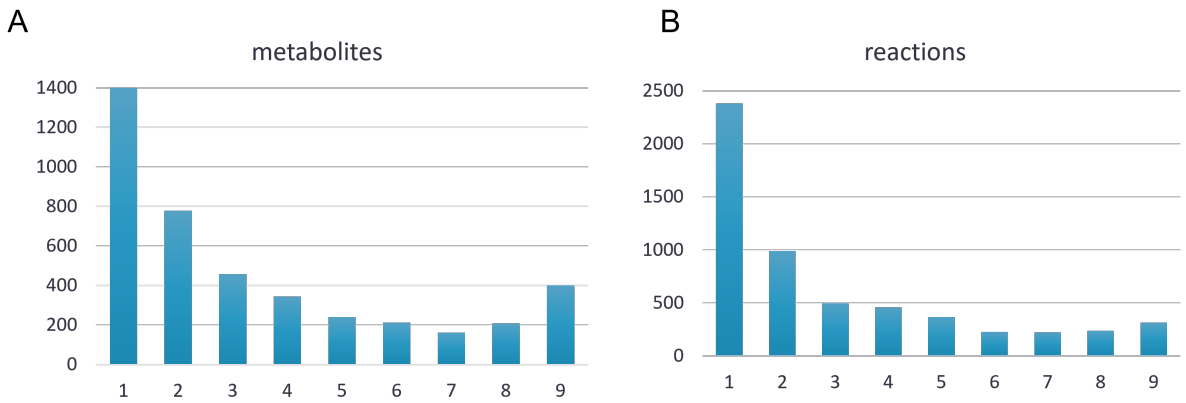


Figure 30: Out-degree frequency data of metabolites (A) and reactions (B) to metabolic models in the database. The number of outgoing relationships towards the metabolic model nodes calculates the out-degree of metabolites and reactions.

4

Reconstruction of a genome-scale metabolic model for *Vitis vinifera*

The work presented in this chapter has also been partially included in the publication submitted (available as a pre-print) and waiting for revision:

Sampaio, M., Rocha, M., and Dias, O. A diel multi-tissue genome-scale metabolic model of Vitis vinifera. Submitted. Pre-print available at: https://doi.org/10.1101/2024.01.30.578056

4.1 Introduction

Vitis vinifera is one of the most important fruit crops in the world. It is cultivated worldwide and has high economic value, mainly due to wine production [260].

In 2022, the world vineyard surface area was estimated to be 7.3 million hectares and world wine production reached 258 million hectolitres. In the same year, and despite inflation, wine exports reached a value of 37600 million euros [3]. In addition, grapes have other purposes, being marketed as fresh and dried fruits and used for juice production.

The grape pulp contains a high level of sugars and phenolic compounds, like flavonoids and stilbenes, with potential health benefits, such as antioxidant and anti-inflammatory activity and cardiovascular protection, thus being currently studied for possible pharmaceutical and cosmetic applications [261]. Resveratrol is the most studied stilbene and is produced by plants in response to pathogenic attacks or ultraviolet (UV) irradiation. This phenol has been intensively studied for its potential cardioprotective, antimicrobial, antioxidant, and anticancer activities. Currently, resveratrol is used in cosmetic formulations due to its anti-ageing and skin-lightening effects.

Therefore, as grapevines and their metabolites have a high economic interest and fruit quality is intrinsically linked to metabolism, the study of grapevine metabolism is essential for understanding its responses to different environmental conditions that may affect grape metabolic composition. Metabolism has been studied by Systems Biology approaches, which use GSMMs to analyse biological systems as a whole, modelling the inner components and their respective interactions [262].

GSMMs are reconstructed from the genome, representing all metabolic reactions taking place within an organism. These models allow performing flux simulations of metabolic phenotypes under different environmental or genetic conditions [263]. Although GSMMs have been extensively used for the metabolic engineering of prokaryotes, several GSMMs are available for plants [264], mainly *Arabidopsis thaliana* [129, 134, 138, 139, 148–150], *Zea mays* [134–136], and *Oryza sativa* [130, 132, 133].

Currently, the reconstruction of plant GSMMs is still very challenging and time-consuming due to the high number of gaps in genome annotations, the large diversity of metabolites, and the extensive compartmentalisation of plant cells [43, 44]. Despite the obstacles, many plant GSMMs have emerged recently, and new approaches have been developed to reconstruct more realistic models that include different plant tissues, through the integration of omics data, as well as the day-night cycle [148, 167, 265]. The integration of omics data allows differentiating the metabolism of different tissues and multi-tissue models allow the analysis of the metabolic interactions between tissues [167]. Moreover, the introduction of a diel cycle in the model captures the metabolic interactions between light and dark phases [148, 150].

Following these advances, we pioneer *V. vinifera* research with the introduction of iMS7199, the first GSMM for the grapevine, developed using the most recent genome version, PN40024.v4 [79]. In addition to the overarching model, tissue-specific models for the leaf, stem, and grape were developed by incorporating RNA-Seq data from these distinct tissues. Furthermore, to capture the changes in grape metabolism, we created two separate models representing the grape in both its green and mature states.

These tissue-specific models were then integrated to construct diel multi-tissue GSMMs, enabling phenotype predictions of grapevine metabolism across the day-night cycle and facilitating the study of inter-tissue metabolic interactions. Using this comprehensive model, we investigated the metabolic responses of the grapevine under varying concentrations of sulfate and nitrate.

4.2 Model Reconstruction

The reconstruction of the GSMM of *V. vinifera* was performed using Python 3.8, CobraPy 0.25, and the repository API for getting the data. A Python class was created to allow the execution of the automatic steps in the reconstruction process. The main steps in model reconstruction are depicted in Figure 31 and are described in the next sections. The code is available at the https://github.com/martaS95/iplantsdb_app GitHub repository.

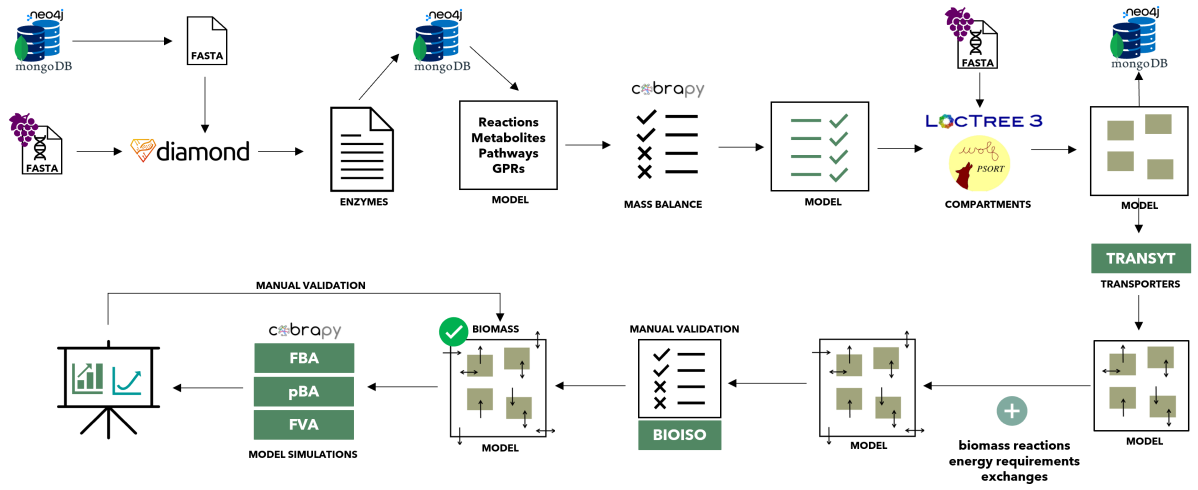


Figure 31: The main steps of the *V. vinifera* GSMM reconstruction. Genome annotation is performed using Diamond with the protein sequences from the database and *V. vinifera* proteins from the genome. The identified enzymes are used to get all metabolic information from the database as well as to define the GPR rules in the model. After, the mass balance of reactions is checked and compartments are predicted using LocTree3 and Wolfpsort tools. Then, transporters were predicted using Transyt, and biomass, exchange, and energy requirement reactions were defined. Finally, the model was manually validated, using BIOISO to verify biomass production. The final model was simulated using different methods, such as FBA and FVA, and the results were analysed to validate the model.

4.2.1 Genome Annotation

The reconstruction of a GSMM starts with genome annotation, which will provide the information needed to build a draft metabolic network. The PN40024.v4 genome of *V. vinifera* and the PN40024.v4.1 annotation version were downloaded from The Grapevine Genomics Encyclopedia (Grapedia) (available at: https://grapedia.org/), which is an open-access platform that integrates all grapevine knowledge, resources, and services for research and industry. This genome was recently published (March 2023) and includes 35922 genes and 41160 proteins [79].

The annotation was done at the protein level using DIAMOND, which performed similarity searches between the proteins of the *V. vinifera* genome and the proteins in *iplants*. The goal was to assign enzymatic functions and associate reactions with *V. vinifera* proteins.

First, the protein sequences in *iplants* were extracted in FASTA format and were used to build a database for DIAMOND. This tool was available on our server and was run through an API, using the genome proteins as input. It also required the definition of some submission parameters, including an e-value threshold, which was set to 1^{-10}, the maximum number of matches returned, which was set to 100, and the alignment mode, which was defined as ultra-sensitive. These parameters were chosen to return the most significant hits (the default e-value was 0.001) with high alignment sensitivity, allowing the tool to detect weak sequence similarities. The results were downloaded, parsed, and saved in JSON format. Only the first hit was considered for each genome protein. A Python class named *Diamond* was implemented, comprising methods to submit the inputs to the DIAMOND's API, download the results, and parse and transform the results into JSON files.

Using only the iplants database to perform the annotation facilitated the identification of the reactions to include in the model. DIAMOND searches were performed against the SwissProt database to determine the completeness of important enzyme annotations in iplants. In the course of these searches, it was found that more than 17000 *V. vinifera* proteins had a match in SwissProt. However, most proteins were involved in other processes like transcription regulation, protein phosphorylation, and nuclear transport, which are not relevant to be included in the metabolic model.

4.2.2 Network Assembly

After running DIAMOND, the mapping between *V. vinifera* genome proteins and the iplant enzymes was available and it was used to get the list of enzymes to include in the model. Then, the Neo4j database was used to get the reactions that are associated with the identified enzymes to also include them in the model. Similarly, the metabolites involved in these reactions and the pathways associated with the reactions were retrieved from the database.

As this data was collected, it was stored in the repository. First, a document in MongoDB and a node in Neo4j were created for the metabolic model. Then, the annotation and the list of enzymes were added to the model document in the appropriate attributes while, in Neo4j, the model node was linked to the enzymes through the creation of "MODEL_HAS_ENZYME" relationships. Similarly, the lists of reactions, pathways, and metabolites were added to the model document in MongoDB (Figure 32), and the relationships between the model node and the reaction and metabolite nodes were created in Neo4j (Figure 33). All these steps were run automatically, using the methods from the *ModelReconstruction* Python class.

The GPR rules were also retrieved automatically from the MongoDB database. For each reaction, the list of associated enzymes was selected and, for each enzyme, it was verified if the enzyme was a protein complex. This was done by checking the attribute "components" in the database document for the identified enzymes. If this attribute had a value, the enzyme was considered to be a protein complex.

In that case, all the monomers (components) composing the protein complex need to be present for the reaction to occur. In this case, the AND rules were applied. Otherwise, the OR rules were applied and the different enzymes associated with a reaction were considered to be isoenzymes. The GPR rules were then saved into a CSV file and loaded to the metabolic model document in MongoDB (Figure 32).

Figure 32: Metabolic model document in MongoDB. All data to include in the model is saved in the attributes *annotation*, *enzymes*, *reactions*, *metabolites*, *pathways*, and *GPR rules*.

After having the list of metabolites, reactions, proteins, pathways, and GPR rules to include in the metabolic model, the next step was to check the mass and charge balance of the reactions. Hence, the list of unbalanced reactions was retrieved and manually analysed. In some cases, the formula of certain metabolites was missing in the database and was manually added using other sources, such as MetaNetX [266] or ChEBI [267]. In other cases, only a proton or a water molecule was missing or in excess either on the reactants or the products of the reaction. These were added or removed accordingly.

Many reactions were composed of classes of metabolites. In the cases where the reaction was unbalanced, the class was replaced by an instance metabolite whenever possible. For instance, the metabolite class "Ubiquinones" was replaced by the instance "UBIQUINONE-10". The list of all changes is presented in Supplementary File 2. Metabolites with variable formulas but necessary for the model to function, such as *Fatty acid* or *Acyl-carrier-protein* were kept in the model.

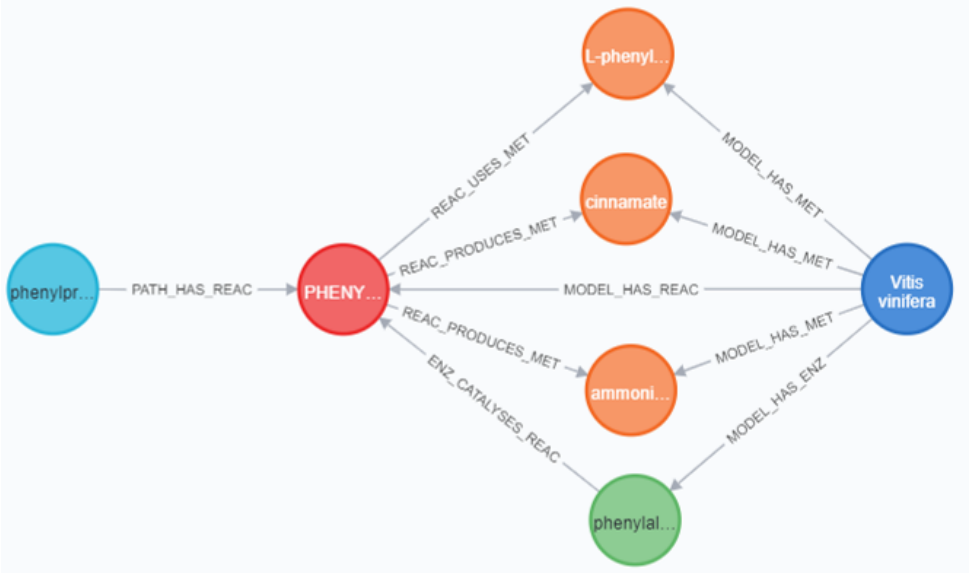


Figure 33: Simplified example of the relationships of the Metabolic Model node in Neo4j. This node is linked to the enzyme phenylalanine ammonia lyase (green) through the relationship MODEL_HAS_ENZ and it is linked to the associated reaction (red) and its metabolites (orange) through the relationships MODEL_HAS_REAC and MODEL_HAS_MET, respectively.

Non-essential reactions where balancing was not possible were eliminated (Supplementary File 3). Most of them corresponded to glucan transformations, such as hydrolysis, which converts a glucan with n monomers into a glucan with $n-1$ monomers. For these reactions, the same metabolite class was on both sides of the reaction. Other deleted reactions had protein, DNA, and RNA compounds as reactants or products. The transport reactions retrieved from the database were also deleted as the same metabolite was on both sides of the reaction and it was not possible to define the compartments of the metabolite automatically. Then, the central metabolic pathways were briefly analysed and the missing spontaneous reactions were added to the list of reactions.

The next task was the compartmentalisation of the network, which consisted of assigning the reactions to the subcellular location where they take place. As plant cells present a large number of compartments, this task is particularly important and it substantially increased the size and complexity of this network.

WolfPsort [46] and LocTree3 [47] were used to predict the subcellular location of the proteins based on the protein sequence, but manual curation of the results was required. The results obtained from these tools were parsed and the location of the reactions was inferred from the predicted protein location using the GPR rules. In the cases where protein location was unknown, reactions were assigned to the cytosol. The thylakoidal and mitochondrial intermembrane compartments were added manually for the photosynthesis and oxidative phosphorylation reactions, respectively.

This information was saved in MongoDB in the attribute *models* of metabolites and reactions, through the repository's API. As similar reactions or metabolites occurring in different compartments need to have different identifiers reflecting the compartment they are on, a new identifier was created by adding a suffix for the compartment to the *entry_id* of the object. This identifier is then added to the *models* attribute with the corresponding compartment (Figure 34).

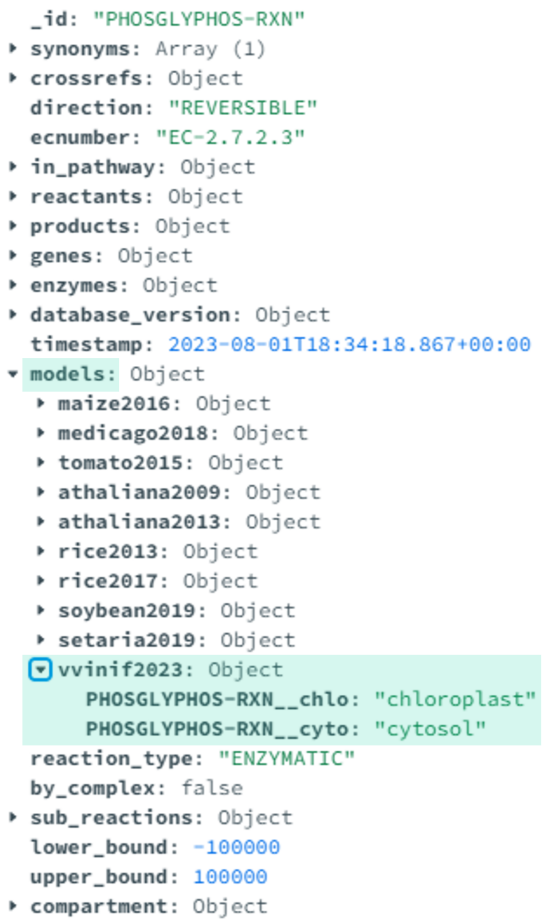


Figure 34: MongoDB document for the phosphoglycerate kinase reaction (PHOSGLYPHOS-RXN), highlighting how the compartment data is saved. In the attribute *models*, it is possible to see which models contain the reaction and in which compartments. For the *V. vinifera* model (*vvinif2023*), this reaction is located at the chloroplast and cytosol.

After compartmentalisation, transport reactions needed to be identified to ensure the connectivity of the network. This was accomplished using the TranSyT tool [268], which retrieves information from TCDB and performs the annotation using BioSynth through TC numbers. This tool is available on *merlin* [166], but it was run using the tool's API to perform transporter annotation. All transporters were validated for mass and charge balance. Additional transporters were manually included in the model when required for biomass production or the production of metabolites of interest. When no information was available, the transport reaction was added without gene association to ensure model functionality.

Then, the next step was to convert the network into a model and perform validation. For this, the biomass equation was formulated, the drains were added, and the model was manually revised and gap-filled to ensure biomass production.

4.2.3 Model Assembly

As mentioned before, the biomass equation represents the biomass formation from complex macromolecules. Ideally, experimental data should be used to define the biomass composition of an organism.

However, as this data is not available for *V. vinifera*, the biomass composition and the energy requirements were defined based on previously published plants GSMMs and available literature [139, 160, 165, 269]. The details for the biomass composition are available in Supplementary File 4.

Biomass is composed of macromolecules labelled "e-metabolites", which are required for cell growth. Based on the previous models, the biomass equation for *V. vinifera* model includes RNA, DNA, proteins, carbohydrates, lipids, co-factors, and cell wall components. Reactions for producing each of these macromolecules were created.

The monomer contents for the production of DNA, RNA, and proteins were calculated from the genome sequence using the e-Biomass tool [270]. This tool allows the estimation of biomass composition in amino acid and nucleotide content from genomic information, without the need for experimental data. With this information, the reactions for producing DNA, RNA, and proteins were created.

The reactions for the production of carbohydrates, lipids, cell wall components, and co-factors were adapted from *A. thaliana* [139] and *Q. suber* [165] models. The e-Cofactor metabolite includes several universal cofactors, such as NAD(H), vitamins, and pigments. The content of carbohydrates was adapted to reflect the grape composition described in the literature. Although fatty acids are not directly included in the biomass reaction, they are components of lipids and are needed for biomass production. Hence, a metabolite representing the average fatty acid composition (*Fatty_Acids*) was created as well as the reaction to produce it. The molecular weight of this metabolite allowed determining the molecular weight of the lipids.

Similarly, growth and non-growth energy requirements were retrieved from previous plant models, mainly *A. thaliana* [139] and *Q. suber* [165] models, as no experimental data was available for *V. vinifera*. Hence, the growth energy requirement was added to the biomass reaction whereas the non-growth energy was included in the model in a simple ATP hydrolysis reaction.

The main exchange reactions were defined based on previous models that include the essential metabolites for biomass production. These are light, iron(II), magnesium, sulfate, nitrate, phosphate, water, oxygen, carbon dioxide, protons, amino acids, and sugars, such as sucrose, glucose, fructose, and maltose. Later, more exchange reactions were manually added for produced metabolites, according to the transporters in the metabolic network and the expected behaviour.

The draft model was then exported in the SBML format to be validated, using the method from the *ModelReconstruction* class. This method connects to MongoDB's API to get all the data needed for creating the model, including the list of metabolites, reactions, pathways, GPR rules, and compartments.

4.2.4 Model Validation

The first step of model validation was to ensure that biomass was produced. BioISO [271] was used to accomplish this goal, as it verifies the production of each biomass substrate to see which ones are missing. Each gap affecting biomass production was manually analysed and filled when necessary for the production of a biomass precursor. During this process, in addition to finding and filling the gaps of missing enzymatic or transport reactions, errors in the genome annotation and reaction reversibility

were also detected and fixed. Literature and biological databases, such as KEGG [27], BRENDA [31], and UniProt [30], were consulted to assist this process.

Then, dead-end metabolites and blocked reactions were identified using CobraPy. Each blocked reaction was analysed and fixed resorting once again to other databases. When no information was available, the blocked reactions were kept in the model. In addition, some reactions were added to the model to assure the production of certain metabolites that are known to be produced by *V. vinifera*, according to the literature or metabolomics datasets, even when no gene associations were found.

Also, it was assured that growth did not occur without photon and carbon sources, and no futile cycles were present. Finally, it was verified that the model was able to simulate the different metabolic processes of photosynthesis, photorespiration, and respiration, as desired.

4.2.5 Phenotype Predictions

Phenotype predictions were performed by pFBA, using two different strategies as applied in other plant models [139, 165], and based on *A. thaliana* experimental measures [272]. The first consisted of fixing the biomass growth rate to $0.11 h^{-1}$ and the minimisation of the photon/sucrose uptake was set as the objective function. The second strategy defined the biomass growth rate as the objective function and fixed the photon uptake to $100\ mmmol_{photon}.gDW^{-1}.h^{-1}$ for photosynthesis and photorespiration and the sucrose uptake to $1\ mmmol_{sucrose}.gDW^{-1}.h^{-1}$ for respiration. Photorespiration was simulated by constraining the carboxylation (RIBULOSE-BISPHOSPHATE-CARBOXYLASE-RXN) and oxygenation (RXN-961) reactions by Rubisco with a flux ratio of 3:1 [139, 165]. The fixed biomass growth rate, the uptake rates of photon and sucrose, and the RUBISCO ratio were retrieved from the AraGEM model [139] of *A. thaliana* as no rates were available for *V. vinifera* in the literature.

4.2.6 Tissue-specific Models

The reconstructed iMS7199 model is generic and includes all reactions known to occur in *V. vinifera*, regardless of the tissue, which can lead to wrong interpretations as plants are composed of a wide variety of cell types with specific metabolisms. Thus, certain pathways can be active in some tissues and inactive in others. Therefore, after curation of the generic model of *V. vinifera*, omics data were integrated with this model to create tissue-specific models for stem, leaf, and berry.

4.2.6.1 Omics Data

Available RNA-Seq transcriptomics data from different tissues were used to create specific GSMMs for stem, leaf, and berry. As there was not a single study with RNA-Seq data for these three tissues, two different studies were used. Leaf and stem data were retrieved from the healthy samples of *V. vinifera Cabernet Sauvignon* cultivar from the study of Massonnet et. al [273], which is available on GEO under the accession GSE97900. The goal of this study was to analyse the transcriptional response of grapevines to *Neofusicoccum parvum* infection. RNA-Seq data for *C. Sauvignon* berries was obtained from the study of

Fasoli et al. [274] (GEO accession: GSE98923), which included berry samples at different developmental stages. In this study, both transcriptomics and metabolomics data were retrieved from the berries to identify key events controlling berry formation and ripening.

These samples were retrieved from GRape Expression ATlas (GREAT)). The RNA-Seq data was already normalised in transcripts per million (TPM) and the reads were mapped to the new genome (PN40024.v4) [78]. In the case of berries, the data was discretized to obtain samples representing the green and mature states of grapes. Samples until time point 4 were considered to be in the green state, while samples after time point 4 were considered to be mature. Although sample time 4 was collected seven days after veraison, it was still considered to be in the green phase.

Given the need to use samples from different studies to create tissue-specific models, an additional analysis was performed to see if samples from different studies had intrinsic noise created by the sequencing platforms and experimental procedures used, or if they could be compared to each other. Therefore, additional leaf and berry samples from other studies (GSE89113, GSE74428, GSE97960, GSE76256, GSE89113, and GSE74428) were retrieved from GREAT and joined with the aforementioned samples in the same dataset. Then, a comprehensive analysis was performed using t-SNE to visualise all the data and verify if the data was grouped by sequencing platform or study. Before applying t-SNE, the data values were transformed by log2 and scaled to z-scores.

This analysis has shown that data normalisation in GREAT allows samples from different studies to be used and compared. Hence, a dataset with the chosen samples from stem, leaf, and berry and respective metadata was built. Finally, the gene identifiers in the datasets were mapped to the protein identifiers present in the model to allow for omics integration.

4.2.6.2 Models

Tissue-specific models for stem, leaf, green berry, and mature berry were reconstructed through the integration of the RNAseq dataset with the iMS7199 model, using the FastCORE algorithm [275] implemented in Troppo [276].

First, the biomass composition of stem, leaf, and berry, and the energy requirements for each tissue were defined based on previously published plant GSMMs [139, 160, 165] and available literature [269, 274, 277]. In addition, the biomass of the mature berry was adjusted according to the literature and the metabolomics data of the berry samples [274].

Then, the reconstructed generic model and the omics dataset were used as input for the FastCORE algorithm. The biomass reaction was included in the protected reactions of the algorithm so that all tissue-specific models would be able to produce biomass.

In addition, different strategies were used to define the thresholds for data integration. The default strategy does not apply any preprocessing step to the omics dataset before integration with the model. The only threshold defined is the one used by the integration algorithm. Recently, Lewis et. al [278] proposed three thresholding approaches that prepare the omics dataset before integration, defining additional thresholds (Figure 35):

- **global**: in this strategy, the genes with an expression above the global threshold are considered active while the others are inactive. The threshold value is the same for all the genes.

- **local T1**: this strategy defines a global threshold and genes whose expression is below this value are considered to be inactive. Genes with expressions above this threshold may be active and are further analysed by comparing their expressions with a local threshold. This is a gene-specific threshold that accounts for the expression levels of each gene over all the samples.

- **local T2**: this strategy is similar to local T1 but uses two global thresholds, upper and lower. Genes whose expression is below the lower threshold are considered to be inactive while genes whose expression is above the upper threshold are considered to be active. Genes with expressions between these two thresholds are further analysed and classified according to the local threshold.

Figure 35: Representation of the three thresholding strategies for omics integration: global, local T1 and local T2. Adapted from [278].

These global and local thresholds usually represent the percentiles of the gene expression data. The different threshold values tested in this work are presented in Table 10 for the different strategies. The choice of the threshold values was based on the data and the results from Lewis et. al [278].

Table 10: Threshold values tested for each integration strategy for creating tissue-specific models. The global and local thresholds correspond to the percentiles of the gene expression data.

Strategies	integration thresholds	global threshold 1	global threshold 2	local threshold
Default	6, 8, 10	-	-	-
Global	0	50%, 75%	-	-
Local T1	0	50%	-	50%, 75%
Local T2	0	25%	75%	50%, 75%

The different strategies for the creation of tissue-specific models were compared by generating flux sampling data with the Artificial Co-ordinate Hit and Run (ACHR) sampler [279] from CobraPy with a thinning factor of 100 and a sample size of 3000. From these, 1000 random samples were selected to facilitate data visualisation. Reactions with a flux of 0 across all samples were removed from the dataset and data was scaled to z-scores. The resulting datasets from the different models were visualised using t-SNEs with a perplexity of 50 to see the effect of the threshold strategy on the flux distributions of the models.

4.2.6.3 Differential Flux Analysis

After selecting the tissue-specific models, differential flux analysis (DFA) between them was performed using the approach of Nanda et. al [280]. In this approach, sample fluxes for the tissue models were generated using ACHR sampler from CobraPy, with a thinning factor of 100 and a sample size of 10000 for each model. Pairwise Kolmogorov-Smirnov tests were used to compare the flux distribution of the tissue models. The flux change (FC) of each reaction between two models was also calculated as shown in the formula 4.1. $\overline{S}_{model1}$ and $\overline{S}_{model2}$ represent the mean of the flux distributions for a reaction in *model1* and *model2*, respectively. Reactions with an absolute FC value less than 0.82 (equivalent to a 10-fold change in flux) were considered insignificant. For reactions that are absent in a model, their flux is assumed to be zero in that model, and bootstrapping is used to estimate the 95% confidence interval of their fluxes. If zero is outside the interval, the reactions are considered to have differential flux in the two models. In addition, the p-values of altered reactions were adjusted by a Benjamini-Hochberg correction, with a significance level of 0.05.

$$FC = \frac{\overline{S}_{model1} - \overline{S}_{model2}}{|\overline{S}_{model1} + \overline{S}_{model1}|} \tag{4.1}$$

The differential pathways between models were obtained using hypergeometric enrichment tests that select the pathways that are actually over-represented due to the higher number of altered reactions and not by chance. Once again, t-SNE was used to visually compare the sampling flux data of the tissue models. Before applying t-SNE, the flux data was filtered, keeping only the reactions with significantly different flux between models, and scaled to z-scores.

4.2.7 Diel Multi-tissue Models

The final tissue-specific models for stem, leaf, and berry were merged and connected to create multi-tissue models. These models allow the study of complex metabolic behaviours of higher organisms that involve interactions between multiple tissues.

A Python class named *Multitissue* was created for this end, following the previous approach [149]. It includes the creation of two common pools to connect the tissues: common pool 1 between stem and leaf and common pool 2 between stem and berry. Transport reactions between tissues and common pools were manually added when required for model functionality. In addition, a reaction for the total biomass was defined, representing the sum of the biomass of all tissues.

As no model for the root was created, the uptake of minerals was assumed to take place in the stem. Exchanges of water, oxygen, and carbon dioxide were allowed in all tissues, and light absorption was done only by the leaf. Two multi-tissue models were created, one with the berry in the green phase and the other with the mature berry.

Then, diel models were also created to account for light and dark phases. All reactions and metabolites from the multi-tissue models were duplicated for each phase, and new reactions were added to allow the

exchange of some metabolites between the two phases. These are called storage metabolites and include the 20 amino acids, nitrate, citrate, malate, glucose, sucrose, fructose, and starch, which can be produced in one phase and used in the other, as previously described [148].

Phenotype predictions of diel multi-tissue models were also performed with pFBA using the first strategy mentioned in the *Phenotype predictions* subsection for photorespiration conditions, but with a photon uptake of 300 mmmol.gDW^{-1}.h^{-1} as flux values were very low with 100 mmmol.gDW^{-1}.h^{-1}. As in other plant diel models [148, 150, 165], the nitrate uptake was constrained to a ratio of 3:2 in the light/ dark cycle.

4.3 Model Reconstruction Results

4.3.1 Genome Annotation

4.3.1.1 Enzymes

The first GSMM for *V. vinifera* was reconstructed from the PN40024.v4 genome (annotation version 1) [79]. DIAMOND similarity searches against *iplants* resulted in 10840 protein matches, representing 26% of the 41160 proteins in the genome, which is in line with the percentage of metabolic genes described for the *A. thaliana*'s genome (between 25-30%) [281].

From these matches, 811 enzymes were not associated with any reaction in the *iplants* database. Hence, only 10029 proteins were included in the model, reflecting 1577 different EC numbers. After manual curation, some annotation errors were fixed, some reactions were added to the model and others were removed, reducing the number of proteins in the model to 6421, representing 1358 different EC numbers. The distribution of the proteins in the model by EC number class is shown in Table 11. 2% of the proteins in the model were not associated with an EC number while some proteins were associated with more than one class of EC numbers.

Table 11: Distribution of the proteins in the model by EC number class.

EC class	EC numbers (%)
EC-1: Oxidoreductases	33.7
EC-2: Transferases	31.4
EC-3: Hydrolases	18.1
EC-4: Lyases	9.6
EC-5: Isomerases	3.8
EC-6: Ligases	3.6
EC-7: Translocases	2.1

Oxidoreductases and transferases are the most representative classes in the model, including 65% of the proteins. Oxidoreductases are responsible for the oxidation and reduction of several substrates using cofactors, mainly NAD+ and NADH, and are essential for maintaining the redox balance in cells. Transferases facilitate the transfer of functional groups between molecules. The model includes different

types of transferases, such as kinases, methyltransferases, and aminotransferases, which are important for the metabolism of carbohydrates, nucleotides, and amino acids.

Then, hydrolases represent 18% of the proteins in the model, which are mainly involved in the hydrolysis of ester and glycosidic bonds, which are important for lipid and carbohydrate metabolism, respectively. Lyases, isomerases, ligases, and translocases represent less than 10% of the proteins in the model each. Most lyases in the model cleave carbon-carbon bonds (EC 4.1), like decarboxylases, or carbon-oxygen bonds (EC 4.2), like dehydratases, being involved in several important pathways, such as glycolysis and photosynthesis.

Regarding isomerases, the main classes in the model are intramolecular transferases (mutases), which transfer functional groups from one part of a molecule to another, like the enzyme phosphoglucomutase, and intramolecular oxidoreductases, which are involved in the production of stereoisomers, such as dihydroxyacetone phosphate (DHAP) and D-glyceraldehyde 3-phosphate in glycolysis. Most ligases in the model form carbon-oxygen bonds (EC 6.1), carbon-sulfur bonds (EC 6.2), and carbon-nitrogen bonds (EC 6.3). The first subclass includes only aminoacyl-tRNA ligases, which catalyse the attachment of amino acids to the corresponding tRNA molecule during protein synthesis. The second class includes proteins that are mainly involved in the addition of a coenzyme A molecule to the substrate. The proteins of the third class are mainly involved in the biosynthesis of amino acids.

Finally, translocases catalyse the movement of ions or molecules across membranes, being important for maintaining overall cellular homeostasis. In the model, most translocases are involved in the active transport of protons (H+) across the membranes.

The model contains 82 incomplete EC numbers, 2.4.1 and 1.14.14 being the most represented. The first represents hexosyltransferases that are involved in the biosynthesis of complex carbohydrates, such as oligosaccharides and cytokinins, as well as glycoproteins, and glycolipids, which are essential components of cellular membranes. The latter are oxidoreductases that use a reduced flavoprotein as a donor, being involved in the biosynthesis of plant hormones and alkaloids.

4.3.1.2 Compartments

Eight different compartments were identified using WolfPSort [46] and Loctree3 [47] tools. Manual curation of the results was required as these tools exhibited contradictory outcomes. For instance, some enzymes catalysing the reactions of the sphingolipid biosynthesis, like serine C-palmitoyltransferase (SERINE-C-PALMITOYLTRANSFERASE-RXN) and dihydroceramide fatty acyl 2-hydroxylase (RXN-7796) reactions, were predicted to be on the endoplasmic reticulum by Loctree3 and in the chloroplast or cytosol by WolfPSort. In this case, the annotations collected from UniProt were considered and the location of these enzymes was defined to be the endoplasmic reticulum. When no information was available, the protein was assumed to be in the cytosol. Many proteins were predicted to occur in more than one compartment.

The proteins in the model were distributed by the cytosol, chloroplast, mitochondria, peroxisome, endoplasmic reticulum, Golgi complex, vacuole, and extracellular space (Table 12). As expected, most proteins are in the cytosol, followed by the chloroplast. The thylakoidal and mitochondrial intermembrane

compartments were added manually for the photosynthesis and oxidative phosphorylation reactions, respectively.

Table 12: Distribution of the proteins in the model by compartment.

Compartment	Number of proteins
cytosol	8951
chloroplast	3412
mitochondria	1595
peroxisome	723
endoplasmic reticulum	2625
Golgi complex	429
vacuole	405
extracellular space	1135

4.3.1.3 Transporters

TranSyt [268] was used to automatically predict the transport proteins in the *V. vinifera* genome and get the associated transport reactions. Of the 7199 proteins in the model, 824 were identified as transporters and were associated with 1087 transport reactions that were added to the model. The distribution of these proteins across TC Number classification is shown in Table 13.

Table 13: Distribution of the transport proteins in the model by TC number class.

EC class	EC numbers (%)
TC-1: Channels	31.1
TC-2: Secondary transporters	65.2
TC-3: Primary Active Transporters	26.6
TC-4: Group translocators	0.73

More than half of the transport proteins are secondary carrier-type facilitators. These proteins can catalyse different transport mechanisms using only the energy stored in ion gradients. The transport mechanisms include uniport, in which a single molecule is transported by facilitated diffusion, antiport, where two or more molecules are transported in opposite directions, and symport, when transporting two or more molecules in the same direction. Transport through channels represents 31.1% of the transport reactions included in the model. These proteins have transmembrane channels and catalyse facilitated diffusion without spending energy. Around 26% of the proteins are primary active transporters, which use a primary source of energy to perform active transport of a molecule against a concentration gradient. Finally, less than 0.73% of the proteins catalyse group translocation transports, where the transported substrate is modified during the transport process. In the model, these proteins are involved in the secretion of cellulose. No proteins were identified belonging to the remaining TC classes (TC-5, TC-8, TC-9).

Then, as plant cells are highly compartmentalised, 537 transport reactions were manually added to the model to ensure network connectivity and biomass production. Given the complexity of the network and the high number of new transporters, these reactions were not associated with any genes.

4.3.2 Biomass Composition

As mentioned before, the biomass equation for *V. vinifera* includes seven different reactants, labelled as "e-Metabolites", that represent the complex macromolecules produced by the plant to grow. These include RNA, DNA, proteins, carbohydrates, lipids, co-factors, and cell wall components.

As experimental data was not available for *V. vinifera*, the biomass composition and the energy requirements were defined based on previously published plant GSMMs and available literature [139, 160, 165, 269] (Figure 36 and Supplementary File 4). The differences between biomass compositions observed in the figure can be explained by the fact that the AraGEM model does not include any co-factors in the biomass equation and the percentage of e-lipids is lower than in the other models, as this model only includes the biosynthesis of palmitic acid. Also, the *Q. suber* model has a higher percentage of e-carbohydrates and a lower percentage of e-cell wall components because cellulose and xylose metabolites are included in the e-carbohydrate reaction while, in the other models, these are included in the e-cell wall reaction.

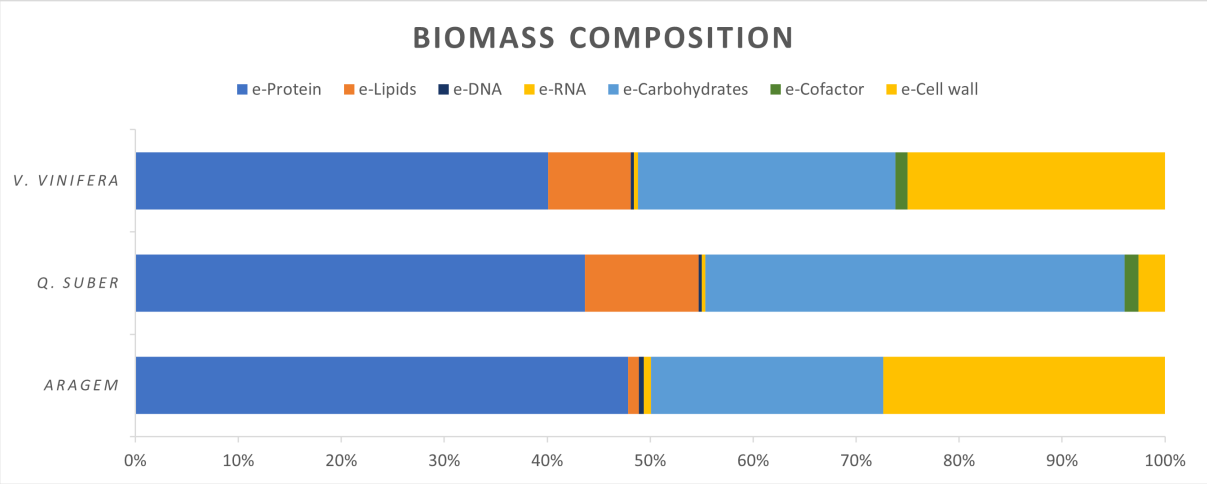


Figure 36: Biomass composition of *A. thaliana* in the AraGEM model, *Q. suber* and *V. vinifera*.

Then, reactions for producing each macromolecule were created. The amino acid and nucleotide compositions were estimated from the genome using the e-biomass tool to create the reactions that produce protein, DNA, and RNA (Tables 14, 15, and and 16, respectively).

Leucine, serine, glycine, and glutamate are the amino acids with higher stoichiometry while cysteine and tryptophan present the lowest concentrations. Glutamate and aspartate were reported to be the most abundant amino acids in plants and methionine and tryptophan were reported as the less abundant [282]. This was not verified here for aspartate and methionine. Regarding nucleotides, dTTP and dATP for DNA, and UTP and ATP for RNA were predicted to be the most abundant in *V. vinifera*.

The reactions for the production of carbohydrates, cell wall, lipids, fatty acids, and co-factors were also adapted from *A. thaliana* [139] and *Q. suber* [165] models. The carbohydrate composition is described in Table 17 and includes starch, UDP-glucose, fructose, sucrose, and the organic acids tartrate, malate, succinate, and citrate. The stoichiometry was adapted to reflect the grape composition described in the

Table 14: Protein composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-DNA)**
L-Alanyl-tRNA (Ala)	Charged-ALA-tRNAs	0.496
L-Arginyl-tRNA (Arg)	Charged-ARG-tRNAs	0.406
L-Asparaginyl-tRNA (Asn)	Charged-ASN-tRNAs	0.335
L-Aspartyl-tRNA (Asp)	Charged-ASP-tRNAs	0.404
L-Cysteinyl-tRNA (Cys)	Charged-CYS-tRNAs	0.152
L-Glutaminyl-tRNA (Gln)	Charged-GLN-tRNAs	0.288
Glutamyl-tRNA (Glu)	Charged-GLT-tRNAs	0.511
Glycyl-tRNA (Gly)	Charged-GLY-tRNAs	0.506
L-Histidyl-tRNA (His)	Charged-HIS-tRNAs	0.191
L-Isoleucyl-tRNA (Ile)	Charged-ILE-tRNAs	0.413
L-Leucyl-tRNA (Leu)	Charged-LEU-tRNAs	0.787
L-Lysyl-tRNA (Lys)	Charged-LYS-tRNAs	0.463
L-Methionyl-tRNA (Met)	Charged-MET-tRNAs	0.190
L-Phenylalanyl-tRNA (Phe)	Charged-PHE-tRNAs	0.319
L-Prolyl-tRNA (Pro)	Charged-PRO-tRNAs	0.378
L-Seryl-tRNA (Ser)	Charged-SER-tRNAs	0.694
L-Threonyl-tRNA (Thr)	Charged-THR-tRNAs	0.362
L-Tryptophanyl-tRNA (Trp)	Charged-TRP-tRNAs	0.102
L-Tyrosyl-tRNA (Tyr)	Charged-TYR-tRNAs	0.207
L-Valyl-tRNA (Val)	Charged-VAL-tRNAs	0.488

Table 15: DNA composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-RNA)**
dTTP	TTP	0.672
dGTP	DGTP	0.355
dCTP	DCTP	0.355
dATP	DATP	0.672

Table 16: RNA composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-protein)**
UTP	UTP	0.590
GTP	GTP	0.458
ATP	ATP	0.574
CTP	CTP	0.398

literature. Also, tartaric acid was added to the model as it was described as the main organic acid found in grapes, followed by malate, and it is not present in any other plant model [269]. Starch is the carbohydrate with the highest stoichiometric coefficient, followed by UDP-glucose.

The cell wall reaction (Table 18) includes cellulose and UDP-D-xylose in higher amounts and the lignin components, 4-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol, with lower coefficients.

Regarding lipids (Table 19), the reaction includes two big classes: glycosyldiacylglycerols (MGDG, DGDG, SL) and glycerophospholipids (PC, PE, PG, PA, PI), which are important components of the cellular membranes. The lipids from the first class are the major constituents of the thylakoid membrane of higher plant chloroplasts. Phosphatidylcholine (PC) is the glycerophospholipid with the highest concentration in the e-lipid composition. Although fatty acids are not directly represented in the biomass equation, they are needed for the biosynthesis of lipids. Hence, a reaction was created to produce a compound named *Fatty_Acids*, which represents the average fatty acid composition. As fatty acids are an important constituent of grape seeds, the fatty acid composition was estimated from experimental concentrations of fatty acid in this tissue (Table 20) [283]. The linoleic acid is the most abundant fatty acid, followed by oleic acid.

Finally, co-factors are metabolites needed by the enzymes to catalyse certain biochemical reactions. The co-factor equation includes universal co-factors, such as NAD+ and coenzyme A, that assist in the biosynthesis of essential metabolites. Also, the e-cofactor is composed of pigments, such as chlorophylls, vitamins, such as biotin, and antioxidants, such as glutathione (Table 21). Some known co-factors, such as thiamin diphosphate and pyridoxal phosphate were not included in the e-cofactor equation as these metabolites do not participate as co-factors in any reaction in the model.

As no information regarding the energetic requirements for *V. vinifera* was found in the literature, it was retrieved from the model of *S. lycopersicum* [160]. The growth-associated maintenance energy requirement was included in the biomass equation and defined as 53.26 $mmmol_{ATP}.gDW^{-1}$. The non-growth-associated maintenance energy requirement was defined in an ATP hydrolysis reaction with a mandatory flux of $2mmmol_{ATP}.gDW^{-1}.^{h}$, by setting the lower and upper bounds to 2.

Table 17: Carbohydrate composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-carbohydrate)**
UDP-Glucose	CPD-12575	0.212
Fructose	BETA-D-FRUCTOSE	0.160
Sucrose	SUCROSE	0.004
Starch	Starch	0.360
Tartrate	TARTRATE	0.292
Malate	MAL	0.120
Succinate	SUC	0.024
Citrate	CIT	0.016

Table 18: Cell wall composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-cellwall)**
4-coumaryl alcohol	COUMARYL-ALCOHOL	0.160
coniferyl alcohol	CONIFERYL-ALCOHOL	0.192
sinapyl alcohol	SINAPYL-ALCOHOL	0.224
UDP-D-xylose	UDP-D-XYLOSE	1.128
cellulose	CELLULOSE	1.220

Table 19: Lipid composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-lipid)**
1,2-Diacyl-3-beta-D-galactosyl-sn-glycerol (MGDG)	D-Galactosyl-12-diacyl-glycerols	0.603
Digalactosyl-diacylglycerol (DGDG)	Galactosyl-galactosyl-diacyl-glycerols	0.215
Sulfoquinovosyldiacylglycerol (SL)	SULFOQUINOVOSYLDIACYLGLYCEROL	0.064
Phosphatidylcholine (PC)	PHOSPHATIDYLCHOLINE	0.259
Phosphatidylethanolamine (PE)	L-1-PHOSPHATIDYL-ETHANOLAMINE	0.100
Phosphatidylglycerol (PG)	L-1-PHOSPHATIDYL-GLYCEROL	0.180
Phosphatidate (PA)	L-PHOSPHATIDATE	0.092
1-Phosphatidyl-D-myo-inositol (PI)	L-1-phosphatidyl-inositols	0.067

Table 20: Fatty Acid composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-fattyacid)**
Linoleic acid (C18:2)	LINOLEIC_ACID	0.55
Oleic acid (C18:1)	OLEATE-CPD	0.21
Palmitic acid (C16:0)	PALMITATE	0.13
Stearic acid (C18:0)	STEARIC_ACID	0.1
Linolenic acid (C18:3)	LINOLENIC_ACID	0.01

4.3.3 Model Curation and Validation

After including the biomass equation in the network, the model was evaluated to ensure the production of all biomass precursors. Hence, the medium was defined as containing only the compounds needed for growth. These included magnesium (II), iron (II), sulfate, nitrate, phosphate, water, oxygen, and carbon dioxide. The lower bounds of the exchange reactions of these metabolites were set to -10000. Then, depending on the metabolic process under analysis, the uptake of photons or sucrose was allowed for photosynthesis/photorespiration and respiration, respectively. Photon uptake was set to 100 $mmmol_{photon}.gDW^{-1}.h^{-1}$ and sucrose uptake was set to 1 $mmmol_{sucrose}.gDW^{-1}.h^{-1}$, according to AraGEM [139].

BioISO [271] was used to identify the biomass precursors not produced and the gaps or errors that prevent their production. This tool was run iteratively until all biomass precursors were produced. All the pathways, reactions, and enzymes directly or indirectly involved in the biosynthesis of biomass precursors were manually analysed, which was a very time-consuming and laborious task (Figure 37). The model

Table 21: Co-factor composition in *V. vinifera*.

Reactants	**Model identifier**	**Stoichiometry (mmol/g-cofactor)**
NADP+	NADP	0.00081
FAD	FAD	0.00076
coenzyme A	CO-A	0.00078
S-Adenosyl-L-methionine	S-ADENOSYLMETHIONINE	0.00151
NAD+	NAD	0.00090
riboflavin 5'-phosphate	FMN	0.00131
glutathione	GLUTATHIONE	0.00196
ascorbate	ASCORBATE	0.00341
tetrahydrofolate	THF	0.00135
10-formyltetrahydrofolate	10-FORMYL-THF	0.00127
5-methyltetrahydrofolate	5-METHYL-THF	0.00131
5,10-methenyltetrahydrofolate	5-10-METHENYL-THF	0.00131
pantetheine 4'-phosphate	PANTETHEINE-P	0.00167
heme	PROTOHEME	0.00097
biotin	BIOTIN	0.00246
chlorophyll a	CHLOROPHYLL-A	0.00067
chlorophyll b	CHLOROPHYLL-B	0.00067
ubiquinol-9	CPD-9957	0.00075
ubiquinol-10	CPD-9958	0.00069
plastoquinol-9	CPD-12829	0.00080

was evaluated in photosynthetic conditions to ensure that pathways related to photosynthesis were not blocked.

In the first iteration, no biomass precursor was produced. Although the uptake of photons was allowed, the model was unable to use them to produce energy because two reactions from the photosynthesis pathway were missing from the model. In this case, the database mapping between the enzymes and the reactions they catalyse was incorrect or incomplete. This means that, although the enzyme or the monomers of protein complexes were identified in the genome, the reaction was not added to the model. This problem was fixed in the database whenever possible.

After adding these two reactions, the e-DNA and e-RNA production was analysed. The model was unable to produce nucleotides, inhibiting the majority of biosynthetic pathways. Hence, the pathways producing deoxyribonucleotides and ribonucleotides were analysed to identify the missing reactions or metabolites that were preventing the biosynthesis of these compounds and to find out why these reactions were not included in the model. In most cases, the problem was the same as mentioned above.

Then, after having e-DNA and e-RNA, the production of e-Protein was analysed and the tRNA molecule charged with proline was identified as the only one not being produced, which was due to the fact that the proline-tRNA ligase reaction was not in the model. In this case, the reaction had not been added to the model because there was no match between the enzyme associated with the reaction in the database and the proteins in the *V. vinifera* genome. Hence, the sequence of a new enzyme catalysing this reaction

was obtained from UniProt and DIAMOND was run again with this enzyme against the *V. vinifera* genome, which resulted in four matches. Thus, the reaction and the respective GPR rules were added to the model.

The next step was to verify the e-Carbohydrate reaction, where tartrate was missing. The reactions for tartrate production were not in the model, as no enzyme encoding these reactions has yet been identified. Hence, two reactions were added to the model without associated GPR rules.

The reactions involved in the biosynthesis of starch and cellulose had previously been altered when analysing the mass balance of the reactions. The generic compounds in these reactions were replaced or adapted to allow the mass balancing and production of these two biomass precursors. For instance, the starch synthase reaction had the same generic metabolite, 1,4-alpha-D-glucan, on both sides but could not be removed from the model as it was needed to produce starch. Hence, 1,4-alpha-D-glucan was replaced in the reactants by a molecule of ADP-glucose, as this allowed the mass balancing of the reaction. Similarly, the reactions producing cellulose also had cellulose on both sides. Hence, the cellulose in the reactants was replaced by three molecules of UDP-glucose or GDP-glucose in the reactions cellulose synthase UDP- and GDP-forming, respectively.

Sinapyl alcohol production was blocked, preventing the biosynthesis of the cell wall. The steps taken to fix this are represented in Figure 37. In this case, a reaction for the production of a sinapyl alcohol precursor was missing and was added to the model (RXN-1143 - caffeate O-methyltransferase).

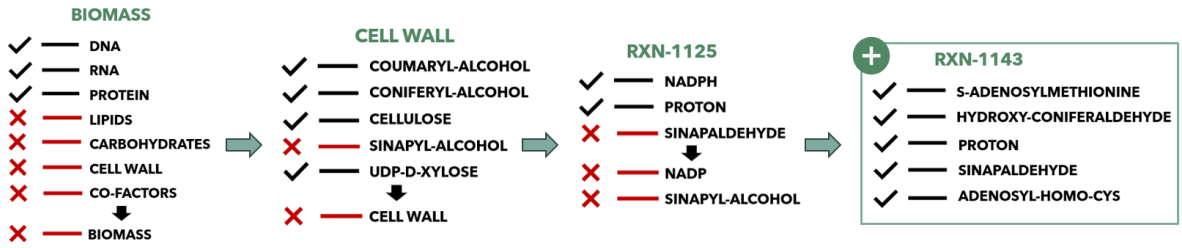


Figure 37: An example of the steps taken during manual curation to fix blocked biomass. In this case, cell wall production was analysed and sinapyl alcohol, a cell wall component, was missing, despite having a reaction to produce it. This reaction (RNX-1125) was analysed and its reactant sinapaldehyde was missing and there was no reaction in the model for its production. Hence, the gap was identified and filled by adding the reaction RXN-1143 to the model. The production of all reactants of this reaction was verified and the GPR rules were manually defined.

Regarding lipids, none of the precursors were being produced. Most fatty acids were not being produced either, which prevented lipid production. This reflects the poor annotation of enzymes related to lipid metabolism. Hence, reactions were manually added to allow the production of oleic acid, palmitic acid, stearic acid, and linolenic acid. In fact, the reaction for producing the latter fatty acid was retrieved from KEGG as no viable reaction was available in Plantcyc or Metacyc for producing this metabolite. GPR rules were also manually added by running DIAMOND again but with new enzyme sequences whenever needed. Then, the production of diacylglycerol was manually analysed as this metabolite was involved in the biosynthesis of most lipids. After assuring the production of fatty acids and diacylglycerol, phosphatidylethanolamine, and phosphatidylglycerol were still not being produced as their producing reactions were missing from the model. These reactions as well as the respective GPR rules were added to the

model.

Finally, in the first analysis of the e-cofactor reaction, most cofactors were missing from the model. Most reactions involved in the biosynthesis of ubiquinol-10 and chlorophyll a and b were manually added to the model as no enzyme matches were obtained in the annotation step. Similarly, reactions involved in the biosynthesis of glutathione, biotin and folate transformations were also missing and were added to the model. Hence, each missing metabolite was analysed recursively until finding the base gap responsible for its blocking.

When biomass was finally produced by the model, the remaining blocked reactions were analysed to identify gaps in important pathways that should be filled in. The pathways with more gaps were those related to lipid metabolism, including the biosynthesis of sterols, such as phytosterol, brassinosteroid, and cholesterol, the biosynthesis and oxidation of fatty acids, and the biosynthesis of other plant hormones, such as jasmonic acid, and gibberellins. These pathways were poorly annotated and a manual analysis allowed adding missing reactions to the model and finding the respective GPR rules.

Then, as the annotation of secondary pathways is also very incomplete in the database, these were extensively analysed. Some gaps were detected in the early steps of the phenylpropanoid biosynthesis pathway as well as in the pathways for flavonol, flavone, flavonoid, and isoflavonoid biosynthesis. These gaps were filled even when no GPR rules were found for the reactions, as the metabolites produced in these pathways are the precursors for the production of important secondary metabolites of *V. vinifera*.

A list of the most common secondary metabolites in *V. vinifera* was created based on the literature and available metabolomics data [269, 274, 284–286]. The reactions for producing these metabolites were included in the model whenever possible, even when there were no known enzymes to catalyse the reaction or matches in the *V. vinifera* genome. The reactions added to the model are listed in Supplementary File 5. For most of them, enzyme associations were found and the GPR rules were defined. Different classes of secondary metabolites were considered, including flavonols, hydroxybenzoic acids, flavan-3-ols, anthocyanins, monoterpenes, flavones, and stilbenes. Some important secondary metabolites were already in the model, such as resveratrol, cyanidins, and quercetin (Supplementary File 5). Many others were not added to the model as no reactions for their production were available in the databases, including procyanidins, resveratrol derivatives, like viniferins, and some anthocyanin glycosides.

At the end of the manual curation process, some reactions were still blocked as there was not enough information or evidence to fill the gaps that were causing the problems. These reactions were kept in the model as, in the future, new information may emerge that would allow filling these gaps and activate the reactions.

In total, 2339 reactions in the model are blocked (Supplementary File 6). This large number can be explained in part by the lack of transport reactions between cellular compartments, which prevents the metabolites from being secreted by the cell through the exchange reactions, inhibiting their production. An example is the phenylalanine—tRNA ligase reaction, whose enzymes were predicted to be in the cytosol, chloroplast, and mitochondria. However, no reactions were identified to mediate the transport of the charged tRNA from the chloroplast or mitochondria to the cytosol. Hence, these two reactions are blocked. In addition, many reactions involved in the biosynthesis of lipids, lipid-derived metabolites, such as sterols,

and some secondary metabolites are blocked due to the lack of information available to better annotate these pathways.

In short, after the manual curation of the model, many reactions were manually added to the model, others were removed, and some were altered to be mass-balanced and not include generic metabolites, ensuring biomass production. The replaced class metabolites are described in Supplementary File 2, all reactions removed from the initial annotation are listed in Supplementary File 3, the reactions added are listed in Supplementary File 7, and reactions made irreversible are listed in Supplementary File 8.

4.3.4 Model Properties

The reconstructed generic model of *V. vinifera* named iMS7199 includes 5399 reactions (1624 transporters and 244 exchanges), and 5141 metabolites, across eight compartments: cytosol, chloroplast, mitochondria, endoplasmic reticulum, peroxisome, Golgi apparatus, vacuole, and extracellular space. In this model, the GPR rules are defined using the genome protein identifiers instead of genes as genome annotation was performed using protein sequences. As genes can encode more than one protein, the model includes 7199 protein identifiers that represent the 6018 genes in the *V. vinifera* genome. This model is mass-balanced and is able to simulate growth in phototrophic and heterotrophic conditions, by setting the photon and carbon dioxide or sucrose as the only energy or carbon source, respectively. In addition, it requires the uptake of nitrate, phosphate, sulfate, iron, magnesium, and water to produce biomass.

The statistics of the *V. vinifera* model as well as other relevant plant models are presented in Table 22. Analysing the table, only the *Q. suber* [165] model has more genes, reactions, and metabolites than the *V. vinifera* model. The other models are much smaller, even the *G. max* model [163], which has a large number of genes.

Table 22: Statistics of the *V. vinifera* model and other eight plant GSMMs

	Reactions	Metabolites	Genes	Compartments
A. thaliana (Cheung et al., 2013)	2769	2739	2857	5
Z. mays (Bogart et al., 2016)	1268	1121	2140	8
O. sativa (Chatterjee et al., 2017)	1136	1330	3602	4
S. lycopersicum (Yuan et al., 2015)	2143	1998	3410	5
M. truncatula (Pfau et al., 2018)	2909	2780	3403	8
G. max (Moreira et al., 2019)	3001	2814	6127	5
S. viridis (Shaw et al. 2019)	2473	2429	3376	5
Q. suber (Cunha et al., 2021)	6230	6481	7871	8
V. vinifera (iMS7199)	5399	5141	7199	8

The reactions of *V. vinifera* were compared with those from the other models, except for *Q. suber*, which has different model identifiers. Drains, transporters, biomass pseudo-reactions, and compartments were not considered, resulting in 2769 reactions of the iMS7199 model (Figure 38).

The *G. max* model shares the highest number of reactions with the *V. vinifera* model, corresponding to around 41% of all reactions from both models. This model is followed by the ones of *S. viridis* and *A.*

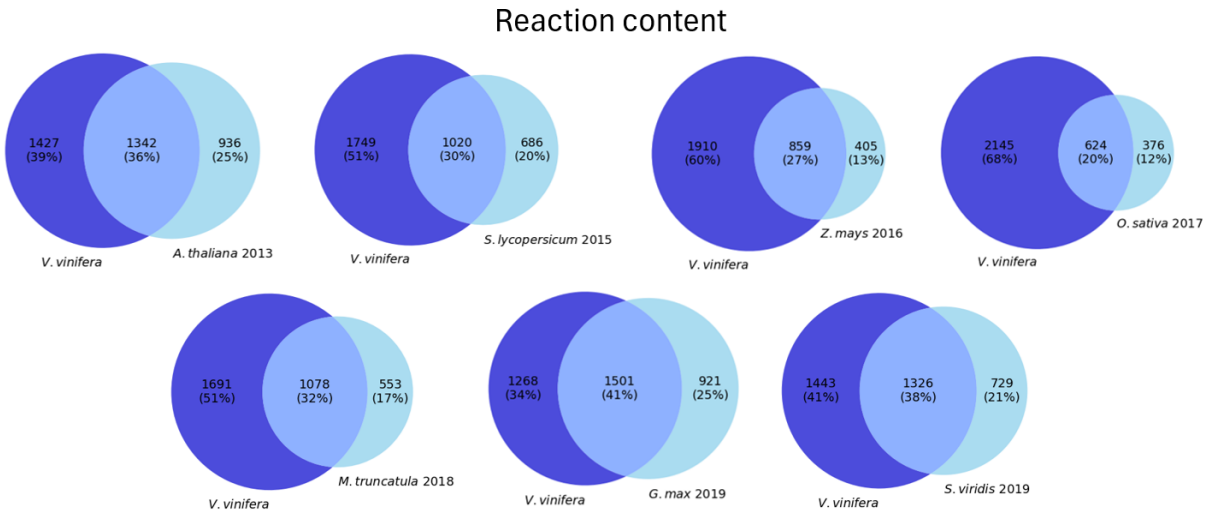


Figure 38: Venn diagrams comparing the reaction content of *V. vinifera* model with other seven plant models.

thaliana, which share 1324 (38%) and 1339 (36%) reactions with iMS7199, respectively. The most distant model is the one from *O. sativa*, sharing only 624 reactions (20%).

In total, *V. vinifera* has 785 reactions that are not present in any other model. These reactions were analysed to identify the associated pathways and gene annotation. The pathways with more unique reactions are presented in Figure 39. Reactions without pathway associations were not considered.

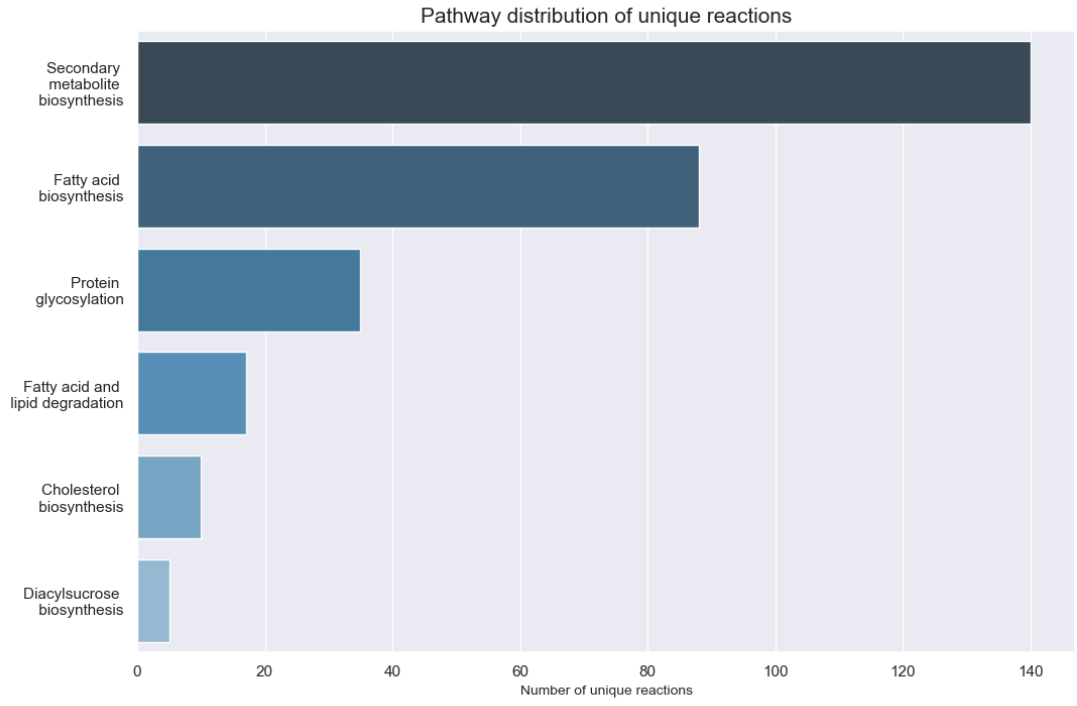


Figure 39: Pathway distribution of the reactions included in the *V. vinifera* model and not in the other plant models analysed.

As shown in Figure 39, the biosynthesis of secondary metabolites is the pathway class associated with more unique reactions, around 140, followed by fatty acid biosynthesis (88 reactions), protein glycosylation (35 reactions), and fatty acid and lipid degradation (17 reactions). These pathway classes comprise several specific pathways. Other specific pathways with more than four unique reactions include cholesterol biosynthesis and diacylsucrose biosynthesis. Hence, the *V. vinifera* model represents a great advance compared to the previous plant models, as it comprises new reactions, especially for the secondary metabolism, which is often underrepresented in plant models. On the other hand, the other plant models include 661 secondary reactions not available in iMS7199. This number can be explained by the fact that 380 reactions are not associated with GPR rules in the models. Moreover, 93 are sub-reactions of others that are included in the *V. vinifera* model. Of the 188 reactions that have GPR rules, 155 have a corresponding enzyme sequence in the iplants database, which had a match in the DIAMOND annotation but were not the first hit for any query protein (Supplementary File 9). This data is made available and may be used in the future to improve the model by performing further manual curation.

Regarding genome annotation, 27% of the proteins that catalyse these unique reactions were annotated based on the genome of *A. thaliana*. These proteins or reactions were probably not in the metabolic databases when the *A. thaliana* models were reconstructed, which can explain why they are missing from these models. Besides *A. thaliana*, 12% of the unique proteins matched human proteins, and around 27% were annotated based on proteins from more than 100 different plant species, including *Solanum lycopersicum*, *Solanum tuberosum*, *Catharanthus roseus*, *Petunia x hybrida*, *Medicago truncatula*, *Glycine max*, and *V. vinifera*.

As *A. thaliana* is a reference organism for plants, much enzymatic and metabolic information of *A. thaliana* is available in databases, like PlantCyc and UniProt. On the other hand, data for more complex plants is scarce. Therefore, it was expected that a large percentage of *V. vinifera* proteins would be annotated based on homologous proteins from *A. thaliana*. However, as *V. vinifera* is a much more complex plant, gene annotations may be wrong or missing. In addition, several proteins are similar to human proteins, which was also expected as various pathways, mainly related to lipid metabolism, are more characterised in humans than in plants.

4.3.4.1 Secondary Metabolic Pathways

Secondary metabolites are economically very important as they have many relevant applications. However, the pathways that produce them are very complex and diverse, and the knowledge in this subject is still very limited [44]. The production of the main secondary metabolites in the model, terpenoids and phenylpropanoids , is schematised in the Figures 4 and 3, respectively, of the State of the Art chapter.

Grapes are known to have a high content of phenolic compounds and different grape varieties usually have different phenolic compositions, which leads to different wine flavours and aromas [287]. The reconstructed *V. vinifera* model contains complete pathways for the biosynthesis of several terpenoids and phenylpropanoids, which include flavonoids, such as quercetin, myricetin, kaempferol (and derivatives), and anthocyanins, like cyanidin and delphinidin. Anthocyanins usually accumulate during grape

maturation and are responsible for the grape colour in red grapevine varieties, being absent in white varieties [288]. Another important group of phenylpropanoids in grapes are stilbenes, such as resveratrol, which protects grapes from intense UV light. As resveratrol is an antioxidant agent, it has high economic importance, being used in the pharmaceutical and cosmetic industry [261]. The complete pathway of resveratrol biosynthesis is described and included in the model, but there are many gaps in the biosynthetic pathways of resveratrol derivatives, such as viniferins, which are not included in the model but are important for wine flavour and aroma.

This model also contains complete pathways for the biosynthesis of other aroma compounds, such as linalool, 1,3,5-trimethoxybenzene (TMB), and 3,5-dimethoxytoluene, the latter two being only described for *Rosa chinensis*.

In addition, complete secondary pathways for the biosynthesis of plant hormones, such as jasmonates, cytokinins, gibberellins, ethylene, and auxins, are available in the model. For instance, jasmonates are known to regulate seed germination, and flower and fruit development, as well as to defend the plants against some pathogens [289]. Cytokinins usually control cell growth and differentiation [290]. Although the reconstructed GSMM does not represent the action of these hormones, it can show the metabolic potential of the network to produce these hormones.

In short, the reconstructed model of *V. vinifera* contains several pathways representing secondary metabolism. However, further curation is necessary to fill the existing gaps and increase the number of secondary metabolites in the model.

4.3.5 Phenotype Predictions

The iMS7199 model was simulated using pFBA under different conditions: photosynthesis, photorespiration, and respiration. The first simulation strategy was used, which consisted of fixing the biomass growth rate to $0.11 h^{-1}$ and minimising the photon/sucrose uptake as the objective function.

Table 23 shows the summary of the phenotype predictions, proving that the model is capable of predicting metabolic behaviour under both phototrophic and heterotrophic conditions.

In photosynthesis and photorespiration, the leaf uptakes light, carbon dioxide, water, nitrate, and sulfate to produce biomass, and releases oxygen and phosphate, as expected. Iron II and magnesium (Mg) are also captured but with very low fluxes (less than 1^{-5} mmmol.gDW^{-1}.h^{-1}). Light uptake is significantly higher in photorespiration than in photosynthesis (41.40 mmmol.gDW^{-1}.h^{-1} vs 29.73 mmmol.gDW^{-1}.h^{-1}), which was also expected as the latter process is known to be more efficient for energy production. In both cases, the pathways of the central metabolism are the most active: photosynthesis, Calvin cycle, glycolysis, starch and amino acid biosynthesis, and oxidative phosphorylation. Also, the photorespiration reactions are only active under photorespiration conditions.

During both processes, TCA cycle is incomplete: citrate is converted to isocitrate and this is converted to 2-oxoglutarate, which is used for the biosynthesis of glutamate and glutamine instead of being used to produce succinate. Fumarate is produced from arginine biosynthesis, instead of being produced from succinate, and enters the cycle. This result is consistent with the results observed from other plant models

under light conditions [148, 165] and with isotope labelling experiments, which stated that a cyclic TCA only happens when the demand for ATP is high. The photosynthetic ATP production reduces that demand [291, 292].

In respiration, the leaf uptakes sucrose, nitrate, sulfate, and oxygen, and releases hydrogencarbonate, water, and phosphate. The main active pathways include glycolysis, complete TCA cycle, starch and amino acid biosynthesis, and oxidative phosphorylation, as expected.

Table 23: Summary of the phenotype predictions for photosynthesis, photorespiration, and respiration. This table shows the metabolites that are consumed and produced by the models. The fluxes of the metabolites are in mmol.gDW^{-1}.h^{-1} while for biomass are in h^{-1}.

	photosynthesis	**photorespiration**	**respiration**
	uptake		
SUCROSE	-	-	0.55
Light	29.73	41.4	-
NITRATE	0.37	0.37	0.37
OXYGEN-MOLECULE	-	-	1.79
PROTON	5.89	5.89	3.69
SULFATE	0.02	0.02	0.02
WATER	2.84	2.84	-
CARBON-DIOXIDE	4.41	4.41	-
	production		
OXYGEN-MOLECULE	4.79	4.79	-
WATER	-	-	0.98
HCO3	-	-	2.16
PPI	0.05	0.05	-
Pi	-	-	0.09
e-Biomass	0.11	0.11	0.11

4.4 Tissue-specific Models

4.4.1 RNA-Seq Data

As no study was available with RNA-Seq data for the three tissues of interest, leaf, stem, and berry, there was a need to use data from different studies. As mentioned above, the chosen datasets were retrieved from the GREAT database, where leaf and stem data belonged to the same study [273], and berry data was from another study [274], both from *V. vinifera* Cabernet Sauvignon cultivar. The first study was chosen because it includes data from two different tissues, and the last was chosen as it includes both gene expression and metabolomics data for a large number of grape samples under different developmental stages.

As samples from different studies could have intrinsic noise created by the sequencing platforms or experimental protocols used, an additional analysis was performed to validate if these samples could be

compared and give meaningful insights regarding the metabolic differences between tissues without external bias. Hence, for a more comprehensive analysis, more control samples were retrieved from GREAT for leaf and berry tissues of the Cabernet Sauvignon cultivar (GEO accessions: GSE97960, GSE76256, GSE89113, and GSE74428). The data collected is characterised in Table 24. In total, six studies were used comprising the expression values of 35336 *V. vinifera* genes and 217 samples from three different sequencing platforms and three countries. The gene expression values were normalised as TPM and log-transformed (log2). Most samples are from berries (around 77%) from the study of Fasoli et al. [274], which used the Illumina HiSeq 1000 as the sequencing platform.

Table 24: RNA-Seq studies retrieved from GREAT database to check if their samples are comparable. The studies in bold are the ones selected for omics data integration with the GSMM.

Study (GEO Accession)	Tissue	Country	Sequencing technology	Year	Number of samples
GSE76256	Grape skin	China	Illumina HiSeq 2500	2015	12
GSE97960	Grape pericarp	Italy	Illumina HiSeq 2000	2011	36
GSE98923	**Green and mature grape**	**California**	**Illumina HiSeq 1000**	**2012, 2013, 2014**	**120**
GSE97900	**Stem and leaf**	**California**	**Illumina HiSeq 2500**	**2017**	**42**
GSE89113	Leaf	China	Illumina HiSeq 2000	2015	3
GSE74428	Leaf	China	Illumina HiSeq 2000	2015	4

t-SNE was used to comprehensively visualise the data and verify if the data distribution is influenced by the study or sequencing platform (Figure 40). Analysing the t-SNE projections, the samples of the same tissue grouped well together, which was not verified for the samples of the same study or platform. These results indicate that data normalisation performed by the GREAT authors seems to be effective in removing intrinsic sample noise, allowing meaningful comparisons of samples from different studies. Therefore, a dataset with samples from the two chosen studies was created and filtered to include only the genes in the *V. vinifera* model.

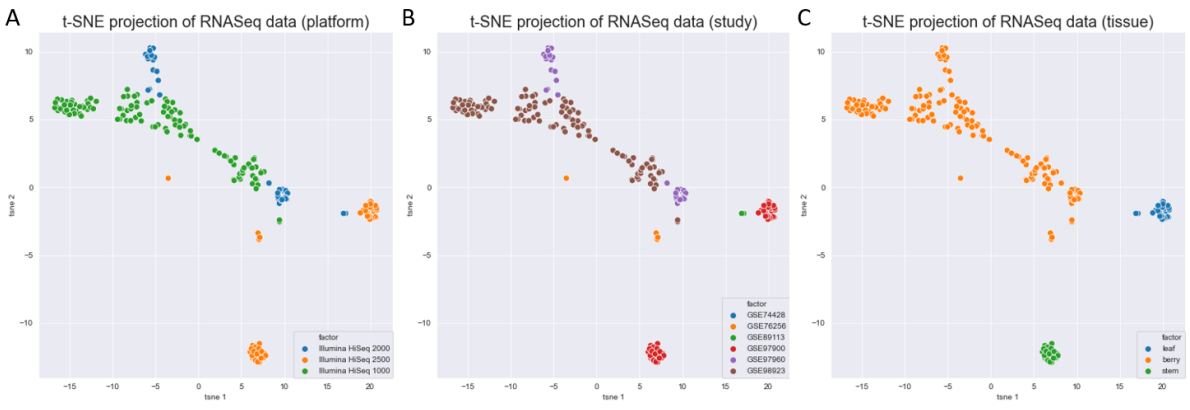


Figure 40: t-SNE projections of the RNA-Seq data coloured by platform (A), study (B), and tissue (C). Data was retrieved from the GREAT database, including the TPMs of 35336 *V. vinifera* genes across 217 samples from six different studies.

In total, the final dataset contained the expression of 6018 genes across 162 samples. The time-point metadata for these samples was grouped into two developmental stages, green and mature, to create a metabolic model for each state. The sample distribution and the t-SNE projections for the final RNA-Seq dataset are shown in Figure 41.

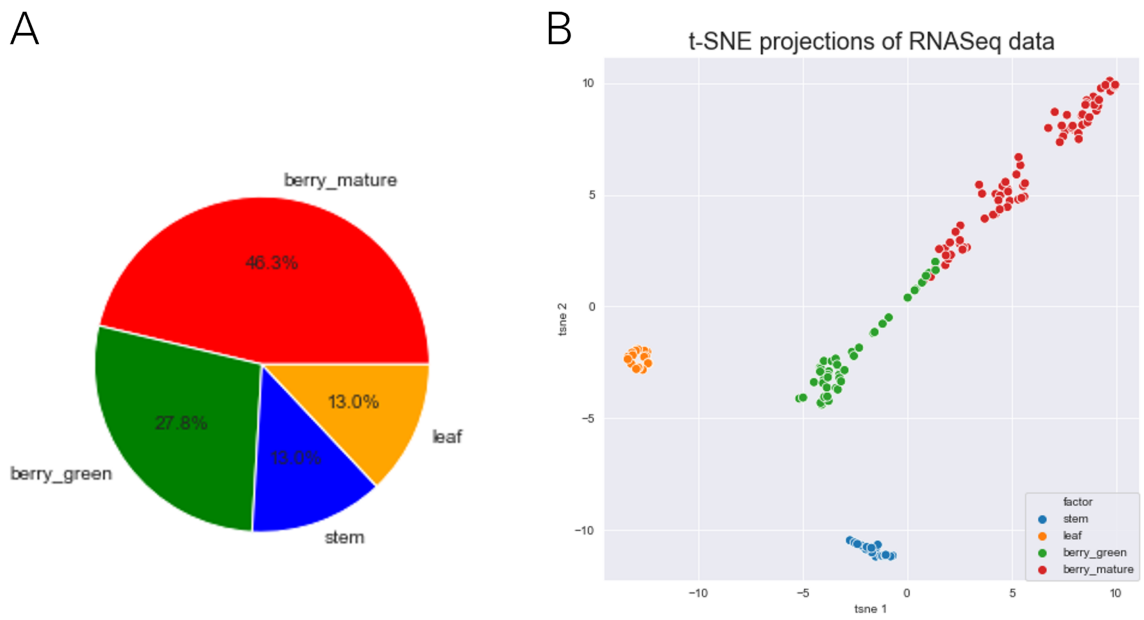


Figure 41: RNA-Seq data for all tissues: leaf, stem, and berry in the green and mature states. A. Distribution of samples across the different tissues. B. t-SNE visualisation of the RNA-Seq data for all tissues: leaf (orange dots), stem (blue dots), and berry in green (green dots) and mature (red dots) states. Data was retrieved from the GREAT database, including the TPM of the *V. vinifera* genes in the model across 162 samples.

Mature berry is the most represented tissue in the dataset with 46% of the samples (75 samples), 28% of the samples are from green berries (45 samples), while stem and leaf represent 13% of the samples each (21 samples). Analysing the t-SNE plot, the data grouped well by tissue: leaf samples are the most well-grouped, followed by stem samples, while berry samples are more scattered. Green berries partially overlap mature berry samples and some separation exists between mature berry samples. The overlap was expected as these samples were retrieved every week from fruit set to maturity. Thus, samples a week before and a week after veraison, which is when the maturation phase starts, may be similar in metabolism.

4.4.2 Biomass Composition

Before integrating the omics dataset with the GSMM of *V. vinifera*, the definition of tissue-specific biomass compositions is crucial for obtaining good models capable of simulating the specific metabolism of each tissue. Ideally, experimental data should be used to define biomass composition for different tissues. However, as these data are not available for *V. vinifera*, the biomass content was estimated based on other plant models and insights from the literature [269, 277].

The biomass compositions of each tissue are represented in Figure 42. The biomass formulation for leaf and green berry was considered to be equal to the generic biomass previously defined, which was used as a reference to define the biomass composition of the other tissues. According to other plant models, like *Q. suber* [165] and *A. thaliana* [149], the stem is expected to have a higher cell wall and carbohydrate content and lower protein and lipid levels. The biomass of mature berry was adjusted according to the literature and the metabolomics data available [274], which states that mature berries have higher sugar and amino acid content than the other tissues.

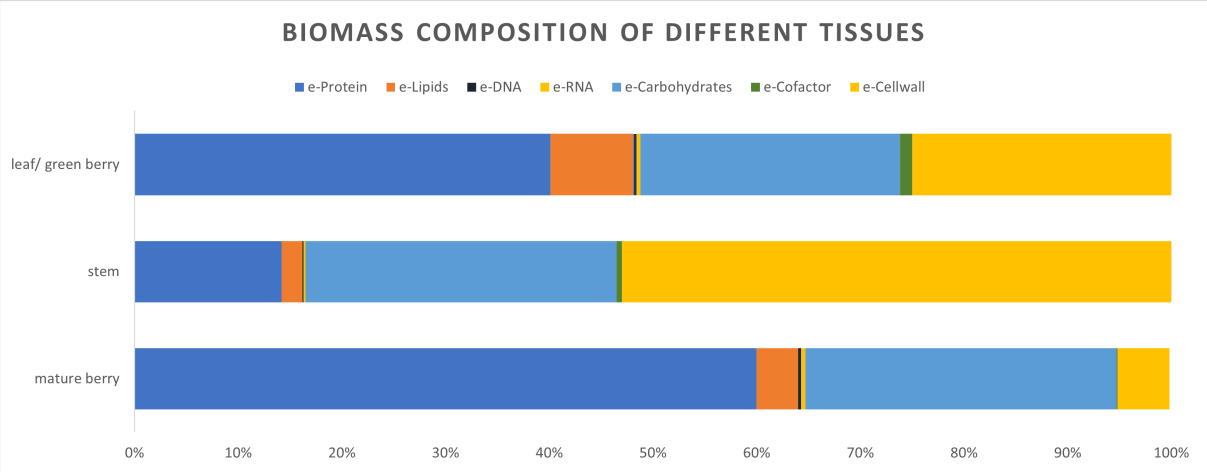


Figure 42: Biomass composition of the leaf and green berry, stem, and mature berry. These values were adapted from available plant models and the literature.

The reactions for producing each macromolecule in the tissue-specific models were considered to be the same as those in the generic model and across tissue models, except for the e-Cofactor reaction, as chlorophylls and plastoquinol were only included in the e-Cofactor reaction of the leaf, and the e-Carbohydrate reaction for mature berry, in which the contents of sugars and organic acids were adjusted to reflect the changes in berry composition during maturation, and some secondary metabolites were added based on metabolomics data [274]. These metabolites are found in mature grapes and mainly include anthocyanins, such as malvidin 3-glucoside, peonidin 3-O-glucoside, and petunidin-3-O-glucoside.

In short, the tissue-specific biomass compositions reflect that leaf and green berries present high levels of carbohydrates and proteins, the stem is mainly composed of carbohydrates and cell wall precursors, and mature berries present high amounts of sugars and proteins but fewer amounts of organic acids.

4.4.3 Omics Integration

The created omics dataset was integrated with the reconstructed generic model of *V. vinifera* to create specific models for different tissues: stem, leaf, green berry, and mature berry. Omics integration was performed using the FASTCORE algorithm and different thresholding strategies. In total, nine models of each tissue were created, and the number of genes, reactions, and unique reactions of the models excluding drains is summarised in Table 25. Three models were created using the default strategy and

different thresholds for the gene expression: 6, 8, and 10. The higher the threshold, the lower the number of reactions and genes included in the models. Thresholds of 8 and 10 resulted in very small models, with less than 1100 reactions, which were not considered for the following analysis. Global models were created using the 50 (global_50) and 75 (global_75) percentiles as global thresholds for pre-filtering the RNA-Seq data. This resulted in two sets of identical models, which included more reactions than the default models but fewer unique reactions across tissues. The local T1 strategy was used with the percentile 50 as the global threshold and both 50 (local T1_50_50) and 75 (local T1_50_75) percentiles as the local threshold, resulting in two different sets of models, the last of which included slightly smaller models but with more unique reactions across tissues. Finally, the local T2 strategy was applied with the 25 and 75 percentiles as global thresholds, and 50 (local T2_25_75_50) and 75 (local T2_25_75_75) as local thresholds, resulting in two equal sets of models that were bigger than local T1 models.

Table 25: Tissue-specific models created with the FASTCORE algorithm implemented in Troppo and different thresholding strategies. Default models were created using three different integration thresholds: 6 (default_t6), 8 (default_t8), and 10 (default_t10). Global models were created using the 50 (global_50) and 75 (global_75) percentiles as global thresholds, which resulted in two equal sets of models. Similarly, the local T2 strategy with the 25 and 75 percentiles as global thresholds, and 50 (local T2_25_75_50) and 75 (local T2_25_75_75) as local thresholds resulted in the same sets of models. Finally, local T1 was used using the percentile 50 as the global threshold and both 50 (local T1_50_50) and 75 (local T1_50_75) percentiles as the local threshold, resulting in two different sets of models.

Models		**Leaf**	**Stem**	**Green berry**	**Mature berry**
default_t6	reactions	2302	2211	2468	2144
	unique reactions	392	169	169	140
	genes	4670	4787	4942	4446
default_t8	reactions	1084	984	1179	1234
	unique reactions	210	95	88	144
	genes	3008	2967	3035	3127
default_t10	reactions	722	590	642	666
	unique reactions	161	15	25	186
	genes	2105	1773	2089	2140
global_50 global_75	reactions	4463	4436	4402	4253
	unique reactions	59	89	8	18
	genes	6889	6869	6861	6640
local T1_50_50	reactions	3603	3420	4094	3289
	unique reactions	60	94	12	19
	genes	6309	6324	6562	5882
local T1_50_75	reactions	3148	2693	3003	2604
	unique reactions	469	314	413	273
	genes	5771	5636	5401	5151
local T2_25_75_50 local T2_25_75_75	reactions	4266	4140	4251	4028
	unique reactions	124	97	26	19
	genes	6701	6602	6657	6312

As the choice for the best tissue-specific models is not straightforward, flux sampling data for all sets of models were generated and t-SNEs were used to visualise this data and check which strategy better distinguishes the different tissues (Figure 43). Overall, all strategies resulted in models whose fluxes are

well separated across tissues in the t-SNE projections. Regarding the models created with the local T1 strategies, the flux samples of the same tissue appear to be better grouped. The projections for default, global, and local T2 strategies create additional sample groups by separating fluxes of the same tissues, mainly of the mature berry and stem. Therefore, the models created with local T1 strategies seem to be the ones that best separate the fluxes by tissue. Although this approach does not provide an objective assessment of which set of tissue models is the best, it seems to be the best way to visualise this data.

The local T1_50_75 model contents were analysed and some important reactions from the core metabolism were removed from the tissue models, such as the transketolase reaction from the pentose phosphate pathway and glyceraldehyde-3-phosphate dehydrogenase and 6-phosphofructokinase from glycolysis. Hence, the models created by the local T2_25_75_50 strategy were chosen for the following analysis as these include all important reactions and this strategy was described by Lewis et. al [278] as the one creating more realistic models.

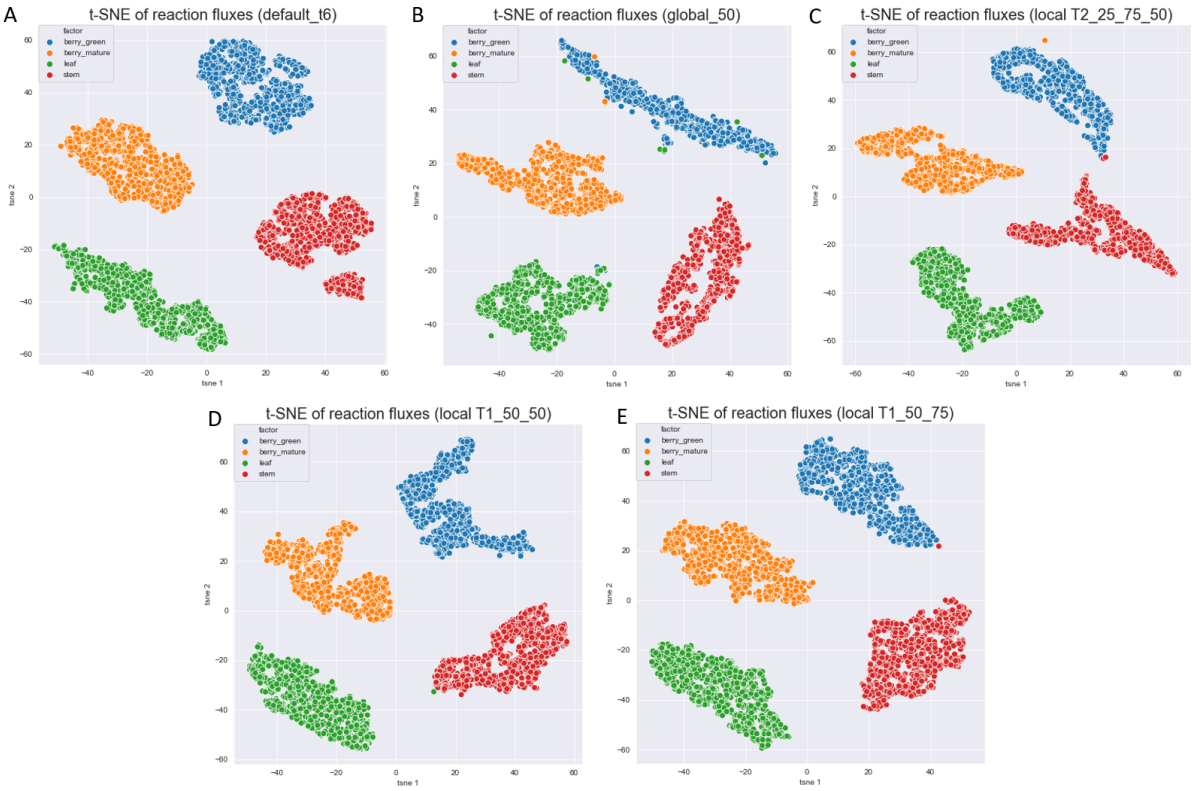


Figure 43: t-SNE projections of the flux sampling data generated by the tissue-specific models created with different strategies: default_t6 (A), global_50 (B), local T2_25_75_50 (C), local T1_50_50 (D), and local T1_50_75 (E). The flux sampling data was obtained using the ACHR sampler with a thinning factor of 100 and a sample size of 3000. From these, 1000 random samples were selected to facilitate data visualisation. Reactions with a flux of 0 across all samples were removed from the dataset and data was scaled to z-scores. The resulting data was visualised with t-SNE using a perplexity of 50. Overall, the flux data grouped well by tissue.

4.4.4 Final Models

According to the previous analysis, the tissue-specific models chosen were the ones created with the local T2 strategy. The statistics of these models are shown in Table 26. All models have the same 244 exchange reactions.

Table 26: Statistics of the generic and tissue-specific GSMMs.

	Generic model	**Leaf**	**Stem**	**Green Berry**	**Mature berry**
Genes	7199	6701	6602	6657	6312
Metabolites	5141	4456	4310	4399	4181
Reactions	5399	4510	4384	4495	4272
Transport	1624	1295	1315	1324	1305
Unique reactions	-	124	97	26	19
Metabolic Reactions	3531	2971	2825	2927	2723
- Cytosol	1434	1154	1092	1113	1059
- Chloroplast	793	745	701	725	684
- Mitochondria	335	313	318	320	314
- Endoplasmic reticulum	568	410	370	417	353
- Peroxisome	165	158	152	153	151
- Vacuole	52	49	47	48	32
- Golgi apparatus	54	41	41	50	50
- Extracellular	130	101	104	101	80

The number of reactions is similar across all tissue-specific models. Even so, the mature berry model is the smallest one, while the leaf model is the largest, having a higher number of reactions in the chloroplast, as well as more unique reactions, as expected. The list of pathways in each tissue model, as well as the number of reactions belonging to each pathway, are available in Supplementary File 10. No significant differences were found between models at the pathway level. Most pathways that differ between tissues only have one or two reactions. For instance, 2,4,6-trinitrotoluene degradation and 2-methylketone biosynthesis pathways are only present in the leaf, and the taxol and oleanolate biosynthesis pathways were only found in the stem. No unique pathways were found in green and mature berries.

The tissue-specific models were simulated using pFBA and the first strategy described before of fixing the biomass growth rate to $0.11h^{-1}$ and minimising the photon/sucrose uptake as the objective function. Additional drain reactions were added to the models to allow for the production of all metabolites in the network as well as essential transporters without associated genes.

A summary of the phenotype predictions is presented in Table 27 and full results are available in Supplementary File 11. The leaf tissue was simulated for all processes: photosynthesis, photorespiration, and respiration, while the other tissues were simulated for respiration only. Photosynthesis can also occur in green berries, but not at significant levels [293]. Hence, in this analysis, the leaf was considered to be the only photosynthetic tissue.

In photosynthesis and photorespiration, the leaf uptakes the same nutrients described before for the

Table 27: Summary of the phenotype predictions for the tissue-specific models of leaf, stem, green berry, and mature berry for photosynthesis, photorespiration (leaf only), and respiration, minimising the uptake of photons or sucrose and fixing biomass rate at $0.11 h^{-1}$. This table shows the metabolites that are consumed and produced by the models. The fluxes of the metabolites are in $mmol.gDW^{-1}.h^{-1}$ while for biomass are in h^{-1}.

	photosynthesis	**photorespiration**	**respiration**			
metabolite	**leaf**			**stem**	**berry green**	**berry mature**
Uptake						
SUCROSE			0.61	0.49	0.61	0.72
Light	32.09	43.76				
CARBON-DIOXIDE	4.41	4.41				
NITRATE	0.37	0.37	0.37		0.37	
OXYGEN-MOLECULE			1.74	1.71	1.74	1.84
PROTON	6.64	6.64	3.73	4.01	3.73	3.71
SULFATE	0.02	0.02	0.02	0.01	0.02	0.02
WATER	3.21	3.21		0.11		
Production						
OXYGEN-MOLECULE	5.54	5.54				
NITRATE						0.70
AMMONIUM				0.04		
HCO3			2.86	1.98	2.86	3.47
PPI	0.05	0.05				
Pi			0.09	0.21	0.09	
WATER			0.55		0.55	1.25
e-Biomass	0.11	0.11	0.11	0.11	0.11	0.11

generic model, but with slightly different values. Light uptake is again significantly higher in photorespiration than in photosynthesis (43.76 $mmmol_{photon}.gDW^{-1}.h^{-1}$ vs 32.09 $mmmol_{photon}.gDW^{-1}.h^{-1}$), and the same central pathways are the most active.

The respiration results were similar to those obtained for the generic model and across tissues. The stem tissue uptakes ammonium and water, and the mature berry has a slightly higher demand for sucrose to produce the same biomass flux.

In summary, the integration of omics data into the generic GSMM created tissue-specific models that try to reflect the differences in gene expression between tissues. However, the number of reactions and the phenotype predictions are not very different between models; thus, a complementary analysis based on differential flux predictions was performed to understand the metabolic differences between the tissues.

4.4.5 Differential Flux Analysis

The ACHR sampler was used to generate 10000 sample fluxes for all reactions from the different tissue-specific models to identify the ones with differential fluxes between models. In total, 764 reactions were found to have altered fluxes between at least two models. The complete flux dataset was then scaled and filtered to include only these reactions. The resulting data was plotted in a t-SNE, shown in Figure 44.

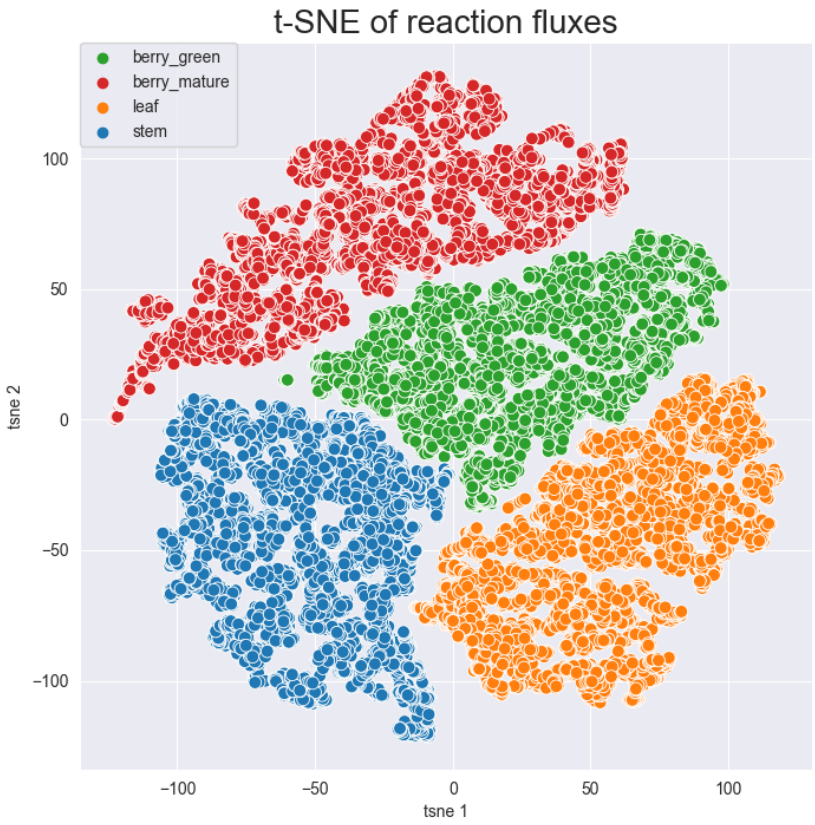


Figure 44: t-SNE visualisation of the sampled reaction fluxes of the tissue-specific models. Data was generated by ACHR sampler, scaled and filtered to only keep the reactions with differential flux between models. Orange dots represent the reaction fluxes from the leaf model, red dots the mature berry, blue dots the stem, and green dots the green berry.

As observed previously, despite the higher number of samples, t-SNE was able to separate well the fluxes from the different tissues as no overlap is evident between samples of different tissues. In addition, there is no group where flux samples group better nor are more separated from the other groups. These results are not fully in line with those observed for the expression data: leaf flux samples do not appear to group better than the samples from the other tissues, and there is no evident overlap between green and mature berry samples. This may occur because only reactions with differential flux between at least two tissue models were considered in the analysis. Overall, flux data seems to separate the tissues better than gene expression data.

Hypergeometric enrichment tests were used to identify the pathways that presented differential flux between each pair of models. These results are available in the Supplementary File 12. Analysing the results, it was clear that smaller pathways were not selected even when only one reaction was not identified as having differential flux. Therefore, this method seems to be more suitable for analysing pathways with a large number of reactions. For this reason, the complete list of reactions with differential flux between the models was also analysed.

Comparing the green and mature berry models, reactions from glycolysis, TCA cycle, and related to nucleotide biosynthesis were identified as having differential flux. In addition, anthocyanin biosynthesis exhibited more flux in the mature berry, as well as some reactions involved in the biosynthesis of quercetin

and derivatives. This was expected as the mature berry has anthocyanins and a higher content of sugars in its biomass composition while having a lower content of nucleotides.

Nucleotide and anthocyanin pathways were also identified as having differential flux between mature berries and the remaining tissues (leaf and stem). Other pathways included glycolysis and gluconeogenesis, photosynthesis in the case of the leaf, and the pentose phosphate and gibberellin inactivation pathways in the case of the stem.

Comparing leaf with stem and green berry, the core pathways were identified as having differential flux, as expected, mainly photosynthesis, glycolysis, gluconeogenesis, and pentose phosphate pathway. In addition, the fluxes for the reactions of 4-aminobenzoate biosynthesis were also altered between these tissues.

In the case of stem and berry green models, besides glycolysis and pentose phosphate pathways, the reactions involved in folate metabolism, gibberellin inactivation, glutathione biosynthesis, and amino acid metabolism were also identified as having differential flux.

In summary, it was expected that the primary metabolic pathways would be identified as having differential flux between tissues, as tissue models have different demands for biomass precursors, and produce energy by different processes: the leaf performs photosynthesis, while the others perform aerobic respiration. Besides these, no relevant pathways were found to characterise the specific metabolism of each tissue.

4.5 Diel Multi-tissue Models

Diel multi-tissue models were created to analyse the metabolic interactions between the leaf, stem, and berry of *V. vinifera* in the light (day) and dark (night) phases of a diel cycle. Two models were created, one using the green berry tissue and the other using the mature berry. The resulting diel multi-tissue models include 32391 and 31999 reactions, and 29064 and 28710 metabolites for green and mature berries, respectively. The multi-tissue diel model is schematised in Figure 45, showing the different tissues, diel phases, and the connections between them.

pFBA was used to simulate the models, as described before for photorespiration conditions. The phenotype predictions are available in Supplementary File 13. A summary of the results is presented in Table 28 and the fluxes for the storage metabolites between light and dark are shown in Table 29.

Significant differences were found between the light and dark phases, mainly in the leaf, as photosynthesis and photorespiration occur in this tissue. The light phase starts with photosynthesis light reactions and carbon dioxide fixation through the Calvin cycle in the leaf. The resulting carbohydrates are then used to produce all biomass precursors. Starch, sucrose, malate, and some amino acids are stored to be used in the leaf during the dark phase. At night, the active pathways include aerobic respiration, starch degradation, glycolysis, pentose phosphate pathway, and citrate biosynthesis through the TCA cycle. Sucrose was expected to be produced at night but instead, the model uses fructose 6-phosphate from starch degradation to start glycolysis. The sucrose requirements for biomass are fulfilled by accumulating very

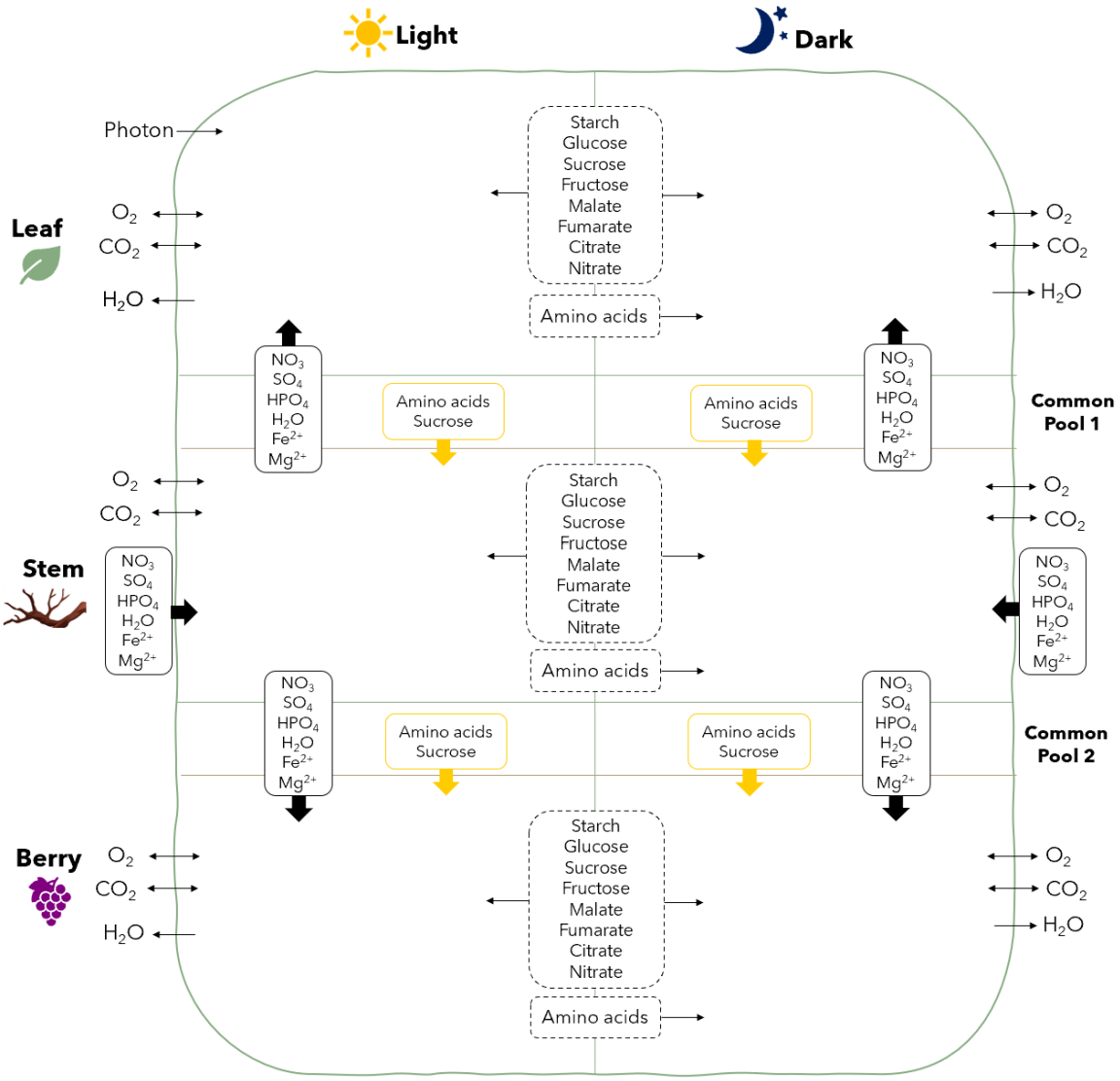


Figure 45: Schematic representation of the reconstructed diel multi-tissue model of *V. vinifera*, including the leaf, stem, and berry tissues and the common pools 1 and 2 in both light and dark phases. Photon uptake was allowed through the leaf in the light phase while mineral nutrients (nitrate, sulfate, phosphate, iron, magnesium) were allowed through the stem in both phases. Exchanges of carbon dioxide, oxygen, and water were allowed in all tissues and phases. Starch, glucose, sucrose, fructose, malate, fumarate, citrate, and nitrate were allowed to accumulate in the light and dark phases (dashed rectangle between phases). Amino acids can be stored in the light and used in the dark. Exchanges of amino acids, sucrose, and minerals were allowed between tissues through common pools.

small quantities of sucrose between light and dark phases (flux value less than 0.02 mmol.gDW^{-1}.h^{-1}). This is an artefact as the model finds sucrose transport to the dark phase less costly than producing it. However, sucrose production at night can be assured by forcing flux in the respective reactions. Starch is the main carbon compound stored in the light and it is degraded in the dark phase to produce energy. This was expected as less energy is needed to mobilise plastidic starch reserves than vacuolar sucrose [148].

Then, the citrate produced at night is stored in the vacuole to be used during the day, entering the TCA cycle. Nitrate is transported from the dark to the light to support nitrogen assimilation, which was predicted to occur only during the day. These results were confirmed by experimental evidence and observed with other plant models [148, 165, 294]. However, the entire TCA cycle was expected to occur in the dark phase. This does not happen in the leaf as all citrate is stored in vacuoles at night to be used during the day. Alpha-ketoglutarate is produced from the degradation of amino acids like glutamate and enters the cycle, which is complete until citrate production. The citrate accumulated, besides feeding the TCA cycle in the light phase, is used for the biosynthesis of acetyl Co-A during the day, which is then used for lipid production. Therefore, the model finds it more efficient to store more citrate to be used during the day than to complete the TCA cycle in the leaf at night.

Ammonium is provided by the stem and transported to the leaf, where it is used for amino acid biosynthesis. In addition, phosphate, sulfate, pyruvate, formate, and glutamate are imported from the stem through common pool 1 in the light and dark phases to feed amino acid and citrate biosynthesis.

On the other hand, sugars and amino acids produced in the leaf are transported to the stem. The active pathways in the stem during the day have much lower fluxes than in the leaf. These include aerobic respiration, sucrose degradation, glycolysis, starch biosynthesis, pentose phosphate pathway, amino acid, and nucleotide biosynthesis, and degradation of beta-alanine and uracil.

The leaf and stem metabolisms in the dark phase are similar and the same metabolites are stored between the light and dark phases. Also, in the stem, the TCA cycle is complete during the night, as expected, but a high percentage of the produced citrate is still stored (around 59%). The berry metabolism is very similar to the stem metabolism, but the reaction fluxes are even lower, except for the reactions related to folate biosynthesis. Formate and pyruvate are produced here and transported to the stem through common pool 2 to be further transported from the stem to the leaf to be used for amino acid biosynthesis. Only starch, methionine, and citrate are exchanged in the berry between light and dark phases. Also, the percentage of storage citrate is much lower, around 1% of the citrate produced at night.

No significant differences were found in the phenotype predictions between the grapevines with green and mature berries. The total biomass rate is slightly higher in the green berry model (0.149 h^{-1}) than in the mature one (0.142 h^{-1}), and generally, the photosynthetic pathways and those related to cellular respiration have lower flux in the mature berry. The pathways related to secondary metabolite biosynthesis, mainly anthocyanins, have flux in the mature and not in the green berry, as expected, but these fluxes are very low; thus, no major differences were observed in the primary metabolism.

Table 28: Summary of the phenotype predictions for the diel multi-tissue models of green and mature berries under photorespiration with biomass maximization as objective function and fixing the photon uptake to 300 mmol.gDW^{-1}.h^{-1}. This Table shows the metabolites that are consumed and produced by the models. The fluxes of the metabolites are in mmol.gDW^{-1}.h^{-1} while for biomass are in h^{-1}.

	photorespiration	
metabolite	**green**	**mature**
uptake		
Light__light	300.000	300.000
NITRATE__light	1.147	1.589
NITRATE__dark	0.764	1.059
OXYGEN-MOLECULE_dark	5.999	5.818
PROTON_light	26.330	27.210
PROTON_dark	16.790	16.020
SULFATE_light	0.098	0.112
SULFATE_dark	0.002	0.001
WATER_light	31.070	29.770
CARBON-DIOXIDE_light	35.830	34.920
production		
OXYGEN-MOLECULE_light	39.810	40.84
WATER_dark	1.455	2.788
CARBON-DIOXIDE_dark	0.000	0.000
HCO3_light	2.995	2.956
HCO3_dark	5.602	4.475
Pi_light	0.527	0.384
Pi_dark	0.527	0.384
total biomass	0.149	0.142

4.5.0.1 Nitrate Assimilation

Nitrogen is one of the most important nutrients that plants need to capture from the soil, as it is a constituent element of several essential compounds, such as nucleic acids and proteins, thus being a limiting factor for plant growth, development, and survival. Previous studies have described that plants adapt to low availability of nitrogen by reducing photosynthesis, retarding growth, and accumulating anthocyanins, which are important compounds for grape and wine flavour and aroma. Hence, low nitrogen content is expected to increase anthocyanin production while decreasing plant growth [269, 295, 296]. A very high content of nitrogen is also expected to decrease growth and anthocyanin production, as the energy is redistributed to nitrogen assimilation.

The *V. vinifera* green diel multi-tissue model was simulated to assess the effect of different nitrate concentrations on its metabolism. Four different flux values for nitrate uptake were tested: 0.1, 0.5, 5, and 10 mmmol.gDW^{-1}.h^{-1}. The choice of these values was arbitrary but the goal was to have two values above and two values below the unrestricted nitrate uptake flux (Table 28). Although plants can also uptake nitrogen from ammonium, ammonium uptake was restricted to zero for simplicity.

The full results are available in Supplementary File 14 and schematised in Figure 46. For nitrate values of 0.1 and 0.5 mmmol.gDW^{-1}.h^{-1}, the maximum flux for total biomass and for the production of all

Table 29: Fluxes of the metabolites stored between light and dark phases in the diel multi-tissue models. Positive fluxes indicate that the metabolites are stored in the light phase to be used in the dark while the metabolites with negative fluxes are stored in the dark to be used during the day. The fluxes are in mmmol.gDW^{-1}.h^{-1}

	reaction	green	mature
leaf	**CIT__vacu_leaf_light_dark_storage**	**-1.354**	**-1.834**
	CYS__cyto_leaf_light_dark_storage	0.009	0.009
	ILE__cyto_leaf_light_dark_storage	0.025	0.024
	MAL__vacu_leaf_light_dark_storage	1.168	1.272
	MET__cyto_leaf_light_dark_storage	0.011	0.011
	NITRATE__vacu_leaf_light_dark_storage	**-0.020**	**-1.059**
	PRO__cyto_leaf_light_dark_storage	0.360	1.143
	Starch__chlo_leaf_light_dark_storage	0.374	0.365
	SUCROSE__vacu_leaf_light_dark_storage	0.020	-
	THR__cyto_leaf_light_dark_storage	0.051	0.028
stem	**CIT__vacu_stem_light_dark_storage**	**-0.261**	**-0.160**
	CYS__cyto_stem_light_dark_storage	0.003	0.003
	ILE__cyto_stem_light_dark_storage	0.009	0.008
	MAL__vacu_stem_light_dark_storage	0.060	0.044
	MET__cyto_stem_light_dark_storage	0.015	0.004
	NITRATE__vacu_stem_light_dark_storage	**-0.745**	-
	PRO__cyto_stem_light_dark_storage	0.276	-
	SUCROSE__vacu_stem_light_dark_storage	-	0.027
	Starch__chlo_stem_light_dark_storage	0.009	-
berry	**CIT__vacu_berry_light_dark_storage**	-	**-0.021**
	CYS__cyto_berry_light_dark_storage	0.009	0.013
	ILE__cyto_berry_light_dark_storage	0.025	0.035
	MAL__vacu_berry_light_dark_storage	-	0.021
	MET__cyto_berry_light_dark_storage	-	0.016
	THR__cyto_berry_light_dark_storage	-	0.031
	PRO__cyto_berry_light_dark_storage	0.162	-
	Starch__chlo_berry_light_dark_storage	0.013	0.011

its components decreased. For instance, the maximum flux value of biomass decreased from 0.149 h^{-1} to 0.08 and 0.02 h^{-1} with a nitrate uptake of 0.5 and 0.1 mmmol.gDW^{-1}.h^{-1}, respectively. Conversely, the fluxes for the production of fatty acids have increased. The maximum flux for the production of fatty acids in the leaf during the day increased from 0.41 to 0.97 and 1.46 mmmol.gDW^{-1}.h^{-1}, with a nitrate uptake of 0.5 and 0.1 mmmol.gDW^{-1}.h^{-1}, respectively. Hence, as expected, the metabolism was redirected to the production of compounds that do not contain nitrogen.

Other pathways with decreased flux include the biosynthesis of phosphopantothenate, 4-aminobenzoate, tetrahydrofolates, nucleotides, and some amino acids, like histidine and tryptophan. The maximum possible flux for storage between light and dark phases also decreased under lower values of nitrate uptake.

On the other hand, there was an increase in the flux of primary pathways, comprising sucrose and starch biosynthesis and degradation, gluconeogenesis, and glycolysis. In addition, the pathways for DNA, RNA, uracil, glutathione, and amino acid degradation were increased to raise nitrogen availability. However,

the biosynthesis of some amino acids was also increased, such as the biosynthesis of phenylalanine, which is the precursor for the phenylpropanoid pathway.

As described in the literature, most secondary metabolic pathways had an increase in the maximum flux, including the biosynthesis of phenylpropanoids (flavonoids, anthocyanins, stilbenes), isoprenoids, and gibberellins. The maximum flux of secondary reactions increased as nitrate uptake decreased. For instance, the flux for resveratrol synthase reaction during the day was 0.41 mmmol.gDW^{-1}.h^{-1} in the unrestricted phenotype prediction and increased to 1.14 mmmol.gDW^{-1}.h^{-1} when restricting the uptake of nitrate to 0.5 mmmol.gDW^{-1}.h^{-1} and to 1.78 mmmol.gDW^{-1}.h^{-1} with a nitrate uptake of 0.1 mmmol.gDW^{-1}.h^{-1}.

For high nitrogen uptake values (5 and 10 mmmol.gDW^{-1}.h^{-1}), most pathways presented lower maximum fluxes than in the unrestricted phenotype prediction, except for some reactions of photosynthesis, chlorophyll, tetrahydrofolate, and ATP biosynthesis, as well as the biosynthesis of some amino acids, such as histidine.

The maximum flux for the total biomass reaction decreased from 0.149 h^{-1} to 0.142 and 0.134 h^{-1} with a nitrate uptake of 5 and 10 mmmol.gDW^{-1}.h^{-1}, respectively. A decrease in the flux was observed in the producing reactions of all biomass components as well as most secondary pathways. The flux of the resveratrol synthase during the day slightly decreased from 0.41 mmmol.gDW^{-1}.h^{-1} to 0.39 and 0.36 mmmol.gDW^{-1}.h^{-1} with a nitrate uptake of 5 and 10 mmmol.gDW^{-1}.h^{-1}, respectively. The maximum flux for the storage of some amino acids, such as serine and glutamate, decreased while it increased for others, such as valine and leucine. Despite these small changes in metabolism, the effects of low nitrate values appear to be more significant than the effects of high nitrate.

The results for the multi-tissue model with the mature berry were very similar. The decrease in total biomass was slightly higher in the green berry model for both low and high levels of nitrate. Conversely, with low nitrate values, the increase in the maximum flux was slightly higher in the mature berry for most secondary reactions.

Therefore, nitrate availability greatly affects plant metabolism, including secondary pathways. Controlling the levels of nitrate at certain stages of grape development is important to help control grape phenolic and sugar content, which will influence the flavour and quality of grapes.

4.5.0.2 Sulfate Assimilation

Sulfur is another important nutrient taken up by plants from the soil in the form of sulfate, and it is the key element of the amino acids cysteine and methionine. Thus, the major part of sulfate is used for protein biosynthesis. Sulfur is also a component of glutathione, which is an important antioxidant agent, and S-adenosyl methionine and coenzyme A, which are cofactors for several enzymes.

Elemental sulfur (S^0) is the oldest pesticide applied to grapevines and it is still widely used nowadays, being particularly effective against powdery mildew disease, one of the most common diseases affecting grapevines that is caused by the fungus *Erysiphe necator*. In addition, sulfur dioxide (SO_2) is usually used as a conservative of table grapes or in winemaking to prevent oxidation and microbial contamination.

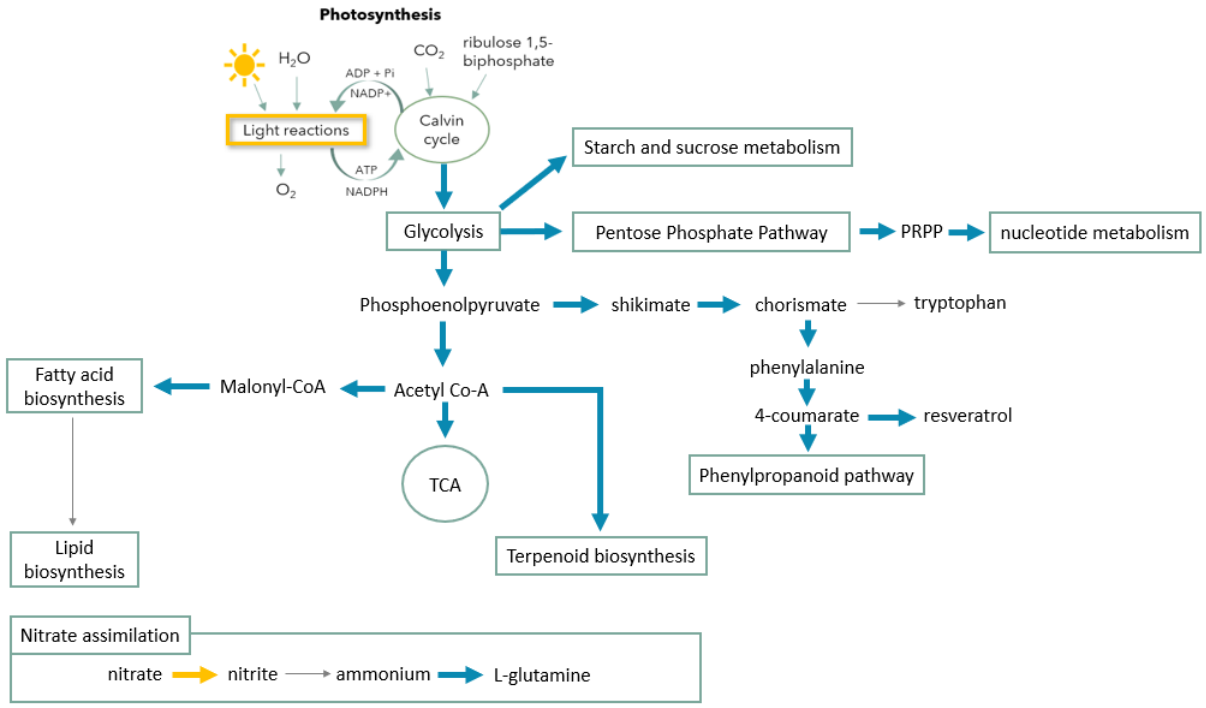


Figure 46: Simplified schema of the main metabolic pathways in the model affected by varying nitrate levels. Pathways with increased maximum flux under low nitrate conditions are highlighted with a thick blue arrow while pathways with increased flux under high nitrate conditions are highlighted with a thick yellow arrow. Pathways with decreased flux in both conditions are represented by a thin grey arrow.

Plant exposure to high sulfur levels can lead to the accumulation of sulfur-derived compounds or affect the metabolism of phenolic compounds, which can change the flavour, aroma, and texture of grapes and wine [269, 297]. It was observed that residual sulfur on berries can lead to the formation of undesirable flavors, such as hydrogen sulfide (H_2S), during wine fermentation [297].

The same approach was used to assess the effect of different sulfate concentrations in the *V. vinifera* model. In this case, FVA was performed with two sulfate uptake values, 0.01 and 10 mmol.gDW^{-1}.h^{-1}, and compared to the unrestricted phenotype predictions. Similar results were obtained for the multi-tissue with green and mature berry. Thus, only the results for the green multi-tissue are described. The full results are available in Supplementary File 15 and schematised in Figure 47.

With high sulfate (10 mmol.gDW^{-1}.h^{-1}), the maximum flux for biomass production decreased from 0.149 to 0.138 h^{-1}. Similarly, the production of all biomass components also decreased. As expected, the maximum fluxes of the reactions involved in sulfate assimilation and oxidation, and glutathione biosynthesis have increased. Surprisingly, in the model, the biosynthesis of cysteine and methionine decreased with high sulfate levels. During sulfate reduction, the reaction that produces H_2S has a higher maximum flux but the reaction that uses it to produce cysteine has a lower flux, which leads to a big increase in the flux of the H_2S exchange reaction. This could mean that when plants are exposed to high sulfur levels they try to adapt to these conditions by adjusting their metabolism, leading to the accumulation of H_2S or other sulfur compounds that can alter the flavour and aroma of the grapes. In addition, the primary pathways

carried less flux in high sulfur conditions. The fatty acid metabolism has also decreased, as well as the metabolism of nucleotides, amino acids, and chlorophylls. Some photosynthesis light reactions presented a higher minimum flux. However, the downstream reactions from the Calvin cycle had lower maximum flux values as well as the aerobic respiration pathway. The biosynthesis of secondary metabolites was also limited by higher levels of sulfur, which will also affect the flavour of grapes. As the biosynthesis of amino acids decreased, the maximum flux for amino acid storage also decreased with high sulfate levels.

When plants are under a sulfate deficiency, the production of biomass and all its components also decreases. For a sulfate uptake of 0.01 mmmol.gDW^{-1}.h^{-1}, the maximum production of biomass decreased from 0.149 to 0.018 h^{-1}. In addition, the biosynthesis of phosphopantothenate, cysteine, methionine, and coenzyme A, also decreased.

As observed previously for nitrogen deficiency, in low sulfur conditions, the plant has an excess of carbon and nitrogen skeletons, which are not being used for protein biosynthesis and are available for the synthesis of secondary metabolites. Therefore, there was an increase in the production of secondary metabolites, plant hormones, and amino acids that are precursors of secondary compounds, like phenylalanine. For instance, the maximum flux for resveratrol synthase reaction during the day increased from 0.41 to 1.75 mmmol.gDW^{-1}.h^{-1}. In addition, there was a great increase in the maximum flux for the storage of all amino acids, except for cysteine and methionine. Again, the changes in metabolism are more significant under low sulfate values. Hence, as for nitrogen, sulfur levels in the soil can greatly influence grapevine metabolism and affect the flavour and aroma of grapes by sulfide or sulfur-compounds accumulation or changes in the phenolic content in grapes.

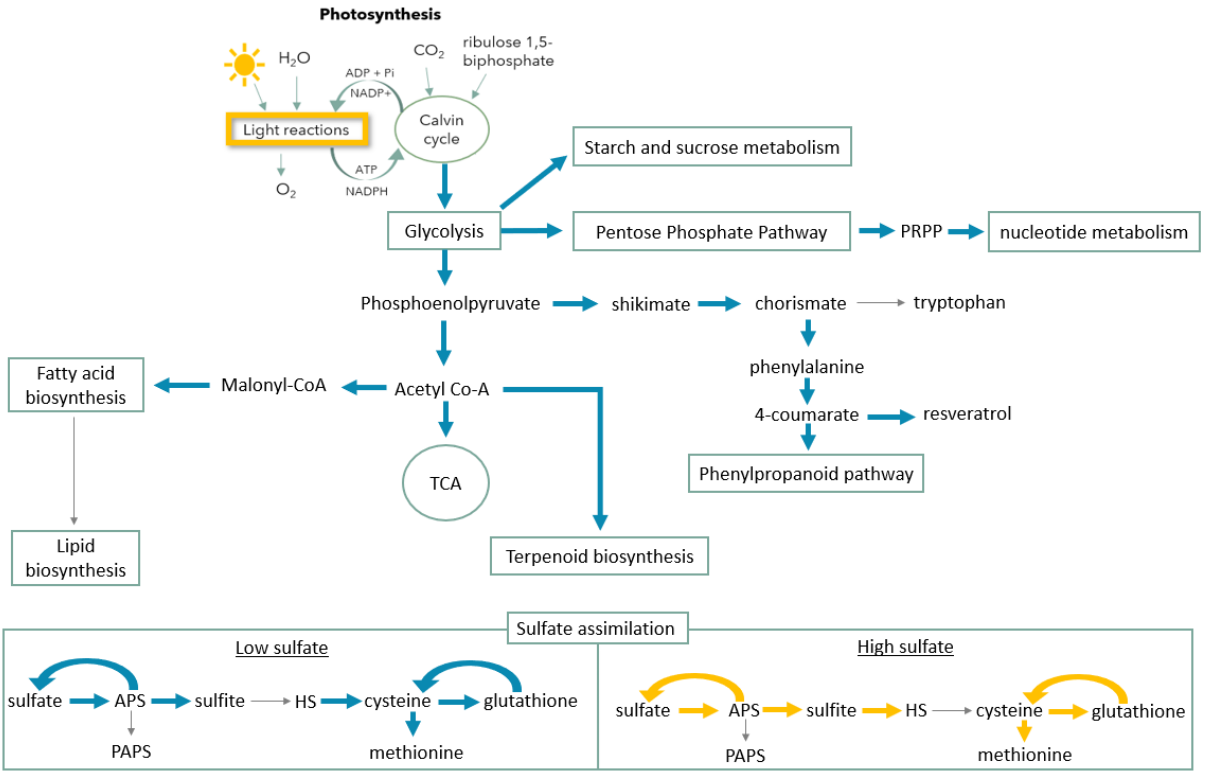


Figure 47: Simplified schema of the main metabolic pathways in the model affected by varying sulfate levels. Pathways with increased maximum flux under low sulfate conditions are highlighted with a thick blue arrow while pathways with increased flux under high sulfate conditions are highlighted with a thick yellow arrow. Pathways with decreased flux in both conditions are represented by a thin grey arrow.

5

Combining omics and genome-scale metabolic models with machine learning

The analysis of fluxomics with machine learning of this chapter was included in the publication submitted (available as a pre-print) and is waiting for revision:

Sampaio, M., Rocha, M., and Dias, O. A diel multi-tissue genome-scale metabolic model of Vitis vinifera. Submitted. Pre-print available at: https://doi.org/10.1101/2024.01.30.578056

The multi-omics data integration in this chapter was included in the publication submitted to the *10th IFAC International Conference on Foundations of Systems Biology in Engineering* and is waiting for revision:

Sampaio, M., Rocha, M., and Dias, O. A multi-omics approach for understanding grape metabolism throughout development. Submitted.

The development of novel technologies has allowed the collection of different molecular data from the same samples, such as gene expression, DNA methylation, protein abundance, and metabolite levels, resulting in multiple omics (multi-omics) datasets. Integrating multi-omics allows the identification of interactions between omics layers that can help identify complex biological relationships responsible for specific phenotypes. These interactions between omics are not considered when analysing omics data from a single type. However, as omics datasets have different biological contexts, numerical scales, formats, and noisiness, their integration is still a challenging task. ML has been used to process, analyse, and integrate different types of omics. ML models can perform data reduction, identify underlying patterns in data, and select relevant features, allowing the extraction of new biological insights from the data. However, ML can be hard to interpret, which can lead to wrong inferences.

The combination of ML and GSMM has been explored recently and holds much promise as these approaches are complementary and the combination of both could help overcome the intrinsic limitations of each [8–11, 168]. On the one hand, ML is crucial for interpreting large omics datasets and extracting knowledge, but does not take into account the existing knowledge and can be difficult to interpret. On the other hand, GSMMs are reconstructed based on prior knowledge about the metabolic pathways and are easy to interpret but generally present many gaps in the networks and do not include the regulatory effects. Therefore, combining both approaches can help generate more accurate predictions and result in more meaningful insights from the data.

With this in mind, the fluxomics data obtained from reconstructed berry models were analysed by ML. Although the quality of the fluxomics data depends greatly on the quality of the model network, the model can be constrained with gene expression data to produce more realistic simulations.

A significant number of samples is required to train ML models to provide a more diverse and realistic representation of the problem and improve the model's ability to generalise to new data. However, omics datasets usually have few samples and a high number of features, which results in complex overfitted models with poor generalisation power. Therefore, the idea of using ML to analyse fluxomics data focuses mainly on the study of the impact of the features on the output class, rather than the creation of accurate predictive models, due to the low data dimensions that make model validation difficult.

The RNA-Seq dataset used to create the grape-specific models in the previous chapter is the largest public RNA-Seq dataset of *V. vinifera*, containing 219 grape samples with replicate measurements for each biological sample at different developmental stages [274]. Additionally, metabolomics data is available for the same samples. In the previous chapter, only samples from Cabernet Sauvignon were used. In this analysis, the samples of Pinot Noir cultivar were also included to increase the number of samples in the dataset.

Therefore, ML was applied to fluxomics data generated from *V. vinifera* GSMMs that were constrained by the RNA-Seq data to predict the grape developmental state, green or mature. ML was also applied to the RNA-Seq and metabolomics datasets separately, to compare the results with those obtained with fluxomics and see whether fluxomics data can add more value to the classification task.

As these grape samples included both gene expression and metabolomics data, and now also generated fluxomics data, the relationships between genes, metabolites, and reactions during grape development were further analysed by integrating the different types of omics available. The rationale for integrating fluxomics from the GSMMs with experimental omics is to improve the interpretability of the omics analysis and provide a cellular state closer to the phenotype, which can lead to meaningful insights into the biological processes linking the genotype to the phenotype.

5.1 Machine Learning and Omics data

5.1.1 Creation of the fluxomics dataset

The complete RNA-Seq dataset from Fasoli et al. [274] was used to create the fluxomics data for each sample. It contains 219 grape samples with replicate measurements for each biological sample at different developmental stages. The dataset was transformed by log2 and the mean value of all replicates was calculated to represent each sample. In total, the dataset included 73 samples, 55% from Cabernet Sauvignon and 45% from Pinot Noir, and it was filtered to include only the expression of the 6018 genes in the model.

As performed before, the time-point metadata for these samples was discretized into two developmental stages, green and mature, which represent the output class to be predicted by the models. Samples until time point 4 were considered to be in the green state, while samples after time point 4 were considered mature. Hence, this dataset presents a small class imbalance, where 59% of samples belong to the mature state and 41% are green. Then, the final RNA-Seq dataset was integrated with the generic *V. vinifera* model to create GSMMs representing each sample in the dataset. Thus, 73 models were created using the FASTCORE algorithm [275] as described in the previous chapter for the tissue-specific models and using the same strategy.

The resulting sample-specific models were simulated using FVA and the Flux Capacity (FCa) of each reaction was calculated by subtracting the maximum and minimum flux obtained for each reaction while keeping 80% of the maximum biomass value. Before running FVA, all reactions were made irreversible to facilitate the interpretation of results. Thus, all reversible reactions were split into two irreversible reactions, forward and backward reactions, whose lower bound is equal to or greater than zero. The reactions absent in the models were considered to have an FCa of 0.

$$Flux\ Capacity(FCa) = Flux_{\text{max}} - Flux_{\text{min}} \tag{5.1}$$

5.1.2 Machine learning pipeline

The analysis of omics with ML was performed in Python (v. 3.11), using mainly the package scikit-learn (v. 1.2.2). The scripts are available at the https://github.com/martaS95/MM_ML GitHub repository. First,

the original RNA-Seq data was log-transformed (by log2) and included the expression of 35336 genes of *V. vinifera*.

For the unsupervised analysis (Figure 48), datasets were filtered to remove the features with low variance across all samples. The replicate samples in RNA-Seq and metabolomics were kept in the datasets. The fluxomics dataset was standardized to z-scores. t-SNE was applied to visualise the data distribution.

For the supervised analysis (Figure 49), the mean value of all replicates of each sample was calculated for the RNA-Seq and metabolomics datasets. The 73 samples were divided into train and test sets using cross-validation with 10 folds and repeated 10 times. In each iteration, feature selection was performed using *VarianceThreshold* with a threshold of 0.001, and, for RNA-Seq and fluxomics datasets, as the number of features was still high, the *SelectKBest* function was used to select the 500 most relevant features based on ANOVA F-values. For metabolomics, only the *VarianceThreshold* filter was applied. In addition, the resulting datasets were also scaled by *StandardScaler*. Then, an ML model fitted the train data and predicted the output classes for the test set. Given the reduced number of samples in the dataset, simpler ML architectures were chosen. These included Logistic Regression (LR), KNN, decision trees (DT), SVM, and RF. These models were assessed by different metrics, such as recall (equation 2.9), precision (equation 2.10), balanced accuracy (equation 2.12), and f1 score (equation 2.13), which were averaged across all train-test splits.

The importance of each feature in predicting the output was analysed by calculating the SHAPley Additive exPlanations (SHAP) values for each classifier. SHAP values are defined based on the contribution of each feature to the prediction of each sample and are used to increase the interpretability of ML models. Features with larger absolute SHAP values have a larger effect on the prediction [298]. The SHAP values for each fold of the repeated cross-validation were calculated, and the average SHAP values for each sample were obtained to give a more stable representation of the feature contributions.

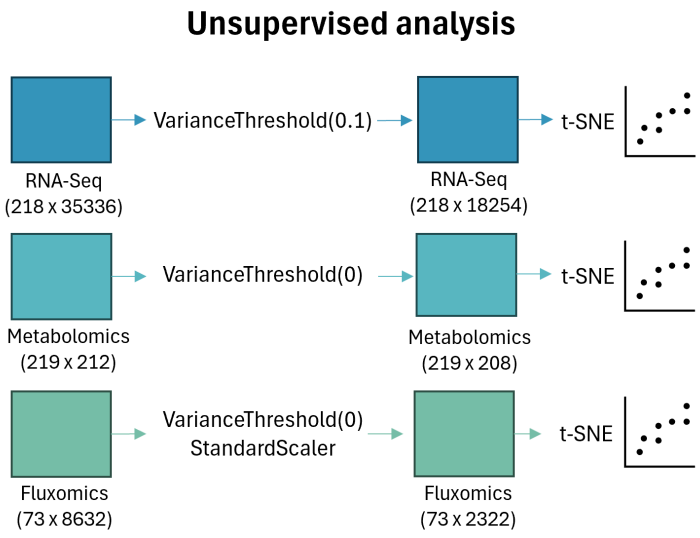


Figure 48: Schema used for the unsupervised analysis of omics data. The original datasets were filtered using *VarianceThreshold* and the fluxomics dataset was scaled. Then, t-SNE was used to visualise the data.

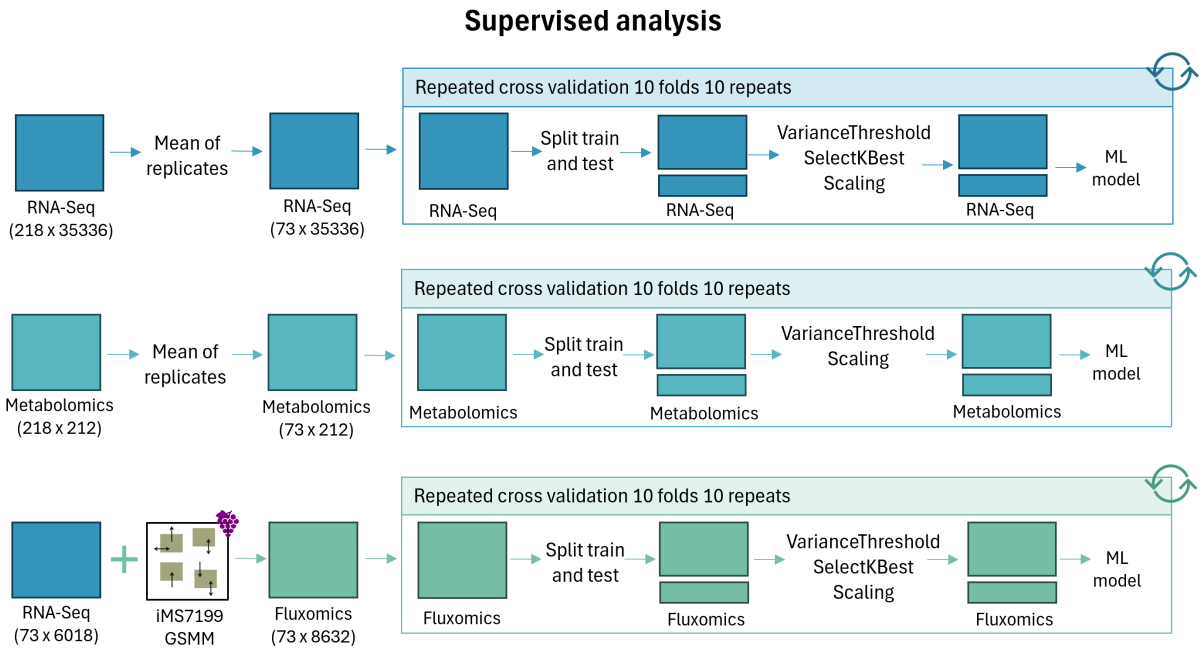


Figure 49: ML pipeline for the supervised analysis. Repeated cross-validation with 10 folds and 10 repeats was applied to split the train and test sets several times and increase the robustness of the results. These sets were filtered and scaled and the ML model was applied in each iteration.

5.1.3 Results of fluxomics analysis

Fluxomics data was obtained from the 73 GSMMs and analysed by ML to predict the grape developmental phase, green or mature. For unsupervised analysis, data was preprocessed and explored using t-SNE for visualisation. The *VarianceThreshold* filter greatly reduced the dataset from 8632 to 2322 features.

The results of the t-SNE are plotted in Figure 50. A clear separation between the two states is shown as all green samples are located at the top of the plot, while most mature samples are at the bottom. However, six mature samples were grouped with the green ones. These comprise Cabernet Sauvignon samples from time points 5 and 6 and Pinot Noir samples from time point 5. As samples of time point 4 were still considered green phase, these results indicate that not many differences exist at the fluxomics level between green and early mature samples. Even so, fluxomics seems to distinguish well the remaining green and mature samples.

For the supervised analysis, repeated stratified cross-validation with 10 folds and 10 repeats was used to evaluate the models. The average evaluation results for all models are shown in Table 30.

According to these results, the models performed well in predicting the grape developmental stage with this fluxomics dataset. The model's performances across the different folds are robust, indicating that the models can learn meaningful patterns in the data and handle data variations well. Overall, RF obtained higher values for all metrics, while the SVM model presented the worst performance. Despite the results being good and the models being able to generalise correctly on the test set of each fold, the dataset is very small (73 samples), which is a common problem when working with omics data. Larger datasets are

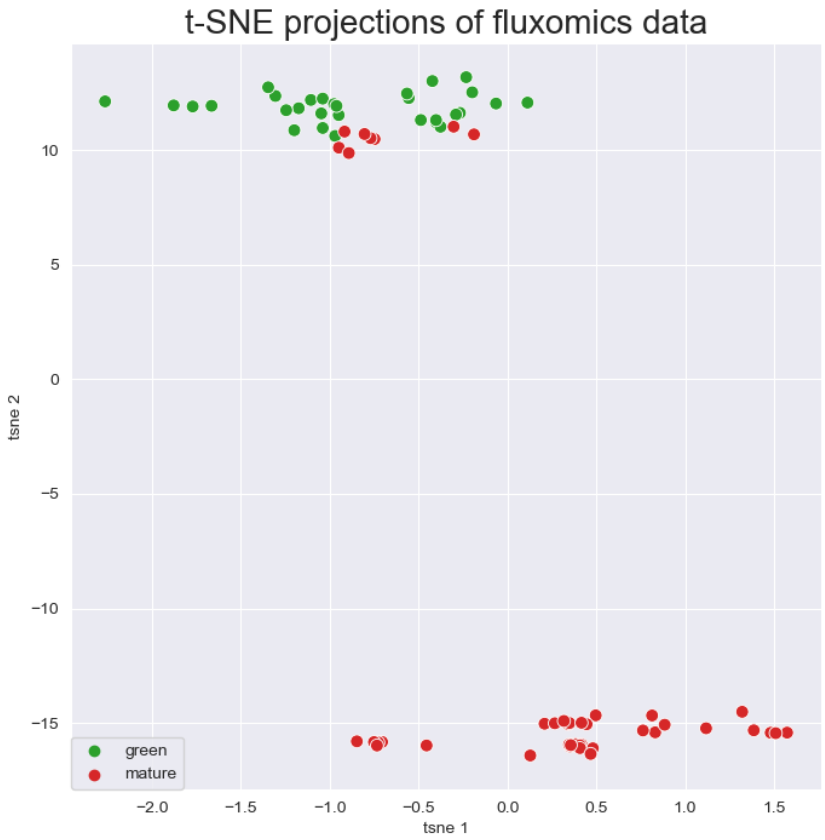


Figure 50: t-SNE visualisation of the fluxomics data obtained from the 73 context-specific models. Samples are coloured by grape developmental stage.

Table 30: Evaluation results of the ML models trained with fluxomics data for the metrics balanced accuracy, precision, recall, and F1 score.

	LR	**KNN**	**DT**	**SVM**	**RF**
balanced accuracy	0.96	0.95	0.96	0.93	0.97
precision	0.97	0.96	0.98	0.96	0.98
recall	0.96	0.98	0.95	0.94	0.97
F1 score	0.96	0.97	0.96	0.94	0.97

needed to create better predictive models and draw more conclusions from the data. Nevertheless, these models represent a good starting point for understanding which reactions contribute most to the model's predictions.

Hence, SHAP values were calculated for the two best models, RF and KNN, and the most contributing reactions are shown in Figure 51.

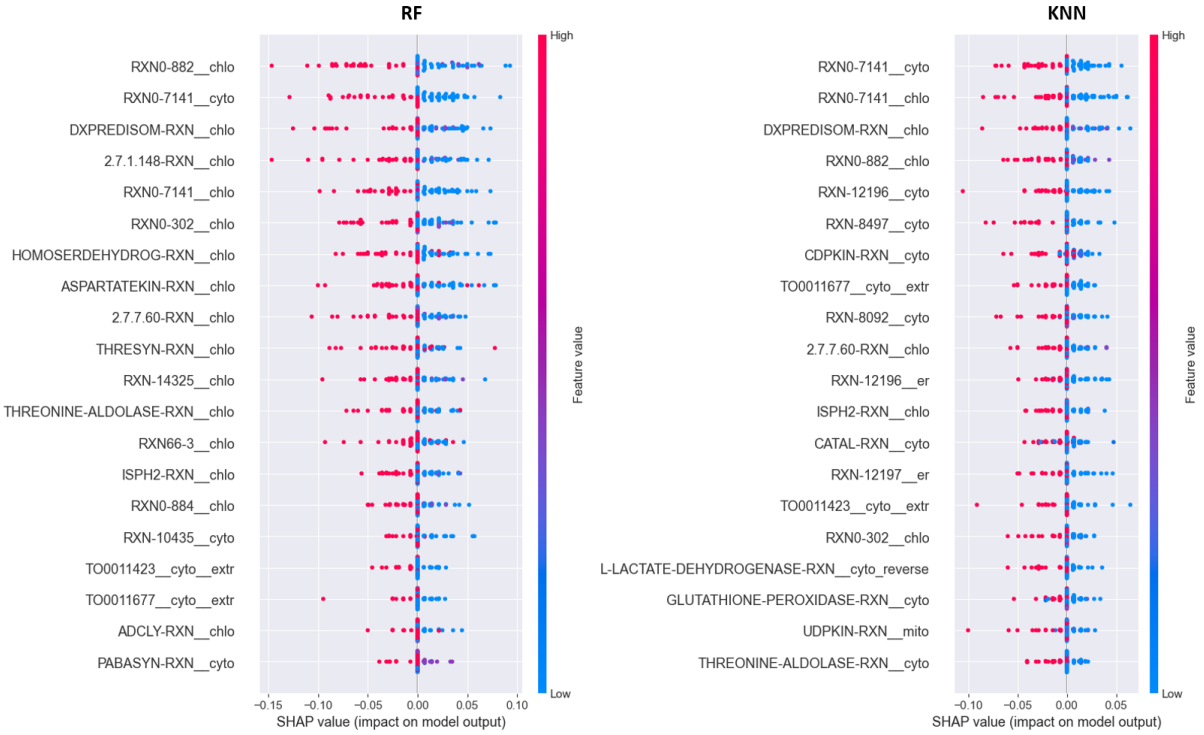


Figure 51: Beeswarm plot of SHAP values for the reactions that contribute most to RF's (leaft) and KNN's (right) predictions. Features are ordered from higher (top) to lower (bottom) effects on the predictions. The dots represent a single observation and the colour indicates if the observation has a higher (pink) or a lower (blue) feature value compared to the other observations.

Overall, when the reactions presented high FCa values, they had negative SHAP values, leading the model to predict the green state, while with lower FCa they exhibited higher positive SHAP values, leading the model to predict the mature state. Hence, all the reactions identified here presented an average FCa value higher in the green than in the mature grapes. There are some exceptions to this trend, such as the chloroplastic THREONINE-ALDOLASE-RXN and THRESYN-RXN reactions that show positive SHAP values when presenting high flux for the RF model, indicating that these reactions have high flux in some samples classified as mature by the models.

For the RF model, RXN0-882 in the chloroplast has the highest contribution to the predictions. High fluxes of this reaction have a negative impact on the model (SHAP value around -0.15), classifying the samples as green while lower fluxes have a positive impact (SHAP value close to 0.10), classifying the samples as mature. For the KNN model, the RXN0-7141 in the cytosol is the reaction that most contributes to the predictions. Similarly, high fluxes of this reaction have a negative contribution (SHAP value lower than -0.05), while lower fluxes have positive SHAP values (SHAP value higher than 0.05). Hence, high

fluxes of this reaction are associated with the green state, while low fluxes are linked to the mature state.

Of the 20 reactions identified for each model, 10 have a high impact on both models, indicating that the results are reliable and robust and that these features are important for predicting the output. Most of these reactions are involved in the MEP pathway, which is responsible for the biosynthesis of the terpenoid precursors (Figure 3). The other reactions in common are associated with threonine degradation into glycine (THREONINE-ALDOLASE-RXN), and the transport of glycerides (TO0011677 and TO0011423).

The reactions identified only for the RF model are also involved in the MEP pathway (2.7.1.148-RXN and RXN0-884), ethanol degradation (RXN66-3), and the biosynthesis of jasmonate derivatives (RXN-10435), nucleotides (RXN-14325), 4-aminobenzoate (ADCLY-RXN and PABASYN-RXN), and the homoserine and threonine amino acids (THRESYN-RXN, ASPARTATEKIN-RXN, HOMOSERDEHYDROG-RXN).

Analysing the reactions identified only for the KNN model, RXN-8092 is similar to the reaction RXN66-3 already identified for RF. The remaining are related to several different pathways including fermentation (L-LACTATE-DEHYDROGENASE-RXN), the degradation of triphosphate nucleotides (RXN-12196 and RXN-12197) and reactive oxygen species (CATAL-RXN and GLUTATHIONE-PEROXIDASE-RXN), and the biosynthesis of fatty acids (RXN-8497) and nucleotides (CDPKIN-RXN and UDPKIN-RXN).

The accumulation of terpenoids in grapes typically starts before veraison, which can explain why the reactions associated with the biosynthesis of terpenoid precursors had higher fluxes in the green state. However, terpenoid biosynthesis intensifies after veraison, which is not observed in the fluxes of these reactions. Fasoli et. al [274] has also identified terpene metabolism as a negative biomarker for the onset of ripening. In addition, the abscisic acid (ABA) signalling is increased at veraison, and ABA is derived from carotenoids, whose biosynthesis starts with the MEP pathway. Thus, there is strong evidence that genes or reactions from the MEP pathway could be used as biomarkers for the onset of ripening.

In the green phase, as grapes are rapidly growing, the metabolism of amino acids, nucleotides, lipids, and oxidative stress is expected to be more active than in the mature phase. 4-aminobenzoate is a precursor for the biosynthesis of various metabolites, such as tetrahydrofolates, which are involved in several processes like photorespiration, amino acid metabolism, and protein biosynthesis. These pathways are also expected to be more active in the green phase. This fact may explain why the reactions related to these pathways are important for the model's predictions. However, it is not clear why threonine metabolism is more important for the model than the metabolism of other amino acids.

Although most reactions that have a higher contribution to the model's predictions are associated with pathways that appear to be more active in the green phase, these pathways were not the most expected as the major differences between green and mature grapes are their content in organic acids, sugars, and phenolic compounds. This could be due to the limitations in the metabolic models, mainly in the GPR rules or biomass definition, which forces the production of some biomass precursors to allow the simulations, reducing the differences between the fluxes of these reactions in the different phases.

This analysis provided an overview of the most important reactions for the models to classify the samples as green or mature, considering all samples. For a more detailed view of what happens at the onset of ripening, the samples collected at veraison (time point 3) were selected and the SHAP values were analysed for these samples, using the RF model. The mean absolute SHAP value for the veraison

samples is shown in Figure 52.

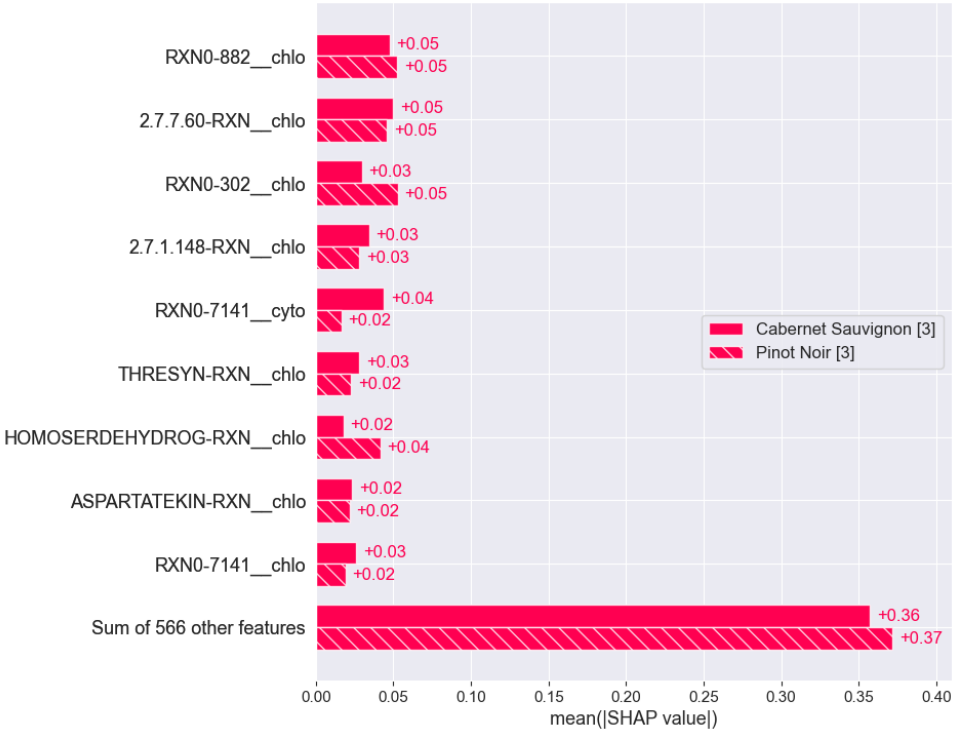


Figure 52: Mean absolute SHAP values of the reactions contributing the most for the classification of the Cabernet Sauvignon and Pinot Noir samples at veraison (time point 3).

No new reactions were found with this analysis. All reactions identified here were already identified in the previous analysis and are associated with the MEP pathway and the biosynthesis of threonine.

In addition, no major differences were identified between the two cultivars. Regarding the MEP pathway, the reaction RXN0-302 contributes more to the classification of Pinot Noir samples while the reaction RXN0-7141 has a higher impact on the classification of Cabernet Sauvignon samples. The HOMOSERDEHYDROG-RXN reaction from the threonine biosynthetic pathway seems to have a higher impact in classifying the samples of Pinot Noir. Therefore, reactions from MEP and threonine biosynthesis pathways should be further studied for the identification of biomarkers for the onset of ripening.

5.1.4 Results of RNA-Seq analysis

This time, ML was trained with RNA-Seq data to check whether the results were similar to the ones obtained with fluxomics and if the same pathways were identified.

Regarding unsupervised analysis, the t-SNE plot for the RNA-Seq dataset is shown in Figure 53. Green samples grouped well, while two groups of mature samples were observed. In fact, these mature samples are grouped by cultivar. This means that the gene expression patterns of Cabernet Sauvignon and Pinot Noir are more similar in the green phase than in the mature phase. Nevertheless, the samples grouped well by developmental state, and only a small overlap was observed between green and mature samples, which is expected as this is a time-course experiment.

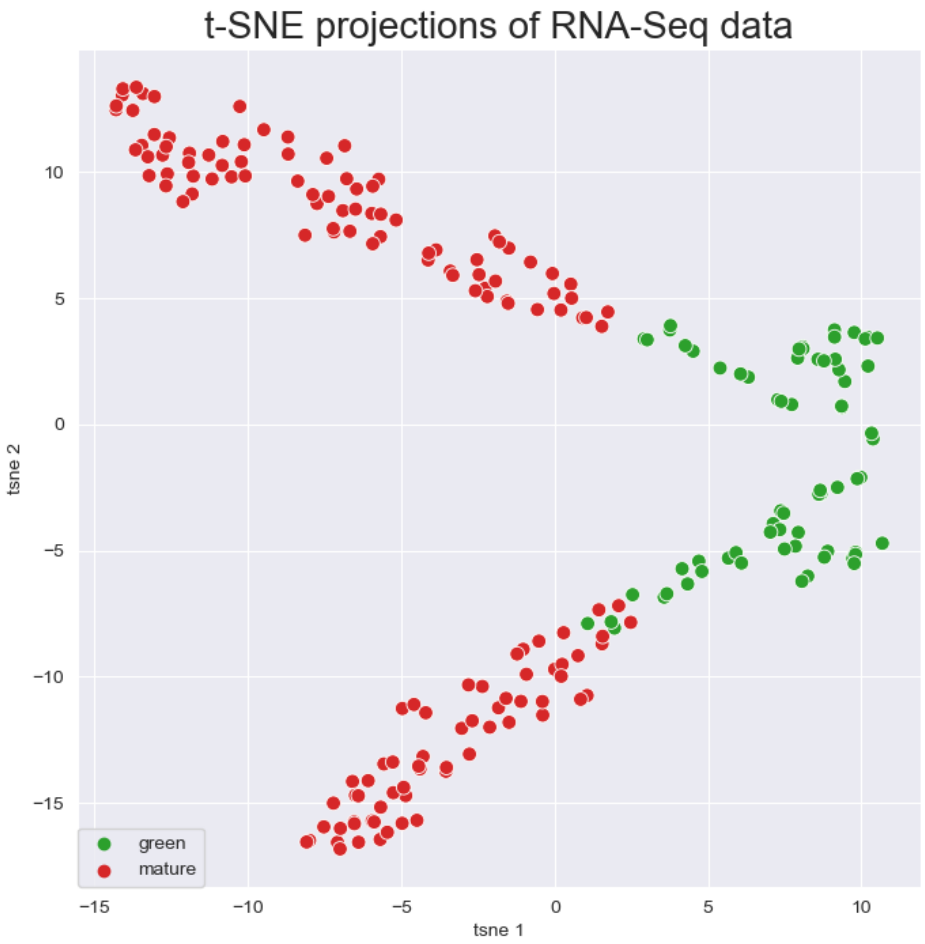


Figure 53: t-SNE visualisation of the RNA-Seq data considering replicates. Samples are coloured by grape developmental stage.

For supervised analysis, the same ML model architectures were applied and the average evaluation results for all models are shown in Table 31. Once again, the models obtained good results, being able to correctly classify most test samples at each fold and repetition. These results are similar to those obtained with fluxomics data, but this time KNN and SVM are the models with the highest scores and perform slightly better than the best models obtained from fluxomics data.

Table 31: Evaluation results of the ML models trained with RNA-Seq data for the metrics balanced accuracy, precision, recall, and F1 score.

	LR	**KNN**	**DT**	**SVM**	**RF**
balanced accuracy	0.96	0.97	0.96	0.97	0.97
precision	0.96	0.96	0.97	0.97	0.97
recall	0.98	1.00	0.96	1.00	0.98
F1 score	0.97	0.98	0.96	0.98	0.97

Then, SHAP values were calculated for the two best models, and the most contributing genes are shown in Figure 54.

As observed before, high values of these genes correspond to negative SHAP values, indicating that these are overexpressed in the green phase, except for five genes, Vitvi12g01876, Vitvi12g02659, Vitvi12g01880, Vitvi17g00621, and Vitvi06g01203, which are overexpressed in the mature phase and are related to seed storage proteins, fatty acid degradation, and biosynthesis of glucosinolates.

The genes that contribute the most to model classification mainly express transcription factors, ribosomal proteins, seed storage proteins, and proteins responsible for cell proliferation, homeostasis, and response to oxidative stress. Hence, these genes are related to regulatory processes and are not directly involved in metabolism. The metabolic genes identified in this analysis are related to jasmonate inactivation (Vitvi08g00944), hydrocarbon (Vitvi07g01182), carotenoid (Vitvi16g01099), and glucosinolate (Vitvi06g01203) biosynthesis, asparagine tRNA charging (Vitvi11g00224), and fatty acid (Vitvi17g00621) and phenolic (Vitvi06g01487) degradation. Some of these pathways were also identified with fluxomics, namely fatty acid metabolism and amino acid biosynthesis. However, the main pathways described in the literature as being differentially active during grape development, such as photosynthesis and phenylpropanoid biosynthesis, were not identified here [274, 286, 299]. Therefore, this analysis suggests that the most important genes for distinguishing green and mature grapes are involved in transcription regulation and cell growth. These findings could only be obtained with the RNA-Seq dataset, as the fluxomics dataset analysed before only included information on metabolic genes. However, the interpretation of these results is more difficult as many genes are still uncharacterized, and the effect of their overexpression on the phenotype is not as evident as analysing the metabolism directly.

5.1.5 Results of metabolomics analysis

Similarly, ML was also applied to metabolomics data to see which information could be retrieved from this data. Regarding unsupervised analysis, the t-SNE plot for the metabolomics dataset is shown in Figure 55. The green samples appear to be closer, indicating that different cultivars have a metabolic content more similar in the green phase than in the mature. However, some green samples are far from the other green samples and overlap the mature samples. This was different from the overlap observed with fluxomics where some mature samples were closer to green samples. In addition, this overlap is more extensive than the ones observed from RNA-Seq and fluxomics data, suggesting that it is more difficult to distinguish the green and mature samples with metabolomics data.

The supervised ML models were evaluated with the same metrics to confirm this assumption (Table 32). The results show that the models trained with metabolomics have slightly worse metrics than the models trained with the other datasets, but still present good results. RF was the model with the highest scores and performed better with metabolomics than with the other datasets. Hence, SHAP values were calculated for RF and the metabolites most contributing to the classification of green and mature samples are shown in Figure 56.

In the case of metabolites, high levels usually correspond to positive SHAP values, indicating that these metabolites present high levels in mature grapes. Exceptions to this trend are the metabolites threonic acid, malic acid, and tartaric acid, which present higher levels in the green phase. This was expected as

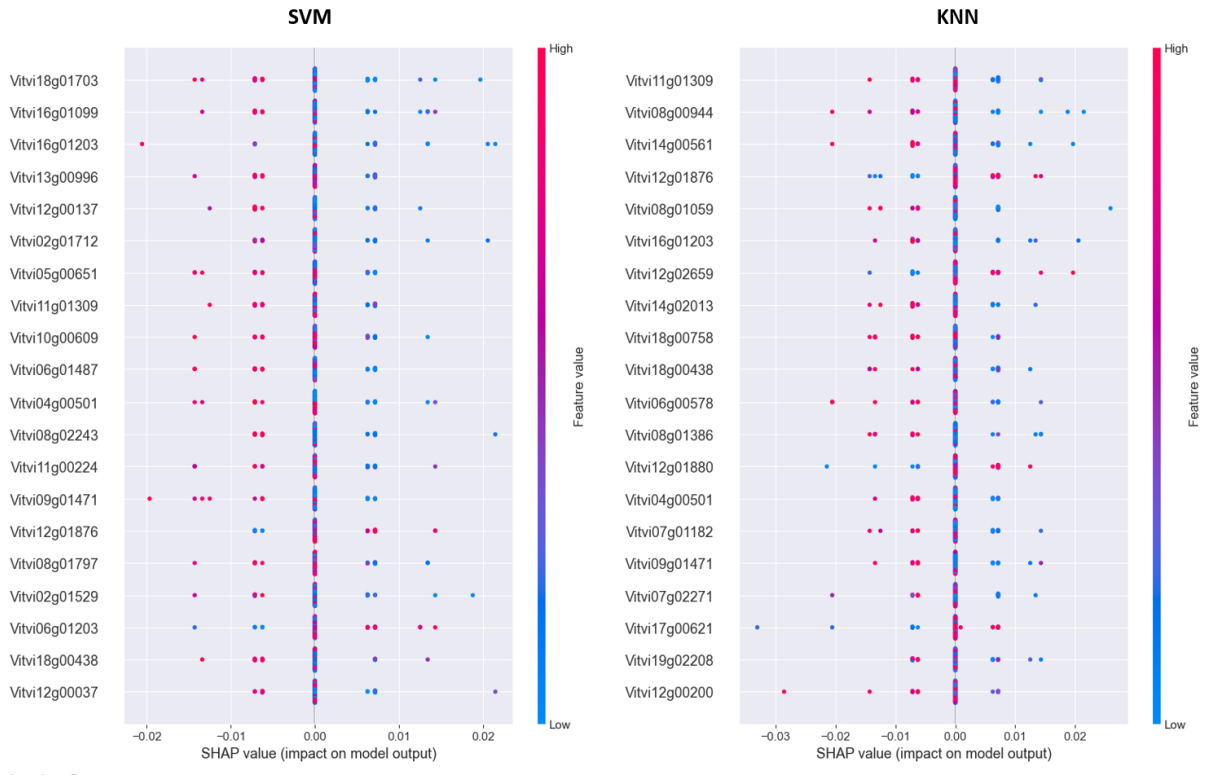


Figure 54: Beeswarm plot of SHAP values for the genes that contribute most to SVM's (left) and KNN's (right) predictions. Features are ordered from higher to lower effects on the predictions. The dots represent a single observation and the colour indicates if the observation has a higher (pink) or a lower (blue) feature value compared to the other observations.

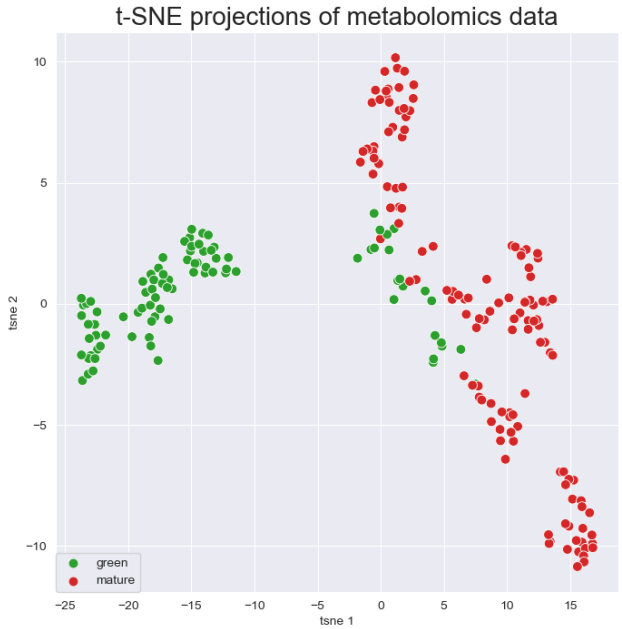


Figure 55: t-SNE visualisation of the metabolomics data of the 208 metabolites and considering replicates. Samples are coloured by grape developmental stage.

Table 32: Evaluation results of the ML models trained with metabolomics data for the metrics balanced accuracy, precision, recall, and F1 score.

	LR	KNN	DT	SVM	RF
balanced accuracy	0.94	0.92	0.91	0.93	0.96
precision	0.95	0.95	0.93	0.94	0.96
recall	0.97	0.92	0.93	0.97	1.00
F1 score	0.96	0.93	0.92	0.95	0.98

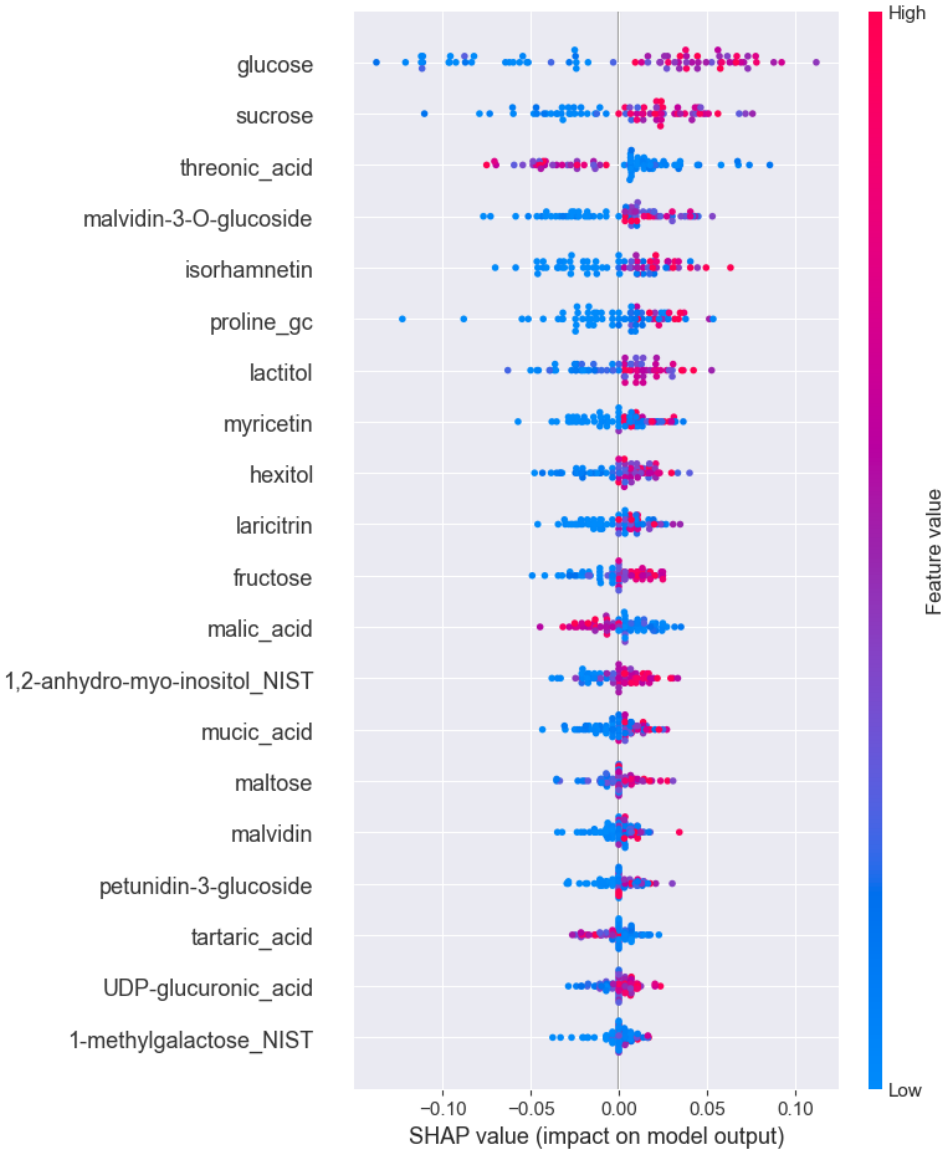


Figure 56: Beeswarm plot of SHAP values for the metabolites that contribute most to RF's predictions. Features are ordered from higher to lower effects on the predictions. The dots represent a single observation and the colour indicates if the observation has a higher (pink) or a lower (blue) feature value compared to the other observations.

the levels of organic acids decrease as the grape matures. However, the mucic acid presents high levels in most mature samples. The most contributing metabolites are mainly sugars and phenylpropanoids, such as malvidin-3-glucoside, which are expected to be more abundant in the mature grape. Proline is the only amino acid identified here and presents high levels in the mature samples. This was also described by Fasoli et. al [274]. The results obtained from metabolomics data are easier to interpret and are more in line with the differences between green and mature grapes described in the literature. Therefore, metabolomics data reflects the ultimate differences closer to the phenotypes that define the green and mature grapes and are described in the literature [274, 286, 299]. However, they do not give insights into the mechanisms that lead to these differences.

In conclusion, each omics dataset was valuable for classifying the grape developmental state, as each omics type led to unique insights that helped to characterise the green and mature states.

5.2 Multi-omics data integration

The three omics datasets used in the previous section, RNA-Seq, metabolomics, and predicted fluxomics, were then integrated to study the interactions between variables of different datasets (Figure 57).

Similar to the previous ML analysis, samples of Pinot Noir were also included in this analysis to increase the number of samples and samples were discretized into green and mature, which again represent the class to be predicted by the model. In summary, the RNA-Seq dataset included the expression of the 35336 genes, the metabolomics dataset comprised the measurements of 212 putative metabolites and the fluxomics dataset contained the flux capacity of 8632 reactions.

The multiview analyses were performed using R and the mixOmics package, which includes several multi-omics integration methods and plots that helped visualise the results [195]. The method chosen was DIABLO [194], which besides integrating omics also performs supervised classification. It requires the train and test dataset for each omics type as input as well as the output class for the train and test samples.

First, the RNA-Seq dataset was split into train and test, using stratified 5-fold cross-validation to ensure the robustness of the results. The features were selected by applying two filters to the training set, as performed in the previous section. The train and test sets of the metabolomics and fluxomics datasets were selected based on the division previously obtained for the RNA-Seq data. As the number of metabolic features is low, only the *VarianceThreshold* filter was applied to the metabolomics dataset. Therefore, the final datasets included 500 genes, 208 putative metabolites and 500 reactions for the same samples.

DIABLO also requires a design matrix as input defining the weights to assign to the correlations between omics datasets, which will establish a balance between maximising the correlation between datasets and maximising the model's ability to distinguish the outcome groups. Values higher than 0.5 will prioritise the correlation while lower values will prioritise predictions. Therefore, before running DIABLO, pairwise correlations between the features of different datasets were calculated using PLS, as recommended by the authors, to define the values of the design matrix.

After defining the design matrix, a draft DIABLO model was created to tune the number of components using 5-fold cross-validation repeated 10 times. An additional feature selection step was performed to obtain the ideal number of features of each dataset from a set of values passed as input (25, 50, and 100). Repeated cross-validation of 10 folds was used to run this function.

Finally, a DIABLO model was created for each data split, resulting in five models. These models were assessed with the training set, using repeated cross-validation 10 folds. The overall Balanced Error Rate (BER) was calculated for every distance metric, maximum, centroids, and Mahalanobis, and both components, using the weighted vote function, and averaged across data splits. The DIABLO models were then applied to the test dataset, for a more accurate evaluation. The test BERs were obtained for each distance metric and each component and averaged across all models.

The correlations between the variables of the different datasets were visualised for the best model using the diagnostic (*plotDiablo*) and Circos plots (*circosPlot*).

After analysing the correlations between variables, the variables with the highest impact on each component and in the classification task were identified by calculating the average loading weights of each feature on each component across all data splits.

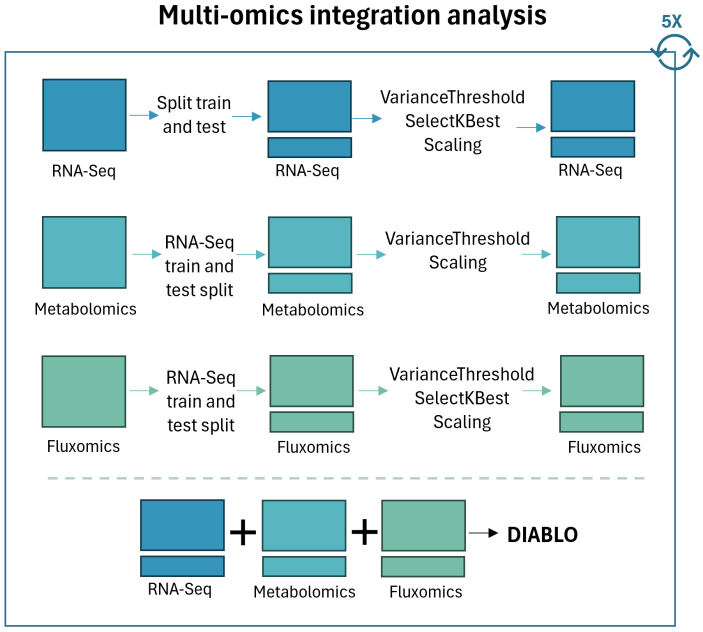


Figure 57: Multi-omics integration analyses performed. RNA-Seq data was split into train and test using stratified 5-fold cross-validation and these samples defined the train and test sets for metabolomics and fluxomics data. The omics datasets were preprocessed individually and integrated with DIABLO.

5.2.1 Results of multi-omics integration

Before integrating all omics datasets with DIABLO, pairwise correlations between the different datasets were calculated using PLS to help define the design matrix for DIABLO. The average correlation results are represented in Table 33.

Table 33: Average pairwise correlation values between the two components of the three datasets obtained using PLS.

	component 1	**component 2**
RNASeq + Metabolomics	0.952	0.837
RNASeq + Fluxomics	0.888	0.777
Metabolomics + Fluxomics	0.891	0.449

RNASeq and Metabolomics datasets present the highest correlation values for both components, 0.952 and 0.837, respectively. For component 1, the correlation is slightly higher between the metabolomics and fluxomics datasets than between RNASeq and fluxomics, but for component 2 the correlation between metabolomics and fluxomics is much lower, around 0.45. Hence, the design matrix was created with 0.5 as the correlation weight.

5.2.1.1 Model evaluation

The performance of the DIABLO models was assessed using only the training set and applying them to the test sets. The average BER was lower than 0.036 for all distance metrics and both components with the training set, using the weighted vote function (Table 34). The results were slightly better using the second components for classification.

Table 34: Average performance on the training set integrating all omics datasets. The overall BERs were calculated for maximum, centroids, and Mahalanobis distances for each component using the weighted vote function and averaged across all train-test splits.

	component 1	component 2
maximum distance	0.036	0.034
centroids distance	0.034	0.033
Mahalanobis distance	0.034	0.032

For the test sets, the average BER was lower than 0.028 for all distance metrics and both components using the weighted vote function (Table 35). Surprisingly, component 2 presents a lower error rate for maximum distance. The results were good as three out of the five DIABLO models correctly classified all test samples. The confusion matrices of the other two models were analysed to identify the misclassified samples (Tables 36 and 37). One model failed to classify a mature sample from Cabernet Sauvignon 2012 time point 5 and another failed to classify a green sample from Pinot Noir 2013 time point 4. These time points are closer to the transition between green and mature phases, meaning that the transcript and metabolic contents in these samples are probably similar to the samples in the other state, which may explain the incorrect classification by the model.

Therefore, the created models performed well in learning and predicting the results in the test set. However, the omics dataset used in this analysis is too small to create good models capable of capturing all real patterns related to grape development and correctly classifying new samples. Nevertheless, this analysis is useful for establishing relationships between variables and identifying possible biomarkers for each stage.

Table 35: Model performance on the test set integrating all omics. The overall BERs were calculated for maximum, centroids, and Mahalanobis distances for each component using the weighted vote function. The results are shown for each data split and were averaged across all train-test splits.

data split	distances	comp1	comp2
1	maximum distance	0.056	0.000
	centroids distance	0.056	0.056
	mahalanobis distance	0.056	0.056
2	maximum distance	0.000	0.000
	centroids distance	0.000	0.000
	mahalanobis distance	0.000	0.000
3	maximum distance	0.000	0.000
	centroids distance	0.000	0.000
	mahalanobis distance	0.000	0.000
4	maximum distance	0.000	0.000
	centroids distance	0.000	0.000
	mahalanobis distance	0.000	0.000
5	maximum distance	0.083	0.083
	centroids distance	0.083	0.083
	mahalanobis distance	0.083	0.083
mean	maximum distance	0.028	0.017
	centroids distance	0.028	0.028
	mahalanobis distance	0.028	0.028

Table 36: Confusion matrix of DIABLO model from data split 1.

		Predicted label	
True label		green	mature
	green	6	0
	mature	1	8

Table 37: Confusion matrix of DIABLO model from data split 5.

		Predicted label	
True label		green	mature
	green	5	1
	mature	0	8

5.2.1.2 Feature correlations

The correlations between features were analysed using the DIABLO model from data split 3 as this was one of the models with the best evaluation results. The feature selection for this model resulted in 25 genes, metabolites, and reactions for both components.

The diagnostic plot was created for each component to assess the correlation between datasets and the predictive capacity of the model (Figure 58). In the first component, RNA-Seq is highly correlated to metabolomics (0.98) and fluxomics (0.97), and the correlation between metabolomics and fluxomics is still high (0.94). All dataset combinations seem to be able to separate well the output classes, but with fluxomics the samples of the same class appear to be closer. In the second component, the most correlated datasets are RNA-Seq and metabolomics (0.87) and the correlation between metabolomics and fluxomics is higher than the correlation between RNA-Seq and fluxomics. These values are lower than the ones of the first component. In addition, the separation between output classes is not that evident, meaning that components 2 appear to have less predictive power than component 1, which was not verified in the model evaluation.

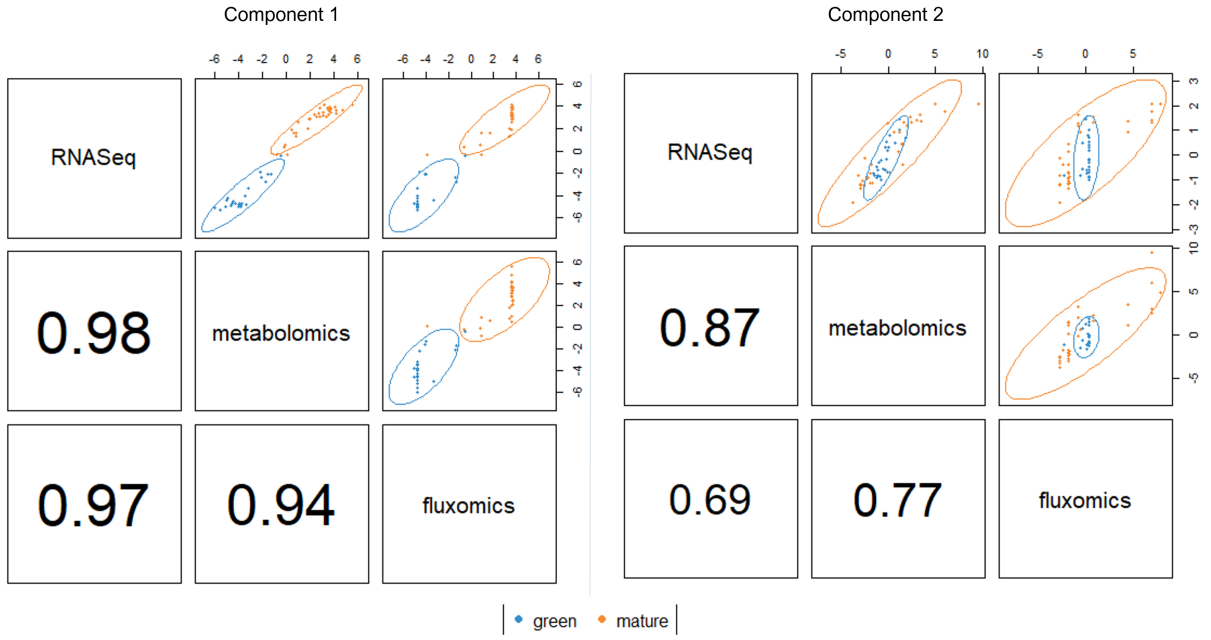


Figure 58: Diagnostic plot from DIABLO for each component of RNA-Seq, metabolomics, and fluxomics datasets. The values correspond to the correlation between datasets. Samples are coloured by grape developmental stage and 95% confidence ellipses are plotted.

The correlations between features from the different datasets were explored and visualised using the Circos plot (Figure 59) to identify interesting relationships between genes, reactions, and metabolites (Supplementary File 16). Only correlations higher than 0.9 are displayed.

As the number of features is high, feature names are not legible on the plot. However, it is possible to see that three metabolites, malate, sucrose and glucose, have absolute correlation values higher than 0.9

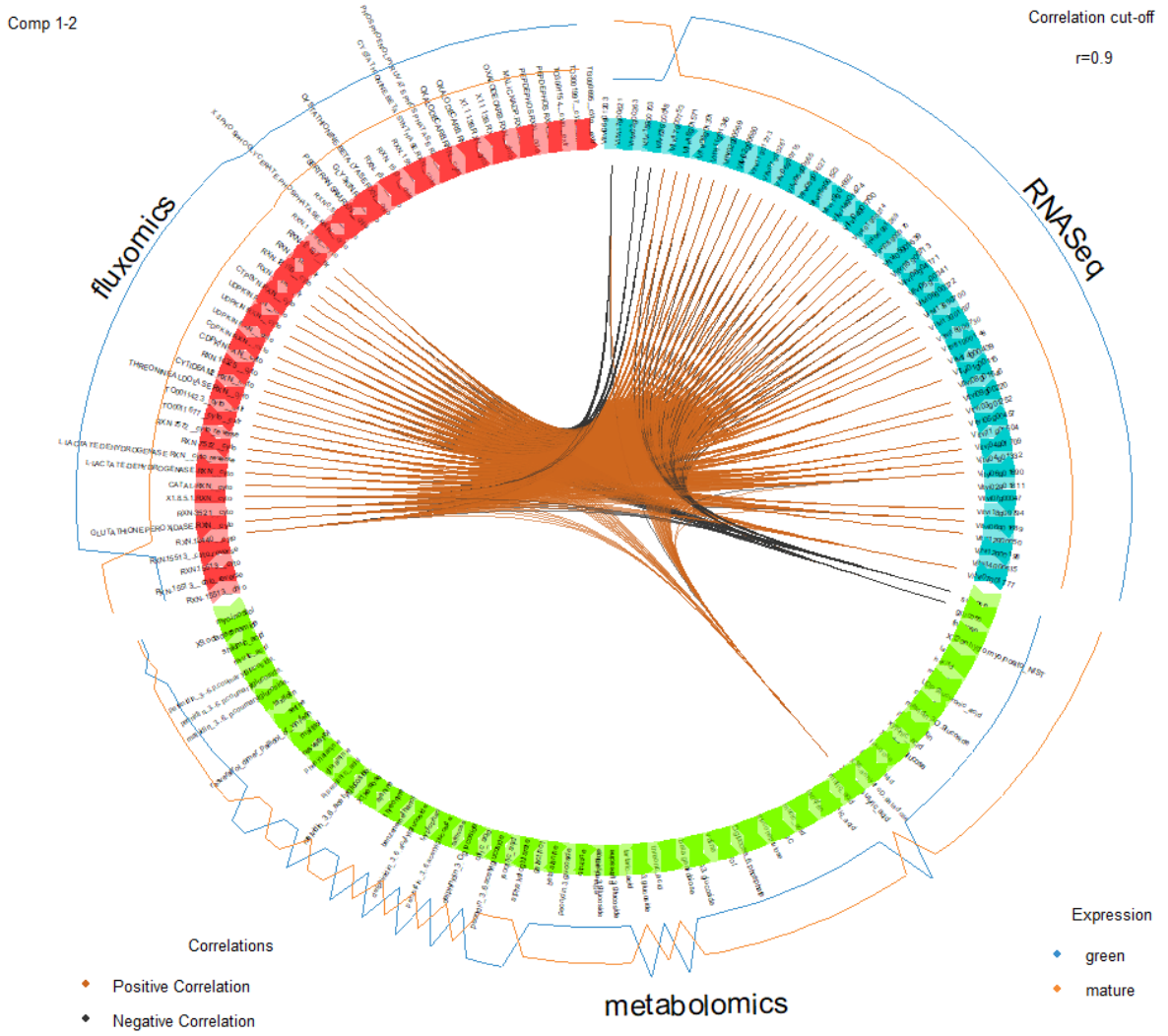


Figure 59: Circos plot from DIABLO model of data split 3 representing correlations greater than 0.9 between genes, metabolites, and reactions. Positive correlations are coloured in orange while negative ones are coloured in black. The value of each variable in the dataset in each output class is represented by the lines around the plot. The blue line is for green samples while the mature samples are in orange.

with most genes and some reactions. Black edges indicate that sucrose and glucose are negatively correlated with most genes and reactions, while orange edges indicate positive correlations between malate and those features. This also indicates that malate is negatively correlated to sucrose and glucose, which was expected as, during grape ripening, the grape concentration of sugars increases while decreasing organic acid concentration.

Malate is positively correlated to genes involved in photosynthesis and photorespiration pathways, primary sugar metabolism, alkaloid and arginine biosynthesis, protein modifications, phenolic degradation, nitrogen assimilation, transport, and cell development regulation. During maturation, these pathways are expected to be less active and the same happens to malate levels, which explains the high levels of correlation between malate and these genes. An exception is the biosynthesis of alkaloids that despite starting in the green phase, typically increases after veraison, which is not observed in malate levels.

Malate is also highly correlated to three genes with no known function, Vitvi11g01345, Vitvi12g00037, and Vitvi16g00931, suggesting that these genes are more active in the green stage. Malate is negatively correlated to only one gene (Vitvi12g02659) that codifies for seed storage proteins, which are expected to increase as the grape matures.

As expected, sucrose and glucose are negatively correlated to genes positively correlated with malate. Sucrose is negatively correlated to other genes involved in transcription and stress response regulation, ribosomal proteins, and metabolic pathways like photosynthesis, Calvin cycle, and the biosynthesis of plant hormones, carotenoids, and asparagine. Two additional genes with no known function (Vitvi08g01107 and Vitvi08g01114) presented high negative correlations with sucrose, suggesting that these genes are overexpressed in the green stage. In addition, sucrose is positively correlated to Vitvi01g00619 and Vitvi17g00621, which are related to the biosynthesis of terpenoids and fatty acid degradation, respectively.

Regarding the reactions, the same patterns are observed. Malate is positively correlated to the reactions CDPKIN-RXN, CTPSYN-RXN, RXN-14325, CYTIDEAM2-RXN, UDPKIN-RXN, RXN-12440, RXN-3521, 1.8.5.1-RXN, GLUTATHIONE-PEROXIDASE-RXN and CATAL-RXN, while sucrose and glucose are negatively correlated to the same reactions, meaning that they are more active in the green stage. These reactions are involved in nucleotide and ascorbate metabolism, and Reactive Oxygen Species (ROS) degradation, which are typically more active in the green stage and are also highly correlated with the same genes described above. Six additional reactions (RXN-969, GPH-RXN, RXN-961, PHOSPHORIBULOKINASE-RXN, THREONINE-ALDOLASE-RXN and GLYCOLATEDEHYDRO-RXN) were identified with absolute correlations higher than 0.9 to the described genes and not to the metabolites. These reactions participate in photorespiration, glycine biosynthesis, and glycolate degradation, which are also expected to be more active in the green stage. Hence, this confirms that the expression of these genes and the flux of these reactions have a similar pattern to malate levels during grape development.

5.2.1.3 Feature importance

The most contributing features of all datasets are shown in Figure 60 and are described in Supplementary File 17. The feature weights were averaged across all train-test splits.

The most contributing genes for component 1 are overexpressed in the green phase. These genes are involved in transport reactions (Vitvi18g01571, Vitvi04g00501, and Vitvi01g00052), photosynthesis (Vitvi08g01245 and Vitvi06g00513), sugar primary metabolism (Vitvi08g01301), ATP biosynthesis (Vitvi12g00284), alkaloid and arginine biosynthesis (Vitvi05g00565 and Vitvi18g00523, respectively), protein modifications (Vitvi17g00557 and Vitvi01g00081), and regulation of cell differentiation (Vitvi02g00666). Three genes have no known function (Vitvi18g02192, Vitvi05g01601, and Vitvi12g00037).

The most contributing metabolites for component 1 are organic acids and sugars. Sugars present higher levels in mature grapes, while organic acids are more abundant in the green phase, as expected. 1,2-anhydro-myo-inositol is a synthetic compound that is not produced by living organisms. Hence, its presence in the metabolomics dataset must be an error in metabolite identification. Malvidin 3-O glucoside

is the only secondary metabolite contributing to component 1.

The most contributing reactions for component 1 present higher flux in the green phase and participate in nucleotide biosynthesis, ROS degradation, histidine metabolism, and lipid transport.

For component 2, only one gene is overexpressed in the mature stage (Vitvi17g00621) and is involved in lipid degradation. The other genes are related to transcription regulation (Vitvi12g00200 and Vitvi18g00758), protein modification (Vitvi19g01952), antioxidant defence (Vitvi06g01349), ribosomal proteins (Vitvi02g01315 and Vitvi05g00053), photosynthesis (Vitvi13g00700), transport (Vitvi01g00647 and Vitvi01g00812), and carotenoid and asparagine metabolism (Vitvi11g00148 and Vitvi18g00730, respectively). Again, three genes identified have no known function (Vitvi03g01603, Vitvi08g01107, and Vitvi08g01114).

The most contributing metabolites for component 2 are more abundant in the mature stage, except the shikimic acid, which has higher levels in the green stage. Most metabolites are secondary metabolites, including flavonoids (taxifolin and myricetin) and anthocyanins (peonidin, petunidin, malvidin and delphinidin derivatives), which are described to be more abundant in mature grapes.

Finally, the most contributing reactions for component 2 are related to L-glutamate transport, glycolysis and gluconeogenesis. These pathways tend to be more active during the green phase of grape development as there is a higher energy demand for cell growth. Therefore, it is unclear why the reaction RXN-15513 (phosphoglycerate mutase) presents a higher flux in the mature phase.

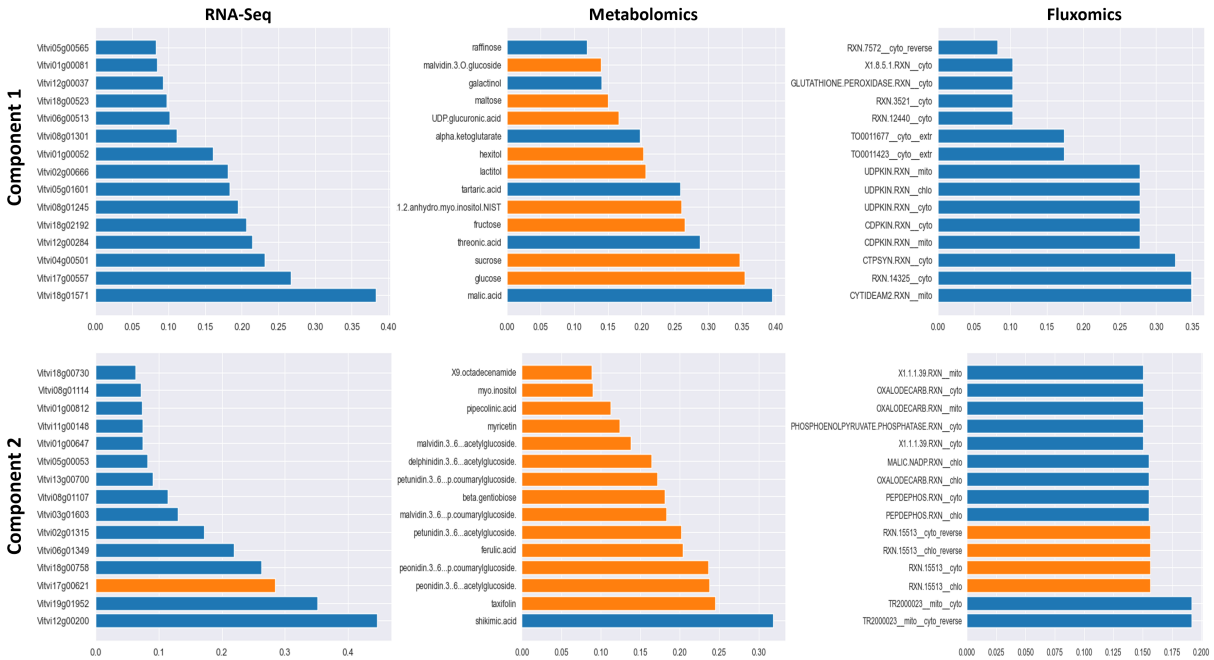


Figure 60: Loading weights for the 15 most important features from RNA-Seq, metabolomics, and fluxomics for both components. The feature weights were averaged across all train-test splits. The features are ordered from higher to lower impact from bottom to top. The colour indicates the class in which the feature has the maximum value (the green state is in blue while the mature is in orange)

5.3 Feature importance summary

The previous analyses led to the identification of several features that can be potential biomarkers of the grape developmental stage. These include genes, metabolites, and reactions identified by the analysis of omics datasets with ML and by the integrative analysis using DIABLO. Summary tables are available in Supplementary File 18.

From the described genes, only Vitvi04g00501 was identified as having a high impact on both ML models and DIABLO model. This gene codifies a putative lipid-binding protein and is overexpressed in the green state. Other genes highly impact both ML models but not the DIABLO model, which was the case of Vitvi16g01203, Vitvi09g01471, Vitvi18g00438, Vitvi12g01876, and Vitvi11g01309. In addition, Vitvi17g00621, Vitvi18g00758, Vitvi12g00200, and Vitvi12g00037 highly impact only one ML model and the DIABLO model. These genes are involved in fatty acid degradation, transcription regulation, and seed storage proteins. Vitvi12g00037 and Vitvi09g01471 have no known function. Vitvi17g00621 and Vitvi12g01876 are the only genes overexpressed in the mature phase. Hence, these genes could be potential biomarkers of each state. However, further analysis is needed to understand why these genes are important for distinguishing green and mature stages.

Regarding metabolites, 13 metabolites were considered to have a high impact on the predictions from ML and DIABLO models. These include sugars, such as glucose and sucrose, organic acids, like threonic and malic acid, and two secondary metabolites, malvidin 3-O-glucoside and myricetin. The metabolite misidentified as 12-anhydro-myo-inositol also greatly impacted the predictions of both models. These metabolites were expected to be important in classifying the grape developmental stage.

The fluxomics analysis did not identify reactions associated with the previous metabolites or genes described. Two glycerol transporters, TO0011677 and TO0011423, from the cytosol to the extracellular space were identified as highly contributing to both ML models and the DIABLO model. The reactions related to the MEP pathway highly impacted both ML models but not the DIABLO model. Lastly, the reactions involved in nucleotide biosynthesis and ROS degradation highly impacted one ML model and the DIABLO model. All reactions presented a mean flux value higher in the green phase. However, these reactions were not the most expected as the main differences between the green and mature states described in the literature are more related to sugars, organic acids, and secondary metabolites.

Thus, the genes, metabolites, and reactions identified here could represent potential biomarkers of the grape's developmental state, and the relationships between these features could help elucidate the underlying mechanisms leading to different grape metabolic compositions. Despite the high correlations between these features, the reactions identified do not directly produce or consume these metabolites nor are they directly linked to the identified genes, which hampers the interpretation of these results. Nevertheless, these models can be the starting point for exploring the unknown mechanisms occurring during grape development.

6

Conclusion

6.1 Main contributions

The present work has developed and applied a set of computational tools to model and study *V. vinifera* metabolism. The main outcomes are:

- the *iplants* repository with plant metabolic data to assist GSMM reconstruction;
- a novel generic GSMM of *V. vinifera* (iMS7199);
- *V. vinifera* tissue-specific models, for the leaf, stem and berry tissues;
- *V. vinifera* context-specific models for green and mature berries;
- a diel multi-tissue GSMM of *V. vinifera* including the metabolism of leaf, stem and berry tissues across the day-night cycle;
- the application of ML for analysing fluxomics data predicted by the GSMMs;
- the multi-omics integration analysis including experimental omics, RNASeq and metabolomics, and generated fluxomics for identifying biomarkers of each developmental stage of grapes.

Despite the variety of frameworks to reconstruct GSMMs, these are difficult to adapt for novel analyses and more complex organisms, like plants, making this task challenging and time-consuming. With this in mind, *iplants*, a repository of plant metabolic data was built in this work, integrating data from MetaCyc and PlantCyc databases, as well as UniProt and NCBI sequence data and the metabolic information of available plant GSMMs. This repository was used to assist in reconstructing the *V. vinifera* GSMM. An API was created to facilitate access to the repository and the retrieval of the necessary data for the reconstruction. This repository can also be used in the future to support the reconstruction of other plant GSMMs.

The main outcome of this work is the iMS7199 model, the first GSMM of *V. vinifera*. This model is based on the latest *V. vinifera* genomic and other omics data, as well as literature, including primary and secondary metabolic pathways, mainly related to flavonoids and hormone biosynthesis. The model can simulate grapevine metabolism under photosynthesis, photorespiration, and respiration. RNA-Seq data was integrated with this generic model to build tissue-specific models for the leaf, stem, green berry, and mature berry of *V. vinifera* Cabernet Sauvignon cultivar.

Multi-tissue models were built by connecting the tissue-specific models, and the diel cycle was introduced in the models by replicating the multi-tissue model for both light and dark phases. Two diel multi-tissue GSMMs were built, one using the green berry tissue and the other using the mature berry tissue, to capture the changes in grape metabolism during development. The models were used to assess the metabolic responses of grapevine to different levels of sulfate and nitrate. The results indicated that less biomass is produced with low nitrate or low sulfate, and more flux is expected in the secondary pathways. Hence, controlling the soil levels of nitrate and sulfate in specific stages of development can help control the phenolic and sugar content in the grapes, which will affect their quality. The metabolic models developed here can be a valuable tool for analysing and predicting the metabolic behaviour of grapevine under different environmental conditions and assessing its metabolic potential and fruit quality, which can be important for wine production.

A first effort of combining ML and GSMMs was also accomplished in this work. Fluxomics data were generated from GSMMs of green and mature grapes and analysed using ML models. The resulting models obtained very good results in predicting the grape developmental stage, with accuracy, precision, recall, and F1 scores higher than 90%. The reactions that contributed the most to the model's predictions are associated with different pathways, including MEP, threonine, and nucleotide metabolism, and presented higher fluxes in the green state. Although these pathways are not the main differences between green and mature grapes described in the literature, the results suggest that their fluxes are significantly different between the two states. In addition, DIABLO, a multi-view method, was used to integrate the phenotype predictions generated by the berry models with RNA-Seq and metabolomics data to identify connections between genes, reactions, and metabolites that can help discover biomarkers of the grape developmental state. These strategies can be applied to study other aspects of metabolism, for instance, drought and disease resistance, when new omics datasets become available.

A deeper understanding of plant metabolic pathways is essential to develop more robust GSMMs. Additionally, the creation of larger omics datasets is crucial for developing more realistic ML models, enabling more advanced analyses such as the identification of biomarkers for disease or environmental stress resistance. This approach not only represents a novel and pioneering effort in integrating omics, GSMMs, and ML in plant metabolism studies, but also showcases the significant potential of applying this strategy for more insightful analyses, as additional data becomes available.

6.2 Publications

The present thesis has contributed to a few publications. The following papers have been submitted and accepted during the development of this work, some of which developed with the collaboration of other researchers in the host research group:

- Sampaio, M., Rocha, M., and Dias, O. (2022). Exploring synergies between plant metabolic modelling and machine learning. Comput. Struct. Biotechnol. J. 20: 1885–1900.

The following has been submitted (available as a pre-print) and is waiting for revision:

- Sampaio, M., Rocha, M., and Dias, O. A diel multi-tissue genome-scale metabolic model of *Vitis vinifera*. Submitted. Pre-print available at: https://doi.org/10.1101/2024.01.30.578056

The following articles were submitted to conferences and are waiting for revision:

- Sampaio, M., Rocha, M., and Dias, O. A multi-omics approach for understanding grape metabolism throughout development. Submitted to the 10th IFAC International Conference on Foundations of Systems Biology in Engineering.
- Sampaio, M., Rocha, M., and Dias, O. *iplants*: a repository of plant metabolic data. Submitted to the 18th International Conference on Practical Applications of Computational Biology and Bioinformatics.

The following articles were also published during the doctoral program but were not included in this thesis:

- Oliveira, A., Cunha, E., Cruz, F., Ribeiro, J., Sequeira, J. C., Sampaio, M., Dias, O. (2022). Towards a Multivariate Analysis of Genome-Scale Metabolic Models Derived from the BiGG Models Database. In: Rocha, M., Fdez-Riverola, F., Mohamad, M.S., Casado-Vara, R. (eds) Practical Applications of Computational Biology and Bioinformatics, 15th International Conference (PACBB 2021). PACBB 2021. Lecture Notes in Networks and Systems, vol 325. Springer, Cham.
- Oliveira, A., Cunha, E., Cruz, F., Ribeiro, J., Sequeira, J. C., Sampaio, M., Sampaio, C., Dias, O. (2022). Systematic assessment of template-based genome-scale metabolic models created with the BiGG Integration Tool. Journal of Integrative Bioinformatics, Volume 19, Number 3, Pages 20220014.

6.3 Future work

The development of the tools mentioned above also raised several future improvements. The data in *iplants* repository could be further curated, for instance, many reactions are not mass balanced and this is required for model reconstruction. Also, it would be interesting to add the sequence data of the genes in plant GSMMs to improve genome annotation. The *iplants* repository could also benefit from a graphical user interface, to facilitate its access by scientists with no programming skills. The ideal goal would be to integrate the repository into a user-friendly framework for reconstructing plant GSMMs. This will require the management of users and their permissions. The functions for reconstructing a draft model from the genome sequence and *iplants* were already implemented in this work.

The reconstructed models of *V. vinifera* can also be improved in several ways. First, the use of experimental data to define the biomass equation and the pseudo reactions for all biomass components will greatly improve the model and lead to more realistic phenotype predictions. The annotation of the *V. vinifera* genome can also be improved by manual curation as only the first hit was included in the model.

Several enzymes of secondary pathways had a match in the genome, but not as the first hit. Hence, the annotation results should be further analysed to assess which new enzymes can be added to the model. In addition, several reactions in the model are still blocked. A more extensive gap-filling is required to fill the gaps and allow flux in these reactions. On the other hand, the blocked reactions should also be analysed to check whether they should stay in the model or be removed. As information about transport reactions is scarce, most transport reactions in the model lack GPR rules and many transporters must be missing, which can be the reason for the blockage of some metabolic reactions. Hence, manual curation of transport reactions would improve the connectivity of the network and the quality of the model. The phenotype predictions of the reconstructed models should also be compared with experimental data to detect and fix the errors in the model and improve the results.

The combination of GSMMs, ML, and omics data can be greatly improved when larger omics datasets become available. Also, other ML models and multi-view algorithms can be explored for multi-omics data integration. This strategy has diverse and interesting potential applications, such as studying grapevine resistance to climate change and diseases, which is economically relevant. The integration of multi-omics data and the combination of ML and metabolic modelling approaches can be used to study the complex behaviour of plants by exploring the interactions between genes, reactions, and metabolites, which can elucidate the mechanisms leading to different plant phenotypes.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. url: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf.

[2] C. T. Argueso et al. "Directions for research and training in plant omics: Big Questions and Big Data". In: *Plant Direct* 3.4 (2019), pp. 1–16. issn: 24754455. doi: 10.1002/pld3.133.

[3] International Organisation of Vine and Wine. "STATE OF THE WORLD VINE AND WINE SECTOR IN 2022". In: (2023).

[4] OEC - The Observatory of Economic Complexity. *Portugal (PRT) Exports, Imports, and Trade Partners*. 2024. url: https://oec.world/en/profile/country/prt/ (visited on 2024-04-22).

[5] C. Gu et al. "Current status and applications of genome-scale metabolic models". In: *Genome Biology* 20.1 (2019), pp. 1–18. issn: 1474760X. doi: 10.1186/s13059-019-1730-3.

[6] B. B. Misra et al. "Integrated omics: Tools, advances and future approaches". In: *Journal of Molecular Endocrinology* 62.1 (2019), R21–R45. issn: 14796813. doi: 10.1530/JME-18-0055.

[7] G. Zampieri et al. "Machine and deep learning meet genome-scale metabolic modeling". In: *PLoS Computational Biology* 15.7 (2019), pp. 1–24. issn: 15537358. doi: 10.1371/journal.pcbi.1007084.

[8] P. Rana et al. "Recent advances on constraint-based models by integrating machine learning". In: *Current Opinion in Biotechnology* 64 (2020), pp. 85–91. issn: 18790429. doi: 10.1016/j.copbio.2019.11.007. url: https://doi.org/10.1016/j.copbio.2019.11.007.

[9] A. Antonakoudis et al. "The era of big data: Genome-scale modelling meets machine learning". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 3287–3300. issn: 20010370. doi: 10.1016/j.csbj.2020.10.011. url: https://doi.org/10.1016/j.csbj.2020.10.011.

[10] Y. Kim, G. B. Kim, and S. Y. Lee. "Machine learning applications in genome-scale metabolic modeling". In: *Current Opinion in Systems Biology* 25 (2021), pp. 42–49. issn: 2452-3100. doi: https://doi.org/10.1016/j.coisb.2021.03.001. url: https://www.sciencedirect.com/science/article/pii/S2452310021000032.

[11] M. Sampaio, M. Rocha, and O. Dias. “Exploring synergies between plant metabolic modelling and machine learning”. In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 1885–1900. issn: 2001-0370. doi: https://doi.org/10.1016/j.csbj.2022.04.016. url: https://www.sciencedirect.com/science/article/pii/S2001037022001301.

[12] J. Stenesh and J. Stenesh. “Introduction to Metabolism”. In: *Biochemistry*. Springer US, 1998, pp. 203–219. doi: 10.1007/978-1-4757-9427-4_8.

[13] H. Kornberg. *metabolism | Definition, Process, & Biology*. url: https://www.britannica.com/science/metabolism (visited on 2020-04-06).

[14] R. Verpoort. “Plant secondary metabolism”. In: *Metabolic Engineering of Plant Secondary Metabolism*. Kluwer Academic Publishers, 2000. isbn: 0792363604. doi: 10.1002/0470869143.kc029.

[15] K. Baghalian, M. R. Hajirezaei, and F. Schreiber. “Plant metabolic modeling: Achieving new insight into metabolism and metabolic engineering”. In: *Plant Cell* 26.10 (2014), pp. 3847–3866. issn: 1532298X. doi: 10.1105/tpc.114.130328.

[16] C. Fang, A. R. Fernie, and J. Luo. “Exploring the Diversity of Plant Metabolism”. In: *Trends in Plant Science* 24.1 (2019), pp. 83–98. issn: 13601385. doi: 10.1016/j.tplants.2018.09.006. url: https://doi.org/10.1016/j.tplants.2018.09.006.

[17] M. Wink. “Introduction: Biochemistry, Physiology and Ecological Functions of Secondary Metabolites”. In: *Biochemistry of Plant Secondary Metabolism: Second Edition* 40 (2010), pp. 1–19. doi: 10.1002/9781444320503.ch1.

[18] A. M. Feist et al. “Reconstruction of Biochemical Networks in Microbial Organisms”. In: *Nature Reviews Microbiology* (2008). doi: 10.1038/nrmicro1949.

[19] J. S. Edwards and B. O. Palsson. “Systems properties of the Haemophilus influenzae Rd metabolic genotype”. In: *Journal of Biological Chemistry* 274.25 (1999-06), pp. 17410–17416. issn: 00219258. doi: 10.1074/jbc.274.25.17410.

[20] I. Thiele and B. Ø. Palsson. “A protocol for generating a high-quality genome-scale metabolic reconstruction”. In: *Nature protocols* (2010). doi: 10.1038/nprot.2009.203.

[21] M. Hucka et al. “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.” In: *Bioinformatics (Oxford, England)* 19.4 (2003-03), pp. 524–31. issn: 1367-4803.

[22] O. Dias et al. “Reconstructing genome-scale metabolic models with merlin”. In: *Nucleic Acids Research* 43.8 (2015), pp. 3899–3910. issn: 13624962. doi: 10.1093/nar/gkv294.

[23] D. Machado et al. “Fast automated reconstruction of genome-scale metabolic models for microbial species and communities”. In: *Nucleic Acids Research* 46.15 (2018-09), pp. 7542–7553. issn: 0305-1048. doi: 10.1093/nar/gky537. url: https://academic.oup.com/nar/article/46/15/7542/5042022.

[24] J. Schellenberger et al. “Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.” In: *Nature protocols* 6.9 (2011), pp. 1290–1307. issn: 1750-2799. doi: 10.1038/nprot.2011.308.

[25] C. S. Henry et al. “High-throughput generation, optimization and analysis of genome-scale metabolic models”. In: *Nature Biotechnology* 28.9 (2010-09), pp. 977–982. issn: 1087-0156. doi: 10.1038/nbt.1672. url: http://www.ncbi.nlm.nih.gov/pubmed/20802497%20http://www.nature.com/articles/nbt.1672.

[26] A. P. Arkin et al. “The DOE Systems Biology Knowledgebase (KBase)”. In: *bioRxiv* (2016-12), p. 096354. doi: 10.1101/096354. url: https://www.biorxiv.org/content/early/2016/12/22/096354.

[27] M. Kanehisa et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Research* 45.D1 (2017-01), pp. D353–D361. doi: 10.1093/nar/gkw1092. url: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1092.

[28] R. Caspi et al. “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”. In: *Nucleic Acids Research* 42.D1 (2014-01), pp. D459–D471. issn: 0305-1048. doi: 10.1093/nar/gkt1103. url: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1103.

[29] E. W. Sayers et al. “Database resources of the National Center for Biotechnology Information”. In: *Nucleic Acids Research* 38.suppl_1 (2010-01), pp. D5–D16. issn: 0305-1048. doi: 10.1093/nar/gkp967. url: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp967.

[30] A. Bateman et al. “UniProt: The universal protein knowledgebase”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D158–D169. issn: 13624962. doi: 10.1093/nar/gkw1099.

[31] S. Placzek et al. “BRENDA in 2017: New perspectives and new tools in BRENDA”. In: *Nucleic Acids Research* 45.D1 (2017-01), pp. D380–D388. issn: 13624962. doi: 10.1093/nar/gkw952.

[32] M. H. Saier et al. “The Transporter Classification Database (TCDB): recent advances”. In: *Nucleic Acids Research* 44.D1 (2016-01), pp. D372–D379. issn: 0305-1048. doi: 10.1093/nar/gkv1103. url: http://www.ncbi.nlm.nih.gov/pubmed/26546518%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702804%20https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1103.

[33] S. Kim et al. “PubChem Substance and Compound databases”. In: *Nucleic Acids Research* 44 (2016). doi: 10.1093/nar/gkv951. url: https://pubchem.ncbi.nlm.nih.gov/vw3d/.

[34] P. Zhang et al. “Creation of a genome-wide metabolic pathway database for populus trichocarpa using a new approach for reconstruction and curation of metabolic pathways for plants”. In: *Plant Physiology* 153.4 (2010-08), pp. 1479–1491. issn: 00320889. doi: 10.1104/pp.110.157396.

[35] S. Naithani et al. “Plant Reactome: a knowledgebase and resource for comparative pathway analysis”. In: *Nucleic Acids Research* 48 (2019), pp. 1093–1103. doi: 10.1093/nar/gkz996. url: https://plantreactome.gramene..

[36] E. Grafahrend-Belau et al. “MetaCrop: a detailed database of crop plant metabolism”. In: *Nucleic Acids Research* 36 (2008). doi: 10.1093/nar/gkm835. url: http://www.arabidopsis.org/.

[37] L. A. Mueller et al. “The SOL Genomics Network. A comparative resource for Solanaceae biology and beyond”. In: *Plant Physiology* 138.3 (2005-07), pp. 1310–1317. issn: 00320889. doi: 10.1104/pp.105.060707.

[38] T. Z. Berardini et al. “The arabidopsis information resource: Making and mining the ”gold standard” annotated reference plant genome”. In: *Genesis* 53.8 (2015-08), pp. 474–485. issn: 1526968X. doi: 10.1002/dvg.22877.

[39] S. F. Altschul et al. “Basic local alignment search tool”. In: *Journal of Molecular Biology* 215.3 (1990-10), pp. 403–410. issn: 00222836. doi: 10.1016/S0022-2836(05)80360-2.

[40] B. Buchfink, C. Xie, and D. H. Huson. “Fast and sensitive protein alignment using diamond”. In: *Nature Methods* 12.1 (2014), pp. 59–60. doi: 10.1038/nmeth.3176.

[41] S. R. Eddy. “Multiple alignment using hidden Markov model”. In: *Intelligent Systems for Molecular Biology* 95 (1995), pp. 114–120. issn: 1553-0833. doi: 7584426.

[42] P. Gupta et al. “Gramene database: Navigating plant comparative genomics resources”. In: *Current Plant Biology* 7-8 (2016-11), pp. 10–15. issn: 22146628. doi: 10.1016/j.cpb.2016.12.005.

[43] L. J. Sweetlove and R. George Ratcliffe. “Flux-balance modeling of plant metabolism”. In: *Frontiers in Plant Science* 2.AUG (2011), pp. 1–10. issn: 1664462X. doi: 10.3389/fpls.2011.00038.

[44] E. Collakova, J. Y. Yen, and R. S. Senger. “Are we ready for genome-scale modeling in plants?” In: *Plant Science* 191-192 (2012), pp. 53–70. issn: 01689452. doi: 10.1016/j.plantsci.2012.04.010. url: http://dx.doi.org/10.1016/j.plantsci.2012.04.010.

[45] O. Emanuelsson et al. “Predicting subcellular localization of proteins based on their N-terminal amino acid sequence”. In: *Journal of Molecular Biology* 300.4 (2000), pp. 1005–1016. issn: 00222836. doi: 10.1006/jmbi.2000.3903.

[46] P. Horton et al. “WoLF PSORT: Protein localization predictor”. In: *Nucleic Acids Research* 35 (2007), pp. 585–587. issn: 03051048. doi: 10.1093/nar/gkm259.

[47] T. Goldberg et al. “LocTree3 prediction of localization”. In: *Nucleic Acids Research* 42.W1 (2014), pp. 350–355. issn: 13624962. doi: 10.1093/nar/gku396.

[48] J. L. Heazlewood et al. “SUBA: the Arabidopsis Subcellular Database”. In: *Nucleic Acids Research* (2006).

[49] S. Reumann et al. “AraPerox. A database of putative arabidopsis proteins from plant peroxisomes”. In: *Plant Physiology* 136.1 (2004-09), pp. 2587–2608. issn: 00320889. doi: 10.1104/pp.104.043695.

[50] I. Rocha, J. Förster, and J. Nielsen. "Design and application of genome-scale reconstructed metabolic models". In: *Methods in Molecular Biology* 416 (2007-12), pp. 409–431. issn: 10643745. doi: 10.1007/978-1-59745-321-9_29.

[51] J. D. Orth, I. Thiele, and B. O. Palsson. "What is flux balance analysis?" In: *Nature Publishing Group* 28.3 (2010), pp. 245–248. issn: 1087-0156. doi: 10.1038/nbt.1614. url: http://dx.doi.org/10.1038/nbt.1614.

[52] A. Varma and B. O. Palsson. "Metabolic flux balancing: Basic concepts, scientific and practical use". In: *Bio/Technology* 12.10 (1994), pp. 994–998. issn: 0733222X. doi: 10.1038/nbt1094-994.

[53] A. P. Burgard, P. Pharkya, and C. D. Maranas. "OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization". In: *Biotechnology and Bioengineering* 84.6 (2003-12), pp. 647–657. issn: 00063592. doi: 10.1002/bit.10803.

[54] I. Rocha et al. "OptFlux: an open-source software platform for in silico metabolic engineering." In: *BMC systems biology* 4.1 (2010), p. 45. issn: 1752-0509. doi: 10.1186/1752-0509-4-45. url: http://www.biomedcentral.com/1752-0509/4/45.

[55] N. E. Lewis et al. "Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models." In: *Molecular systems biology* 6.1 (2010-07), p. 390. issn: 1744-4292. doi: 10.1038/msb.2010.47. url: http://www.ncbi.nlm.nih.gov/pubmed/20664636%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2925526.

[56] R. Mahadevan and C. H. Schilling. "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models". In: *Metabolic Engineering* 5.4 (2003), pp. 264–276. issn: 10967176. doi: 10.1016/j.ymben.2003.09.002.

[57] D. Segrè, D. Vitkup, and G. M. Church. "Analysis of optimality in natural and perturbed metabolic networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.23 (2002-12), pp. 15112–7. issn: 0027-8424. doi: 10.1073/pnas.232349399.

[58] T. Shlomi, O. Berkman, and E. Ruppin. "Regulatory on/off minimization of metabolic flux changes after genetic perturbations." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.21 (2005-05), pp. 7695–700. issn: 0027-8424. doi: 10.1073/pnas.0406346102. url: http://www.ncbi.nlm.nih.gov/pubmed/15897462%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1140402.

[59] A. Ebrahim et al. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". In: *BMC Systems Biology* 7.1 (2013-08), p. 74. issn: 1752-0509. doi: 10.1186/1752-0509-7-74. url: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-74.

[60] D. Machado. *GitHub - ReFramed: metabolic modeling package*. url: https://github.com/cdanielmachado/reframed (visited on 2020-05-18).

[61] H. A. Herrmann et al. “Flux sampling is a powerful tool to study metabolism under changing environmental conditions”. In: *npj Systems Biology and Applications* 5.1 (2019-12), pp. 1–8. issn: 20567189. doi: 10.1038/s41540-019-0109-0.

[62] J. Schellenberger and B. Palsson. *Use of randomized sampling for analysis of metabolic networks*. 2009-02. doi: 10.1074/jbc.R800048200.

[63] R. Mahadevan, J. S. Edwards, and F. J. Doyle. “Dynamic Flux Balance Analysis of diauxic growth in Escherichia coli”. In: *Biophysical Journal* 83.3 (2002), pp. 1331–1340. issn: 00063495. doi: 10.1016/S0006-3495(02)73903-9.

[64] M. Kim and I. Tagkopoulos. *Data integration and predictive modeling methods for multi-omics datasets*. 2018. doi: 10.1039/c7mo00051k. url: https://pubmed.ncbi.nlm.nih.gov/29725673/.

[65] W. M. Aizat, I. Ismail, and N. M. Noor. “Recent development in omics studies”. In: *Advances in Experimental Medicine and Biology*. Vol. 1102. Springer New York LLC, 2018, pp. 1–9. doi: 10.1007/978-3-319-98758-3_1. url: https://pubmed.ncbi.nlm.nih.gov/30382565/.

[66] D. R. Hyduke, N. E. Lewis, and B. O. Palsson. “Analysis of omics data with genome-scale models of metabolism”. In: *Molecular BioSystems* 9.2 (2013), pp. 167–174. issn: 1742206X. doi: 10.1039/c2mb25453k.

[67] R. Leinonen, H. Sugawara, and M. Shumway. “The Sequence Read Archive”. In: *Nucleic Acids Research* (2010). doi: 10.1093/nar/gkq1019. url: http://www.ncbi.nlm.nih.gov/gap.

[68] D. A. Benson et al. “GenBank”. In: *Nucleic Acids Research* (2012). doi: 10.1093/nar/gks1195. url: www.ncbi.nlm.nih.gov/guide/.

[69] K. D. Pruitt, T. Tatusova, and D. R. Maglott. “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins”. In: *Nucleic Acids Research* (2007). doi: 10.1093/nar/gkl842. url: http://www.ncbi.nlm.nih.gov/RefSeq/.

[70] NCBI. *Nucleotide*. url: https://www.ncbi.nlm.nih.gov/nucleotide/ (visited on 2020-06-08).

[71] E. Clough and T. Barrett. “The Gene Expression Omnibus database”. In: *Methods in Molecular Biology*. Vol. 1418. Humana Press Inc., 2016, pp. 93–110. doi: 10.1007/978-1-4939-3578-9_5.

[72] J. Mashima et al. “DNA Data Bank of Japan”. In: *Nucleic Acids Research* 45 (2016), pp. 25–31. doi: 10.1093/nar/gkw1001. url: http://trace.ddbj.nig..

[73] C. Amid et al. “The European Nucleotide Archive in 2019”. In: *Nucleic Acids Research* 48 (2020). doi: 10.1093/nar/gkz1063. url: https://ena-docs.readthedocs.io/en/latest/submit/assembly/.

[74] H. Parkinson et al. “ArrayExpress-a public database of microarray experiments and gene expression profiles”. In: *Nucleic Acids Research* (2007). doi: 10.1093/nar/gkl995.

[75] I. Papatheodorou et al. “Expression Atlas: gene and protein expression across multiple studies and organisms”. In: *Nucleic Acids Research* 46 (2018). doi: 10.1093/nar/gkx1158. url: http://www.ebi.ac.uk/gxa/experiments/E-MTAB-4840.

[76] H. Ohyanagi et al. “Plant omics data center: An integrated web repository for interspecies gene expression networks with NLP-based curation”. In: *Plant and Cell Physiology* 56.1 (2015), e9. issn: 14719053. doi: 10.1093/pcp/pcu188.

[77] T. Kudo et al. “PlantExpress: A Database Integrating OryzaExpress and ArthaExpress for Single-species and Cross-species Gene Expression Network Analyses With Microarray-Based Transcriptome Data”. In: *Plant & cell physiology* 58.1 (2017). issn: 1471-9053. doi: 10.1093/PCP/PCW208.

[78] A. Velt and C. Rustenholz. *GREAT$_PN40024.v4.2_beta$*. 2023. url: https://great.colmar.inrae.fr/app/GREAT_PN40024.v4.2_beta.

[79] A. Velt et al. “An improved reference of the grapevine genome reasserts the origin of the PN40024 highly homozygous genotype”. In: *G3 Genes|Genomes|Genetics* 13.5 (2023-03). jkad067. issn: 2160-1836. doi: 10.1093/g3journal/jkad067. eprint: https://academic.oup.com/g3journal/article-pdf/13/5/jkad067/50151484/jkad067.pdf. url: https://doi.org/10.1093/g3journal/jkad067.

[80] P. Samaras et al. “ProteomicsDB: a multi-omics and multi-organism resource for life science research”. In: *Nucleic Acids Research* 48 (2019), pp. 1153–1163. doi: 10.1093/nar/gkz974. url: https://www.proteomicsdb.org.

[81] Y. Perez-Riverol et al. “The PRIDE database and related tools and resources in 2019: improving support for quantification data”. In: *Nucleic Acids Research* 47 (2019). doi: 10.1093/nar/gky1106. url: http://www.iprox.org/.

[82] E. W. Deutsch. “The PeptideAtlas Project”. In: *Methods in molecular biology (Clifton, N.J.)* Vol. 604. NIH Public Access, 2010, pp. 285–296. doi: 10.1007/978-1-60761-444-9_19. url: http://link.springer.com/10.1007/978-1-60761-444-9%7B%5C_%7D19.

[83] R. Craig, J. P. Cortens, and R. C. Beavis. “Open source system for analyzing, validating, and storing protein identification data”. In: *Journal of Proteome Research* 3.6 (2004-11), pp. 1234–1242. issn: 15353893. doi: 10.1021/pr049882h.

[84] Center for Computational Mass Spectrometry. *MassIVE: Mass Spectromety Interactive Virtual Environment*. url: https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp (visited on 2020-06-08).

[85] Q. Sun et al. “PPDB, the Plant Proteomics Database at Cornell”. In: *Nucleic Acids Research* 37.2 (2008), pp. 969–974. doi: 10.1093/nar/gkn654. url: http://compbio.dfci.harvard.edu/tgi/cgi-bin/.

[86] K. Haug et al. “MetaboLights-an open-access general-purpose repository for metabolomics studies and associated meta-data”. In: *Nucleic Acids Research* (2013). doi: 10.1093/nar/gks1004.

[87] A. J. Carroll, M. R. Badger, and A. Harvey Millar. “The MetabolomeExpress Project: Enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets”. In: *BMC Bioinformatics* 11.1 (2010-07), p. 376. issn: 14712105. doi: 10.1186/1471-2105-11-376. url: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-376.

[88] M. Sud et al. “Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools”. In: *Nucleic Acids Research* 44.8 (2015). doi: 10.1093/nar/gkv1042. url: www.ebi.ac.uk/chebi.

[89] J. Kopka et al. “The Golm Metabolome Database”. In: *Bioinformatics* 21.8 (2005), pp. 1635–1638. doi: 10.1093/bioinformatics/bti236.

[90] S. Robaina Estévez and Z. Nikoloski. “Generalized framework for context-specific metabolic model extraction methods”. In: *Frontiers in Plant Science* 5 (2014-09), p. 491. issn: 1664462X. doi: 10.3389/fpls.2014.00491. url: http://www.ncbi.nlm.nih.gov/pubmed/25285097%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4168813.

[91] D. Machado and M. Herrgård. “Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism”. In: *PLoS Computational Biology* 10.4 (2014). issn: 15537358. doi: 10.1371/journal.pcbi.1003580.

[92] R. P. Vivek-Ananth and A. Samal. “Advances in the integration of transcriptional regulatory information into genome-scale metabolic models”. In: *BioSystems* 147 (2016), pp. 1–10. issn: 18728324. doi: 10.1016/j.biosystems.2016.06.001. url: http://dx.doi.org/10.1016/j.biosystems.2016.06.001.

[93] M. Åkesson, J. Förster, and J. Nielsen. “Integration of gene expression data into genome-scale metabolic models”. In: *Metabolic Engineering* 6.4 (2004-10), pp. 285–293. issn: 10967176. doi: 10.1016/j.ymben.2003.12.002. url: http://www.ncbi.nlm.nih.gov/pubmed/15491858.

[94] S. A. Becker and B. O. Palsson. “Context-Specific Metabolic Networks Are Consistent with Experiments”. In: *PLoS Computational Biology* 4.5 (2008-05). Ed. by H. M. Sauro, e1000082. issn: 1553-7358. doi: 10.1371/journal.pcbi.1000082. url: https://dx.plos.org/10.1371/journal.pcbi.1000082.

[95] B. J. Schmidt et al. “GIM3E: Condition-specific models of cellular metabolism developed from metabolomics and expression data”. In: *Bioinformatics* 29.22 (2013), pp. 2900–2908. issn: 13674803. doi: 10.1093/bioinformatics/btt493.

[96] A. Bordbar et al. “Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics”. In: *Scientific Reports* 7.December 2016 (2017), pp. 1–12. issn: 20452322. doi: 10.1038/srep46249.

[97] C. Colijn et al. “Interpreting Expression Data with Metabolic Flux Models: Predicting Mycobacterium tuberculosis Mycolic Acid Production”. In: *PLoS Computational Biology* 5.8 (2009-08). Ed. by J. A. Papin, e1000489. issn: 1553-7358. doi: 10.1371/journal.pcbi.1000489. url: http://dx.plos.org/10.1371/journal.pcbi.1000489.

[98] M. K. Kim et al. “E-Flux2 and SPOT: Validated Methods for Inferring Intracellular Metabolic Flux Distributions from Transcriptomic Data”. In: *PLoS ONE* (2016). doi: 10.1371/journal.pone.0157101. url: http://most..

[99] C. Angione and P. Lió. “Predictive analytics of environmental adaptability in multi-omic network models”. In: *Scientific Reports* 5 (2015), pp. 1–21. issn: 20452322. doi: 10.1038/srep15147.

[100] A. Navid and E. Almaas. “Genome-level transcription data of Yersinia pestis analyzed with a New metabolic constraint-based approach”. In: *BMC Systems Biology* 6 (2012). issn: 17520509. doi: 10.1186/1752-0509-6-150.

[101] X. Fang, A. Wallqvist, and J. Reifman. “Modeling Phenotypic Metabolic Adaptations of Mycobacterium tuberculosis H37Rv under Hypoxia”. In: *PLoS Computational Biology* 8.9 (2012). issn: 1553734X. doi: 10.1371/journal.pcbi.1002688.

[102] S. Chandrasekaran and N. D. Price. “Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.41 (2010-10), pp. 17845–17850. issn: 00278424. doi: 10.1073/pnas.1005139107.

[103] H. Zur, E. Ruppin, and T. Shlomi. “iMAT: An integrative metabolic analysis tool”. In: *Bioinformatics* 26.24 (2010), pp. 3140–3142. issn: 13674803. doi: 10.1093/bioinformatics/btq602.

[104] R. Agren et al. “Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT”. In: *PLoS Computational Biology* 8.5 (2012). issn: 1553734X. doi: 10.1371/journal.pcbi.1002518.

[105] R. Agren et al. “Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling”. In: *Molecular Systems Biology* 10.3 (2014-03), p. 721. issn: 17444292. doi: 10.1002/msb.145122. url: http://www.ncbi.nlm.nih.gov/pubmed/24646661%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4017677.

[106] T. Shlomi et al. “Network-based prediction of human tissue-specific metabolism”. In: *Nature Biotechnology* 26.9 (2008), pp. 1003–1010. issn: 10870156. doi: 10.1038/nbt.1487.

[107] L. Jerby, T. Shlomi, and E. Ruppin. “Computational reconstruction of tissue-specific metabolic models: Application to human liver metabolism”. In: *Molecular Systems Biology* 6.401 (2010), pp. 1–9. issn: 17444292. doi: 10.1038/msb.2010.56. url: http://dx.doi.org/10.1038/msb.2010.56.

[108] Y. Wang, J. A. Eddy, and N. D. Price. "Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE". In: *BMC Systems Biology* 6 (2012). issn: 17520509. doi: 10.1186/1752-0509-6-153.

[109] N. Vlassis, M. P. Pacheco, and T. Sauter. "Fast Reconstruction of Compact Context-Specific Metabolic Network Models". In: *PLoS Computational Biology* 10.1 (2014). issn: 1553734X. doi: 10.1371/journal.pcbi.1003424.

[110] A. Schultz and A. A. Qutub. "Reconstruction of Tissue-Specific Metabolic Networks Using CORDA". In: *PLOS Computational Biology* 12.3 (2016-03). Ed. by C. D. Maranas, e1004808. issn: 1553-7358. doi: 10.1371/journal.pcbi.1004808.

[111] M. P. Pacheco et al. "Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network". In: *BMC Genomics* 16.1 (2015-10), p. 809. issn: 14712164. doi: 10.1186/s12864-015-1984-4. url: http://www.biomedcentral.com/1471-2164/16/809.

[112] P. A. Jensen and J. A. Papin. "Functional integration of a metabolic network model and expression data without arbitrary thresholding". In: *Bioinformatics* 27.4 (2011), pp. 541–547. issn: 13674803. doi: 10.1093/bioinformatics/btq702.

[113] R. J. Van Berlo et al. "Predicting metabolic fluxes using gene expression differences as constraints". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.1 (2011), pp. 206–216. issn: 15455963. doi: 10.1109/TCBB.2009.55.

[114] H. U. Kim, W. J. Kim, and S. Y. Lee. "Flux-coupled genes and their use in metabolic flux analysis". In: *Biotechnology Journal* 8.9 (2013), pp. 1035–1042. issn: 18606768. doi: 10.1002/biot.201200279.

[115] S. B. Collins, E. Reznik, and D. Segre. "Temporal Expression-based Analysis of Metabolism". In: *PloS one* 8.11 (2012), pp. 1–13. doi: 10.1371/Citation.

[116] N. Töpfer, S. Jozefczuk, and Z. Nikoloski. "Integration of time-resolved transcriptomics data with flux-based methods reveals stress-induced metabolic adaptation in Escherichia coli". In: *BMC Systems Biology* 6.May (2012). issn: 17520509. doi: 10.1186/1752-0509-6-148.

[117] D. Lee et al. "Improving metabolic flux predictions using absolute gene expression data". In: *BMC Systems Biology* 6 (2012). issn: 17520509. doi: 10.1186/1752-0509-6-73.

[118] B. E. Barker et al. "A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data". In: *Computational Biology and Chemistry* 59.0 0 (2016), pp. 98–112. doi: 10.1016/j.compbiolchem.2015.08.002.A.

[119] S. R. Estévez and Z. Nikoloski. "Context-specific metabolic model extraction based on regularized least squares optimization". In: *PLoS ONE* 10.7 (2015), pp. 1–21. issn: 19326203. doi: 10.1371/journal.pone.0131875.

[120] J. F. Moxley et al. “Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.16 (2009), pp. 6477–6482. issn: 00278424. doi: 10.1073/pnas.0811091106.

[121] J. Kim and J. L. Reed. “RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations.” In: *Genome biology* 13.9 (2012), R78. issn: 14656914. doi: 10.1186/gb-2012-13-9-r78. url: http://genomebiology.com/2012/13/9/R78.

[122] S. Rossell, M. A. Huynen, and R. A. Notebaart. “Inferring Metabolic States in Uncharacterized Environments Using Gene-Expression Measurements”. In: *PLoS Computational Biology* 9.3 (2013). issn: 1553734X. doi: 10.1371/journal.pcbi.1002988.

[123] M. Tian and J. L. Reed. “Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis”. In: *Bioinformatics* 34.22 (2018), pp. 3882–3888. issn: 14602059. doi: 10.1093/bioinformatics/bty445.

[124] C. Angione. “Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism”. In: *Bioinformatics* 34.3 (2018), pp. 494–501. issn: 14602059. doi: 10.1093/bioinformatics/btx562.

[125] K. Yizhak et al. “Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model”. In: *Bioinformatics* 26.12 (2010), pp. 255–260. issn: 13674803. doi: 10.1093/bioinformatics/btq183.

[126] M. L. Jenior et al. “Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments”. In: *PLOS Computational Biology* 16.4 (2020), e1007099. issn: 15537358. doi: 10.1371/journal.pcbi.1007099. url: http://dx.doi.org/10.1371/journal.pcbi.1007099.

[127] J. Ferreira et al. “Troppo - A Python Framework for the Reconstruction of Context-Specific Metabolic Models”. In: *Advances in Intelligent Systems and Computing* 1005 (2020), pp. 146–153. issn: 21945365. doi: 10.1007/978-3-030-23873-5_18.

[128] H. Fouladiha and S. A. Marashi. “Biomedical applications of cell- and tissue-specific metabolic network models”. In: *Journal of Biomedical Informatics* 68 (2017), pp. 35–49. issn: 15320464. doi: 10.1016/j.jbi.2017.02.014. url: http://dx.doi.org/10.1016/j.jbi.2017.02.014.

[129] M. G. Poolman et al. “A genome-scale metabolic model of Arabidopsis and some of its properties”. In: *Plant Physiology* 151.3 (2009), pp. 1570–1581. issn: 00320889. doi: 10.1104/pp.109.141267.

[130] M. G. Poolman et al. “Responses to light intensity in a genome-scale model of rice metabolism”. In: *Plant Physiology* 162.2 (2013), pp. 1060–1072. issn: 00320889. doi: 10.1104/pp.113.216762.

[131] A. Chatterjee and S. Kundu. “Revisiting the chlorophyll biosynthesis pathway using genome scale metabolic model of Oryza sativa japonica”. In: *Scientific Reports* 5.April (2015), pp. 1–15. issn: 20452322. doi: 10.1038/srep14975. url: http://dx.doi.org/10.1038/srep14975.

[132] A. Chatterjee et al. “Reconstruction of Oryza sativa indica genome scale metabolic model and its responses to varying RuBisCO activity, light intensity, and enzymatic cost conditions”. In: *Frontiers in Plant Science* 8.November (2017), pp. 1–18. issn: 1664462X. doi: 10.3389/fpls.2017.02060.

[133] M. Lakshmanan et al. “Unraveling the light-specific metabolic and regulatory signatures of rice through combined in silico modeling and multiomics analysis”. In: *Plant Physiology* 169.4 (2015), pp. 3002–3020. issn: 15322548. doi: 10.1104/pp.15.01379.

[134] R. Saha, P. F. Suthers, and C. D. Maranas. “Zea mays irs1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism”. In: *PLoS ONE* 6.7 (2011). issn: 19326203. doi: 10.1371/journal.pone.0021784.

[135] M. Simons et al. “Assessing the metabolic impact of nitrogen availability using a compartmentalized maize leaf genome-scale model”. In: *Plant Physiology* 166.3 (2014), pp. 1659–1674. issn: 15322548. doi: 10.1104/pp.114.245787.

[136] E. Bogart and C. R. Myers. “Multiscale metabolic modeling of C4 plants: Connecting nonlinear genome-scale models to leaf-scale metabolism in developing maize leaves”. In: *PLoS ONE* 11.3 (2016), pp. 1–27. issn: 19326203. doi: 10.1371/journal.pone.0151722. arXiv: 1502.07969. url: http://dx.doi.org/10.1371/journal.pone.0151722.

[137] L. A. Mueller, P. Zhang, and S. Y. Rhee. “AraCyc: A biochemical pathway database for Arabidopsis”. In: *Plant Physiology* 132.2 (2003-06), pp. 453–460. issn: 00320889. doi: 10.1104/pp.102.017236.

[138] C. Y. Cheung et al. “A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions”. In: *Plant Journal* 75.6 (2013), pp. 1050–1061. issn: 09607412. doi: 10.1111/tpj.12252.

[139] C. G. d. O. Dal’Molin et al. “AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis”. In: *Plant Physiology* 152.2 (2010), pp. 579–589. issn: 00320889. doi: 10.1104/pp.109.148817.

[140] B. K. S. Chung et al. “Genome-scale in silico modeling and analysis for designing synthetic terpenoid-producing microbial cell factories”. In: *Chemical Engineering Science* 103 (2013), pp. 100–108. issn: 00092509. doi: 10.1016/j.ces.2012.09.006. url: http://dx.doi.org/10.1016/j.ces.2012.09.006.

[141] R. Siriwach et al. “Drought Stress Responses in Context-Specific Genome-Scale Metabolic Models of Arabidopsis thaliana”. In: *Metabolites* 10.4 (2020), p. 159. issn: 2218-1989. doi: 10.3390/metabo10040159. url: https://www.mdpi.com/2218-1989/10/4/159.

[142] S. Mintz-Oron et al. “Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity”. In: *Proceedings of the National Academy of Sciences of the United States of America* 109.1 (2012), pp. 339–344. issn: 00278424. doi: 10.1073/pnas.1100358109.

[143] N. Töpfer et al. “Integration of genome-scale modeling and transcript profiling reveals metabolic pathways underlying light and temperature acclimation in Arabidopsis”. In: *Plant Cell* 25.4 (2013), pp. 1197–1211. issn: 10404651. doi: 10.1105/tpc.112.108852.

[144] N. Töpfer et al. “Variability of Metabolite Levels Is Linked to Differential Metabolic Pathways in Arabidopsis’s Responses to Abiotic Stresses”. In: *PLoS Computational Biology* 10.6 (2014). issn: 15537358. doi: 10.1371/journal.pcbi.1003656.

[145] S. M. Seaver et al. “Improved evidence-based genome-scale metabolic models for maize leaf, embryo, and endosperm”. In: *Frontiers in Plant Science* 6.MAR (2015). issn: 1664462X. doi: 10.3389/fpls.2015.00142.

[146] S. M. Seaver et al. “High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.26 (2014), pp. 9645–9650. issn: 10916490. doi: 10.1073/pnas.1401329111.

[147] M. Scheunemann, S. M. Brady, and Z. Nikoloski. “Integration of large-scale data for extraction of integrated Arabidopsis root cell-type specific models”. In: *Scientific Reports* 8.1 (2018), pp. 1–15. issn: 20452322. doi: 10.1038/s41598-018-26232-8. url: http://dx.doi.org/10.1038/s41598-018-26232-8.

[148] C. Y. Cheung et al. “A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C3 and crassulacean acid metabolism leaves”. In: *Plant Physiology* 165.2 (2014), pp. 917–929. issn: 15322548. doi: 10.1104/pp.113.234468.

[149] C. G. de Oliveira Dal’Molin et al. “A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems”. In: *Frontiers in Plant Science* 6.JAN (2015), pp. 1–12. issn: 1664462X. doi: 10.3389/fpls.2015.00004.

[150] R. Shaw and C. Y. Cheung. “A dynamic multi-tissue flux balance model captures carbon and nitrogen metabolism and optimal resource partitioning during arabidopsis growth”. In: *Frontiers in Plant Science* 9.June (2018), pp. 1–15. issn: 1664462X. doi: 10.3389/fpls.2018.00884.

[151] W. L. Schroeder and R. Saha. “A Computational Framework to Study the Primary Lifecycle Metabolism of Arabidopsis thaliana”. In: *bioRxiv Systems Biology* (2019-09), pp. 1–61. doi: 10.1101/761189. url: http://biorxiv.org/cgi/content/short/761189v1?rss=1%7B%5C&%7Dutm%7B%5C_%7Dsource=researcher%7B%5C_%7Dapp%7B%5C&%7Dutm%7B%5C_%7Dmedium=referral%7B%5C&%7Dutm%7B%5C_%7Dcampaign=RESR%7B%5C_%7DMRKT%7B%5C_%7DResearcher%7B%5C_%7Dinbound.

[152] A. R. Zomorrodi and C. D. Maranas. "OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities". In: *PLoS Computational Biology* 8.2 (2012-02). Ed. by C. V. Rao, e1002363. issn: 1553-7358. doi: 10.1371/journal.pcbi.1002363. url: https://dx.plos.org/10.1371/journal.pcbi.1002363.

[153] C. G. d. O. Dal'Molin et al. "C4GEM, a genome-scale metabolic model to study C4 plant metabolism". In: *Plant Physiology* 154.4 (2010), pp. 1871–1885. issn: 00320889. doi: 10.1104/pp.110.166488.

[154] R. A. Cañas et al. "Exploiting the genetic diversity of maize using a combined metabolomic, enzyme activity profiling, and metabolic modeling approach to link leaf physiology to kernel yield". In: *Plant Cell* 29.5 (2017), pp. 919–943. issn: 1532298X. doi: 10.1105/tpc.16.00613.

[155] Plant Metabolic Network (PMN). *CornCyc 4.0*. 2013. url: https://www.plantcyc.org/databases/corncyc/4.0 (visited on 2020-05-18).

[156] Gramene. *RiceCyc Database 3.2*. url: http://pathway.gramene.org/gramene/ricecyc.shtml (visited on 2020-05-18).

[157] F. Shen et al. "Transcriptomic and metabolic flux analyses reveal shift of metabolic patterns during rice grain development". In: *BMC Systems Biology* 12.Suppl 4 (2018). issn: 17520509. doi: 10.1186/s12918-018-0574-x.

[158] M. Lakshmanan et al. "Modeling rice metabolism: From elucidating environmental effects on cellular phenotype to guiding crop improvement". In: *Frontiers in Plant Science* 7.NOVEMBER2016 (2016), pp. 1–12. issn: 1664462X. doi: 10.3389/fpls.2016.01795.

[159] E. Grafahrend-Belau et al. "Multiscale metabolic modeling: Dynamic flux balance analysis on a whole-plant scale". In: *Plant Physiology* 163.2 (2013), pp. 637–647. issn: 00320889. doi: 10.1104/pp.113.224006.

[160] H. Yuan et al. "A genome-scale metabolic network reconstruction of tomato (Solanum lycopersicum L.) and its application to photorespiratory metabolism". In: *The Plant Journal* 85.2 (2015-01), pp. 289–304. issn: 09607412. doi: 10.1111/tpj.13075.

[161] K. Botero, S. Restrepo, and A. Pinzón. "A genome-scale metabolic model of potato late blight suggests a photosynthesis suppression mechanism 06 Biological Sciences 0607 Plant Biology". In: *BMC Genomics* 19.Suppl 8 (2018-12). issn: 14712164. doi: 10.1186/s12864-018-5192-x.

[162] T. Pfau et al. "The intertwined metabolism during symbiotic nitrogen fixation elucidated by metabolic modelling". In: *Scientific Reports* 8.1 (2018), pp. 1–11. issn: 20452322. doi: 10.1038/s41598-018-30884-x.

[163] T. B. Moreira et al. "A genome-scale metabolic model of soybean (Glycine max) highlights metabolic fluxes in seedlings". In: *Plant Physiology* 180.4 (2019), pp. 1912–1929. issn: 15322548. doi: 10.1104/pp.19.00122.

[164] R. Shaw and C. Y. Maurice Cheung. "A mass and charge balanced metabolic model of Setaria viridis revealed mechanisms of proton balancing in C4 plants". In: *BMC Bioinformatics* 20.1 (2019), pp. 1–11. issn: 14712105. doi: 10.1186/s12859-019-2941-z.

[165] E. Cunha et al. "The first multi-tissue genome-scale metabolic model of a woody plant highlights suberin biosynthesis pathways in Quercus suber". In: *PLOS Computational Biology* 19.9 (2023-09), pp. 1–27. doi: 10.1371/journal.pcbi.1011499. url: https://doi.org/10.1371/journal.pcbi.1011499.

[166] O. Dias et al. "Reconstructing genome-scale metabolic models with merlin". In: *Nucleic Acids Research* 43.8 (2015), pp. 3899–3910. issn: 13624962. doi: 10.1093/nar/gkv294.

[167] C. Gomes de Oliveira Dal'Molin and L. K. Nielsen. "Plant genome-scale reconstruction: from single cell to multi-tissue modelling and omics analyses". In: *Current Opinion in Biotechnology* 49 (2018), pp. 42–48. issn: 18790429. doi: 10.1016/j.copbio.2017.07.009. url: http://dx.doi.org/10.1016/j.copbio.2017.07.009.

[168] G. Zampieri et al. "A poly-omics machine-learning method to predict metabolite production in CHO cells". In: *2nd International Electronic Conference on Metabolomics*. Vol. 2. 2017, p. 4993. doi: 10.3390/iecm-2-04993.

[169] M. Cuperlovic-Culf. "Machine learning methods for analysis of metabolic data and metabolic pathway modeling". In: *Metabolites* 8.1 (2018). issn: 22181989. doi: 10.3390/metabo8010004.

[170] Z. Costello and H. G. Martin. "A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data". In: *npj Systems Biology and Applications* 4.1 (2018), pp. 1–14. issn: 20567189. doi: 10.1038/s41540-018-0054-3. url: http://dx.doi.org/10.1038/s41540-018-0054-3.

[171] J. Nabi. *Machine Learning —Fundamentals. Basic theory underlying the field of Machine Learning*. 2018. url: https://towardsdatascience.com/machine-learning-basics-part-1-a36d38c7916 (visited on 2020-07-14).

[172] M. Sanjay. *Why and how to Cross Validate a Model?* 2018. url: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f (visited on 2020-07-14).

[173] J. Brownlee. *Hyperparameter Optimization With Random Search and Grid Search*. 2020. url: https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/ (visited on 2024-02-20).

[174] Klosterman, Steve. *Overfitting, underfitting, and the bias-variance tradeoff*. 2019. url: https://towardsdatascience.com/overfitting-underfitting-and-the-bias-variance-tradeoff-83b42fb11efb (visited on 2020-07-14).

[175] J. C. F. Silva et al. "Machine learning approaches and their current application in plant molecular biology: A systematic review". In: *Plant Science* 284.March (2019), pp. 37–47. issn: 18732259. doi: 10.1016/j.plantsci.2019.03.020. url: https://doi.org/10.1016/j.plantsci.2019.03.020.

[176] T. F. M. Carvalho et al. “Rama: A machine learning approach for ribosomal protein prediction in plants”. In: *Scientific Reports* 7.1 (2017-12). issn: 20452322. doi: 10.1038/s41598-017-16322-4.

[177] J. C. F. Silva et al. “Fangorn forest (F2): A machine learning approach to classify genes and genera in the family Geminiviridae”. In: *BMC Bioinformatics* 18.1 (2017-09), p. 431. issn: 14712105. doi: 10.1186/s12859-017-1839-x. url: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1839-x.

[178] A. Moghimi et al. “A Novel Approach to Assess Salt Stress Tolerance in Wheat Using Hyperspectral Imaging”. In: *Frontiers in Plant Science* 9 (2018-08), p. 1182. issn: 1664-462X. doi: 10.3389/fpls.2018.01182. url: https://www.frontiersin.org/article/10.3389/fpls.2018.01182/full.

[179] S. Gutiérrez et al. “On-the-go hyperspectral imaging under field conditions and machine learning for the classification of grapevine varieties”. In: *Frontiers in Plant Science* 9 (2018-07). issn: 1664462X. doi: 10.3389/fpls.2018.01102.

[180] M. Pineda, M. L. Pérez-Bueno, and M. Barón. “Detection of bacterial infection in melon plants by classification methods based on imaging data”. In: *Frontiers in Plant Science* 9 (2018-02). issn: 1664462X. doi: 10.3389/fpls.2018.00164.

[181] V. Gligorijević and N. Pržulj. “Methods for biological data integration: Perspectives and challenges”. In: *Journal of the Royal Society Interface* 12.112 (2015). issn: 17425662. doi: 10.1098/rsif.2015.0571.

[182] Y. Li, F. X. Wu, and A. Ngom. “A review on machine learning principles for multi-view biological data integration”. In: *Briefings in Bioinformatics* 19.2 (2018), pp. 325–340. issn: 14774054. doi: 10.1093/bib/bbw113.

[183] M. D. Ritchie et al. “Methods of integrating data to uncover genotype-phenotype interactions”. In: *Nature Reviews Genetics* 16.2 (2015), pp. 85–97. issn: 14710064. doi: 10.1038/nrg3868. url: http://dx.doi.org/10.1038/nrg3868.

[184] M. Bersanelli et al. “Methods for the integration of multi-omics data: Mathematical aspects”. In: *BMC Bioinformatics* 17.2 (2016). issn: 14712105. doi: 10.1186/s12859-015-0857-9.

[185] S. Huang, K. Chaudhary, and L. X. Garmire. “More is better: Recent progress in multi-omics data integration methods”. In: *Frontiers in Genetics* 8.JUN (2017), pp. 1–12. issn: 16648021. doi: 10.3389/fgene.2017.00084.

[186] I. S. L. Zeng and T. Lumley. “Review of statistical learning methods in integrated omics studies (An integrated information science)”. In: *Bioinformatics and Biology Insights* 12 (2018). issn: 11779322. doi: 10.1177/1177932218759292.

[187] I. Subramanian et al. “Multi-omics Data Integration, Interpretation, and Its Application”. In: *Bioinformatics and Biology Insights* 14 (2020), pp. 7–9. issn: 11779322. doi: 10.1177/1177932219899051.

[188] H. Hotelling. “Relations Between Two Sets of Variates”. In: *Biometrika* 28.3/4 (1936-12), p. 321. issn: 00063444. doi: 10.2307/2333955. url: https://www.jstor.org/stable/2333955.

[189] WOLD H. “Estimation of principal components and related models by iterative least squares”. In: *Multivariate analysis*. Academic Press, 1966, pp. 391–420.

[190] K. A. Lê Cao et al. “Sparse canonical methods for biological data integration: Application to a cross-platform study”. In: *BMC Bioinformatics* 10.1 (2009-01), p. 34. issn: 14712105. doi: 10.1186/1471-2105-10-34. url: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-34.

[191] J. Trygg and S. Wold. “Orthogonal projections to latent structures (O-PLS)”. In: *Journal of Chemometrics* 16.3 (2002-03), pp. 119–128. issn: 0886-9383. doi: 10.1002/cem.695. url: http://doi.wiley.com/10.1002/cem.695.

[192] J. Trygg and S. Wold. “O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter”. In: *Journal of Chemometrics* 17.1 (2003-01), pp. 53–64. issn: 0886-9383. doi: 10.1002/cem.775. url: http://doi.wiley.com/10.1002/cem.775.

[193] M. Bylesjo et al. “Data integration in plant biology: the O2PLS method for combined modeling of transcript and metabolite data”. In: *The Plant Journal* 52.6 (2007-12), pp. 1181–1191. issn: 09607412. doi: 10.1111/j.1365-313X.2007.03293.x. url: http://doi.wiley.com/10.1111/j.1365-313X.2007.03293.x.

[194] A. Singh et al. “DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays”. In: *Bioinformatics* 35.17 (2019-01), pp. 3055–3062. issn: 1367-4803. doi: 10.1093/bioinformatics/bty1054. eprint: https://academic.oup.com/bioinformatics/article-pdf/35/17/3055/50720026/bioinformatics_35_17_3055.pdf. url: https://doi.org/10.1093/bioinformatics/bty1054.

[195] F. Rohart et al. “mixOmics: An R package for ‘omics feature selection and multiple data integration”. In: *PLOS Computational Biology* 13.11 (2017-11), pp. 1–19. doi: 10.1371/journal.pcbi.1005752. url: https://doi.org/10.1371/journal.pcbi.1005752.

[196] B. Lee et al. “Heterogeneous Multi-Layered Network Model for Omics Data Integration and Analysis”. In: *Frontiers in Genetics* (2020). doi: 10.3389/fgene.2019.01381. url: www.frontiersin.org.

[197] N. Di Nanni et al. “Network Diffusion Promotes the Integrative Analysis of Multiple Omics”. In: *Frontiers in Genetics* 11.February (2020), pp. 1–12. issn: 16648021. doi: 10.3389/fgene.2020.00106.

[198] C. Cortes and V. Vapnik. “Support-Vector Networks”. In: *Machine Learning* 20 (1995), pp. 273–297.

[199] A. Ben-Hur et al. “Support Vector Machines and Kernels for Computational Biology”. In: *PLoS Computational Biology* 4.10 (2008-10). Ed. by F. Lewitter, e1000173. issn: 1553-7358. doi: 10.1371/journal.pcbi.1000173. url: https://dx.plos.org/10.1371/journal.pcbi.1000173.

[200] M. Gönen and E. Alpaydın. *Multiple Kernel Learning Algorithms*. Tech. rep. Bogazici University, 2011, pp. 2211–2268.

[201] D. Rajasundaram and J. Selbig. "More effort - more results: Recent advances in integrative 'omics' data analysis". In: *Current Opinion in Plant Biology* 30 (2016), pp. 57–61. issn: 13695266. doi: 10.1016/j.pbi.2015.12.010. url: http://dx.doi.org/10.1016/j.pbi.2015.12.010.

[202] R. Ghan et al. "Five omic technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (Vitis vinifera L.) cultivars". In: *BMC Genomics* 16.1 (2015), pp. 1–26. issn: 14712164. doi: 10.1186/s12864-015-2115-y. url: http://dx.doi.org/10.1186/s12864-015-2115-y.

[203] P. A. Zaini et al. "Molecular Profiling of Pierce's Disease Outlines the Response Circuitry of Vitis vinifera to Xylella fastidiosa Infection". In: *Frontiers in Plant Science* 9 (2018-06), p. 771. issn: 1664-462X. doi: 10.3389/fpls.2018.00771. url: https://www.frontiersin.org/article/10.3389/fpls.2018.00771/full.

[204] D. Rajasundaram et al. "Understanding the relationship between cotton fiber properties and non-cellulosic cell wall polysaccharides". In: *PLoS ONE* 9.11 (2014). issn: 19326203. doi: 10.1371/journal.pone.0112168.

[205] A. Zamboni et al. "Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks". In: *Plant Physiology* 154.3 (2010-11), pp. 1439–1459. issn: 00320889. doi: 10.1104/pp.110.160275.

[206] T. Löfstedt and J. Trygg. "OnPLS-a novel multiblock method for the modelling of predictive and orthogonal variation". In: *Journal of Chemometrics* 25.8 (2011-08), n/a–n/a. issn: 08869383. doi: 10.1002/cem.1388. url: http://doi.wiley.com/10.1002/cem.1388.

[207] V. Srivastava et al. "OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipl- superoxide dismutase Populus plants". In: *BMC Genomics* 14.1 (2013). issn: 14712164. doi: 10.1186/1471-2164-14-893.

[208] A. Anesi et al. "Towards a scientific interpretation of the terroir concept: plasticity of the grape berry metabolome". In: *BMC Plant Biology* 15.1 (2015). issn: 14712229. doi: 10.1186/s12870-015-0584-4. url: http://dx.doi.org/10.1186/s12870-015-0584-4.

[209] A. Acharjee et al. "Integration of multi-omics data for prediction of phenotypic traits using random forest". In: *BMC Bioinformatics* 17.5 (2016). issn: 14712105. doi: 10.1186/s12859-016-1043-4. url: http://dx.doi.org/10.1186/s12859-016-1043-4.

[210] D. C. Wong and J. T. Matus. "Constructing integrated networks for identifying new secondary metabolic pathway regulators in grapevine: Recent applications and future opportunities". In: *Frontiers in Plant Science* 8.April (2017), pp. 1–8. issn: 1664462X. doi: 10.3389/fpls.2017.00505.

[211] J. Jiang et al. “Investigation and development of maize fused network analysis with multi-omics”. In: *Plant Physiology and Biochemistry* 141.June (2019), pp. 380–387. issn: 09819428. doi: 10.1016/j.plaphy.2019.06.016. url: https://doi.org/10.1016/j.plaphy.2019.06.016.

[212] N. D. Nguyen, I. K. Blaby, and D. Wang. “ManiNetCluster: A novel manifold learning approach to reveal the functional links between gene networks”. In: *BMC Genomics* 20.Suppl 12 (2019), pp. 1–14. issn: 14712164. doi: 10.1186/s12864-019-6329-2. url: http://dx.doi.org/10.1186/s12864-019-6329-2.

[213] S. Vijayakumar et al. “Seeing the wood for the trees: A forest of methods for optimization and omic-network integration in metabolic modelling”. In: *Briefings in Bioinformatics* 19.6 (2017), pp. 1218–1235. issn: 14774054. doi: 10.1093/bib/bbx053. arXiv: 1809.09475.

[214] A. Sahu et al. “Advances in flux balance analysis by integrating machine learning and mechanism-based models”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 4626–4640. issn: 20010370. doi: 10.1016/j.csbj.2021.08.004. url: https://doi.org/10.1016/j.csbj.2021.08.004.

[215] M. K. Khaleghi et al. “Synergisms of machine learning and constraint□based modeling of metabolism for analysis and optimization of fermentation parameters”. In: *Biotechnology Journal* August (2021), p. 2100212. issn: 1860-6768. doi: 10.1002/biot.202100212.

[216] A. Folch-Fortuny et al. “Principal elementary mode analysis (PEMA)”. In: *Molecular BioSystems* 12.3 (2016), pp. 737–746. issn: 17422051. doi: 10.1039/c5mb00828j.

[217] S. Bhadra et al. “Principal metabolic flux mode analysis”. In: *Bioinformatics* 34.14 (2018), pp. 2409–2417. issn: 14602059. doi: 10.1093/bioinformatics/bty049.

[218] A. Folch-Fortuny et al. “Dynamic elementary mode modelling of non-steady state flux data”. In: *BMC Systems Biology* 12.1 (2018), pp. 1–15. issn: 17520509. doi: 10.1186/s12918-018-0589-3.

[219] S. Magnúsdóttir et al. “Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota”. In: *Nature Biotechnology* 35.1 (2017), pp. 81–89. issn: 15461696. doi: 10.1038/nbt.3703.

[220] D. DiMucci, M. Kon, and D. Segrè. “Machine Learning Reveals Missing Edges and Putative Interaction Mechanisms in Microbial Ecosystem Networks”. In: *mSystems* 3.5 (2018), pp. 1–13. doi: 10.1128/msystems.00181-18.

[221] I. Shaked et al. “Metabolic Network Prediction of Drug Side Effects”. In: *Cell Systems* 2.3 (2016), pp. 209–213. issn: 24054720. doi: 10.1016/j.cels.2016.03.001. url: http://dx.doi.org/10.1016/j.cels.2016.03.001.

[222] T. Oyetunde et al. “Machine learning framework for assessment of microbial factory performance”. In: *PLoS ONE* 14.1 (2019), pp. 1–15. issn: 19326203. doi: 10.1371/journal.pone.0210558.

[223] J. J. Czajka, T. Oyetunde, and Y. J. Tang. “Integrated knowledge mining, genome-scale modeling, and machine learning for predicting Yarrowia lipolytica bioproduction”. In: *Metabolic Engineering* 67.July (2021), pp. 227–236. issn: 10967184. doi: 10.1016/j.ymben.2021.07.003.

[224] S. M. Schinn et al. “A genome-scale metabolic network model and machine learning predict amino acid concentrations in Chinese Hamster Ovary cell cultures”. In: *Biotechnology and Bioengineering* 118.5 (2021), pp. 2118–2123. issn: 10970290. doi: 10.1002/bit.27714.

[225] A. Antonakoudis et al. “Synergising stoichiometric modelling with artificial neural networks to predict antibody glycosylation patterns in Chinese hamster ovary cells”. In: *Computers and Chemical Engineering* 154 (2021). issn: 00981354. doi: 10.1016/j.compchemeng.2021.107471.

[226] S. Nandi, A. Subramanian, and R. R. Sarkar. “An integrative machine learning strategy for improved prediction of essential genes in Escherichia coli metabolism using flux-coupled features.” In: *Molecular bioSystems* 13.8 (2017-07), pp. 1584–1596. issn: 1742-2051. doi: 10.1039/c7mb00234c. url: http://www.ncbi.nlm.nih.gov/pubmed/28671706.

[227] K. Plaimas et al. “Machine learning based analyses on metabolic networks supports high-throughput knockout screens”. In: *BMC Systems Biology* 2 (2008), pp. 1–11. issn: 17520509. doi: 10.1186/1752-0509-2-67.

[228] L. Li et al. “Predicting enzyme targets for cancer drugs by profiling human Metabolic reactions in NCI-60 cell lines”. In: *BMC Bioinformatics* 11 (2010). issn: 14712105. doi: 10.1186/1471-2105-11-501.

[229] M. Kim et al. “Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli”. In: *Nature Communications* 7 (2016), pp. 1–12. issn: 20411723. doi: 10.1038/ncomms13090. url: http://dx.doi.org/10.1038/ncomms13090.

[230] C. Culley et al. “A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth”. In: *Proceedings of the National Academy of Sciences of the United States of America* 117.31 (2020), pp. 18869–18879. issn: 10916490. doi: 10.1073/pnas.2002959117.

[231] G. Magazzù, G. Zampieri, and C. Angione. “Multimodal regularized linear models with flux balance analysis for mechanistic integration of omics data”. In: *Bioinformatics* (2021). issn: 1367-4803. doi: 10.1093/bioinformatics/btab324.

[232] J. E. Lewis and M. L. Kemp. “Integration of machine learning and genome-scale metabolic modeling identifies multi-omics biomarkers for radiation resistance”. In: *Nature Communications* 12.1 (2021). issn: 20411723. doi: 10.1038/s41467-021-22989-1. url: http://dx.doi.org/10.1038/s41467-021-22989-1.

[233] M. B. Guebila and I. Thiele. “Predicting gastrointestinal drug effects using contextualized metabolic models”. In: *PLoS Computational Biology* 15.6 (2019), pp. 1–21. issn: 15537358. doi: 10.1371/journal.pcbi.1007100. url: http://dx.doi.org/10.1371/journal.pcbi.1007100.

[234] E. S. Kavvas et al. “A biochemically-interpretable machine learning classifier for microbial GWAS”. In: *Nature Communications* 11.1 (2020), pp. 1–11. issn: 20411723. doi: 10.1038/s41467-020-16310-9. url: http://dx.doi.org/10.1038/s41467-020-16310-9.

[235] W. Guo, Y. Xu, and X. Feng. “DeepMetabolism: A Deep Learning System to Predict Phenotype from Genome Sequencing”. In: *arXiv* (2017), pp. 1–7. arXiv: 1705.03094. url: http://arxiv.org/abs/1705.03094.

[236] S. G. Wu et al. “Rapid Prediction of Bacterial Heterotrophic Fluxomics Using Machine Learning and Constraint Programming”. In: *PLoS Computational Biology* 12.4 (2016). issn: 15537358. doi: 10.1371/journal.pcbi.1004838.

[237] E. Brunk et al. “Characterizing strain variation in engineered E. coli using a multi-omics based workflow”. In: *Physiology & behavior* 176.12 (2016), pp. 139–148. doi: 10.1016/j.physbeh.2017.03.040.

[238] A. A. Nagaraja et al. “Flux prediction using artificial neural network (ANN) for the upper part of glycolysis”. In: *PLoS ONE* 14.5 (2019), pp. 10–12. issn: 19326203. doi: 10.1371/journal.pone.0216178. url: http://dx.doi.org/10.1371/journal.pone.0216178.

[239] Y. Cai et al. “Multiclassification Prediction of Enzymatic Reactions for Oxidoreductases and Hydrolases Using Reaction Fingerprints and Machine Learning Methods”. In: *Journal of Chemical Information and Modeling* 58.6 (2018), pp. 1169–1181. issn: 15205142. doi: 10.1021/acs.jcim.7b00656.

[240] M. R. Amin et al. “DeepAnnotator: Genome Annotation with Deep Learning”. In: *ACM-BCB 2018 - Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (2018), pp. 254–259. doi: 10.1145/3233547.3233577.

[241] J. M. Dale, L. Popescu, and P. D. Karp. “Machine learning methods for metabolic pathway prediction”. In: *BMC Bioinformatics* 11.1 (2010-01), p. 15. issn: 1471-2105. doi: 10.1186/1471-2105-11-15. url: http://www.biomedcentral.com/1471-2105/11/15.

[242] I. Boudellioua et al. “Prediction of metabolic pathway involvement in prokaryotic uniprotkb data by association rule mining”. In: *PLoS ONE* 11.7 (2016), pp. 1–16. issn: 19326203. doi: 10.1371/journal.pone.0158896.

[243] J. P. Maheswari. *Breaking the curse of small datasets in Machine Learning. Towards Data Science*. url: https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d (visited on 2022-03-25).

[244] A. Vabalas et al. “Machine learning algorithm validation with a limited sample size”. In: *PLOS ONE* 14.11 (2019-11), e0224365. issn: 1932-6203. doi: 10.1371/JOURNAL.PONE.0224365. url: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224365.

[245] C. F. Caiafa et al. “Machine learning methods with noisy, incomplete or small datasets”. In: *Applied Sciences (Switzerland)* 11.9 (2021). issn: 20763417. doi: 10.3390/app11094132.

[246] S. Vijayakumar, P. K. Rahman, and C. Angione. “A Hybrid Flux Balance Analysis and Machine Learning Pipeline Elucidates Metabolic Adaptation in Cyanobacteria”. In: *iScience* 23.12 (2020). issn: 25890042. doi: 10.1016/j.isci.2020.101818.

[247] I. N. Jamil et al. “Systematic Multi-Omics Integration (MOI) Approach in Plant Systems Biology”. In: *Frontiers in Plant Science* 11.June (2020). issn: 1664462X. doi: 10.3389/fpls.2020.00944.

[248] K. Dolinski and O. G. Troyanskaya. “Implications of Big Data for cell biology”. In: *Molecular Biology of the Cell* 26.14 (2015-07), p. 2575. issn: 19394586. doi: 10.1091/MBC.E13-12-0756. url: /pmc/articles/PMC4501356/%20/pmc/articles/PMC4501356/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4501356/.

[249] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. 2nd. USA: John Wiley & Sons, Inc., 2002. isbn: 0471200247.

[250] Z. El Akkaoui et al. “A Model-Driven Framework for ETL Process Development”. In: *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP*. DOLAP ’11. Glasgow, Scotland, UK: Association for Computing Machinery, 2011, pp. 45–52. isbn: 9781450309639. doi: 10.1145/2064676.2064685. url: https://doi.org/10.1145/2064676.2064685.

[251] E. F. E F. Codd. “Derivability, Redundancy and Consistency of Relations Stored in Large Data Banks”. In: *SIGMOD Rec.* 38.1 (2009-06), pp. 17–36. issn: 0163-5808. doi: 10.1145/1558334.1558336. url: https://doi.org/10.1145/1558334.1558336.

[252] C. J. Date. *A guide to the SQL standard : a user’s guide to the standard relational language SQL*. Addison-Wesley Pub. Co, 1989, p. 228. isbn: 0201502097.

[253] A. Gupta et al. “NoSQL databases: Critical analysis and comparison”. In: *2017 International Conference on Computing and Communication Technologies for Smart Nation, IC3TSN 2017* 2017-October (2018-02), pp. 293–299. doi: 10.1109/IC3TSN.2017.8284494.

[254] J. Carlson. *Redis in action*. Simon and Schuster, 2013.

[255] B. Jose and S. Abraham. “Exploring the merits of nosql: A study based on mongodb”. In: *2017 International Conference on Networks and Advances in Computational Technologies, NetACT 2017* (2017-10), pp. 266–271. doi: 10.1109/NETACT.2017.8076778.

[256] G. Wang and J. Tang. “The NoSQL Principles and Basic Application of Cassandra Model”. In: *2012 International Conference on Computer Science and Service System*. 2012, pp. 1332–1335. doi: 10.1109/CSSS.2012.336.

[257] H. Huang and Z. Dong. “Research on architecture and query performance based on distributed graph database Neo4j”. In: *2013 3rd International Conference on Consumer Electronics, Communications and Networks, CECNet 2013 - Proceedings* (2013), pp. 533–536. doi: 10.1109/CECNET.2013.6703387.

[258] A. Ebrahim et al. “COBRApy: COnstraints-Based Reconstruction and Analysis for Python”. In: *BMC Systems Biology* 7.1 (2013-08), p. 74. issn: 1752-0509. doi: 10.1186/1752-0509-7-74. url: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-74.

[259] U. Sarkans et al. “From ArrayExpress to BioStudies”. In: *Nucleic Acids Research* 49.D1 (2021-01), pp. D1502–D1506. issn: 0305-1048. doi: 10.1093/NAR/GKAA1062. url: https://dx.doi.org/10.1093/nar/gkaa1062.

[260] J. Alston and O. Sambucci. “Grapes in the World Economy”. In: 2019-11, pp. 1–24. isbn: 978-3-030-18600-5. doi: 10.1007/978-3-030-18601-2_1.

[261] N. M. Saad et al. “Resveratrol: Latest scientific evidences of its chemical, biological activities and therapeutic potentials”. In: *Pharmacognosy Journal* 12.6 (2020), pp. 1779–1791. issn: 09753575. doi: 10.5530/pj.2020.12.240.

[262] O. Dias and I. Rocha. “Systems Biology in Fungi”. In: *Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi*. Ed. by R. Paterson. Boca Raton: CRC Press, 2015. Chap. 6, pp. 69–92. isbn: 9781466559868.

[263] A. M. Feist et al. “Reconstruction of Biochemical Networks in Microbial Organisms”. In: *Nature Reviews Microbiology* (2008). doi: 10.1038/nrmicro1949.

[264] C. Gu et al. “Current status and applications of genome-scale metabolic models”. In: *Genome Biology* 20.1 (2019), pp. 1–18. issn: 1474760X. doi: 10.1186/s13059-019-1730-3.

[265] R. Shaw and C. Y. Cheung. “Multi-tissue to whole plant metabolic modelling”. In: *Cellular and Molecular Life Sciences* 77.3 (2020), pp. 489–495. issn: 14209071. doi: 10.1007/s00018-019-03384-y. url: https://doi.org/10.1007/s00018-019-03384-y.

[266] S. Moretti et al. “MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models”. In: *Nucleic Acids Research* 49.D1 (2021-01), pp. D570–D574. issn: 0305-1048. doi: 10.1093/NAR/GKAA992. url: https://dx.doi.org/10.1093/nar/gkaa992.

[267] J. Hastings et al. “ChEBI in 2016: Improved services and an expanding collection of metabolites”. In: *Nucleic Acids Research* 44.D1 (2016-01), pp. D1214–D1219. issn: 0305-1048. doi: 10.1093/NAR/GKV1031. url: https://dx.doi.org/10.1093/nar/gkv1031.

[268] E. Cunha et al. “TranSyT, an innovative framework for identifying transport systems”. In: *Bioinformatics* 39.8 (2023-08), btad466. issn: 1367-4811. doi: 10.1093/bioinformatics/btad466. eprint: https://academic.oup.com/bioinformatics/article-pdf/39/8/btad466/51225339/btad466.pdf. url: https://doi.org/10.1093/bioinformatics/btad466.

[269] Y.-L. Cheng et al. “Grape and Wine Metabolites: Biotechnological Approaches to Improve Wine Quality”. In: *Intech*. Vol. 11. tourism. 2016, p. 13. isbn: 0000957720. url: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics.

[270] S. Santos and I. Rocha. In: *Journal of Integrative Bioinformatics* 13.2 (2016), pp. 1–14. doi: doi: 10.1515/jib-2016-285. url: https://doi.org/10.1515/jib-2016-285.

[271] F. Cruz et al. “BioISO: an objective-oriented application for assisting the curation of genome-scale metabolic models”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2024), pp. 1–13. doi: 10.1109/TCBB.2023.3339972.

[272] G. J. Niemann et al. “Differential chemical allocation and plant adaptation: A Py-MS Study of 24 species differing in relative growth rate”. In: *Plant and Soil* 175.2 (1995-08), pp. 275–289. issn: 0032079X. doi: 10.1007/BF00011364/METRICS. url: https://link.springer.com/article/10.1007/BF00011364.

[273] M. Massonnet et al. “Neofusicoccum parvum colonization of the grapevine woody stem triggers asynchronous host responses at the site of infection and in the leaves”. In: *Frontiers in Plant Science* 8.June (2017). issn: 1664462X. doi: 10.3389/fpls.2017.01117.

[274] M. Fasoli et al. “Timing and order of the molecular events marking the onset of berry ripening in grapevine”. In: *Plant Physiology* 178.3 (2018), pp. 1187–1206. issn: 15322548. doi: 10.1104/pp.18.00559.

[275] N. Vlassis, M. P. Pacheco, and T. Sauter. “Fast Reconstruction of Compact Context-Specific Metabolic Network Models”. In: *PLoS Computational Biology* 10.1 (2014). issn: 1553734X. doi: 10.1371/journal.pcbi.1003424.

[276] J. Ferreira et al. “Troppo - A Python Framework for the Reconstruction of Context-Specific Metabolic Models”. In: *Advances in Intelligent Systems and Computing* 1005 (2020), pp. 146–153. issn: 21945365. doi: 10.1007/978-3-030-23873-5_18.

[277] M. Maia et al. “More than just wine: The nutritional benefits of grapevine leaves”. In: *Foods* 10.10 (2021), pp. 1–16. issn: 23048158. doi: 10.3390/foods10102251.

[278] A. Richelle, C. Joshi, and N. E. Lewis. “Assessing key decisions for transcriptomic data integration in biochemical networks”. In: *PLoS Computational Biology* 15.7 (2019), pp. 1–18. issn: 15537358. doi: 10.1371/journal.pcbi.1007185.

[279] S. Bordel, R. Agren, and J. Nielsen. “Sampling the Solution Space in Genome-Scale Metabolic Networks Reveals Transcriptional Regulation in Key Enzymes”. In: *PLOS Computational Biology* 6.7 (2010-07), pp. 1–13. doi: 10.1371/journal.pcbi.1000859. url: https://doi.org/10.1371/journal.pcbi.1000859.

[280] P. Nanda and A. Ghosh. “Genome scale-differential flux analysis reveals deregulation of lung cell metabolism on SARS-CoV-2 infection”. In: *PLoS Computational Biology* 17.4 (2021), pp. 1–22. issn: 15537358. doi: 10.1371/journal.pcbi.1008860. url: http://dx.doi.org/10.1371/journal.pcbi.1008860.

[281] S. Kaul et al. “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. In: *Nature 2000 408:6814* 408.6814 (2000-12), pp. 796–815. issn: 1476-4687. doi: 10.1038/35048692. url: https://www.nature.com/articles/35048692.

[282] V. Kumar et al. “Differential distribution of amino acids in plants”. In: *Amino Acids* 49.5 (2017-05), pp. 821–869. issn: 14382199. doi: 10.1007/S00726-017-2401-X/TABLES/5. url: https://link.springer.com/article/10.1007/s00726-017-2401-x.

[283] J. Pérez-Navarro et al. “LC-MS/MS analysis of free fatty acid composition and other lipids in skins and seeds of Vitis vinifera grape cultivars”. In: *Food Research International* 125 (2019), p. 108556. issn: 0963-9969. doi: https://doi.org/10.1016/j.foodres.2019.108556. url: https://www.sciencedirect.com/science/article/pii/S096399691930434X.

[284] H. Xia et al. “Comparative analysis of flavonoids in white and red table grape cultivars during ripening by widely targeted metabolome and transcript levels”. In: *Journal of Food Science* 87.4 (2022), pp. 1650–1661. issn: 17503841. doi: 10.1111/1750-3841.16117.

[285] N. Benbouguerra et al. “Stilbenes in grape berries and wine and their potential role as anti-obesity agents: A review”. In: *Trends in Food Science and Technology* 112.April 2020 (2021), pp. 362–381. issn: 09242244. doi: 10.1016/j.tifs.2021.03.060.

[286] R. Flamini et al. “Advanced knowledge of three important classes of grape phenolics: Anthocyanins, stilbenes and flavonols”. In: *International Journal of Molecular Sciences* 14.10 (2013), pp. 19651–19669. issn: 16616596. doi: 10.3390/ijms141019651.

[287] J. Singh et al. *Phenolic Compounds of Grapes and Wines: Key Compounds and Implications in Sensory Perception*. Vol. 11. tourism. 2016, p. 13. isbn: 0000957720. url: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics.

[288] M. Massonnet et al. “Ripening transcriptomic program in red and white grapevine varieties correlates with berry skin anthocyanin accumulation”. In: *Plant Physiology* 174.4 (2017), pp. 2376–2396. issn: 15322548. doi: 10.1104/pp.17.00311.

[289] C. Wasternack and S. Song. “Jasmonates: biosynthesis, metabolism, and signaling by proteins activating and repressing transcription”. In: *Journal of Experimental Botany* 68.6 (2016-12), pp. 1303–1321. issn: 0022-0957. doi: 10.1093/jxb/erw443. eprint: https://academic.oup.com/jxb/article-pdf/68/6/1303/11202969/erw443.pdf. url: https://doi.org/10.1093/jxb/erw443.

[290] J. J. Kieber and G. E. Schaller. “Cytokinins”. In: *The Arabidopsis Book / American Society of Plant Biologists* 12 (2014-01), e0168. issn: 1543-8120. doi: 10.1199/TAB.0168. url: /pmc/articles/PMC3894907/%20/pmc/articles/PMC3894907/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3894907/.

[291] L. J. Sweetlove et al. “Not just a circle: flux modes in the plant TCA cycle”. In: *Trends in plant science* 15.8 (2010-08), pp. 462–470. issn: 1878-4372. doi: 10.1016/J.TPLANTS.2010.05.006. url: https://pubmed.ncbi.nlm.nih.gov/20554469/.

[292] T. C. Williams et al. “Metabolic Network Fluxes in Heterotrophic Arabidopsis Cells: Stability of the Flux Distribution under Different Oxygenation Conditions”. In: *Plant Physiology* 148.2 (2008-10), pp. 704–718. issn: 0032-0889. doi: 10.1104/PP.108.125195. url: https://dx.doi.org/10.1104/pp.108.125195.

[293] A. Garrido et al. “Fruit Photosynthesis: More to Know about Where, How and Why”. In: *Plants* 12.13 (2023). issn: 2223-7747. doi: 10.3390/plants12132393.

[294] P. P. Gauthier et al. “In folio isotopic tracing demonstrates that nitrogen assimilation into glutamate is mostly independent from current CO2 assimilation in illuminated leaves of Brassica napus”. In: *The New phytologist* 185.4 (2010-03), pp. 988–999. issn: 1469-8137. doi: 10.1111/J.1469-8137.2009.03130.X. url: https://pubmed.ncbi.nlm.nih.gov/20070539/.

[295] J. Liang and J. He. “Protective role of anthocyanins in plants under low nitrogen stress”. In: *Biochemical and Biophysical Research Communications* 498.4 (2018), pp. 946–953. issn: 0006-291X. doi: https://doi.org/10.1016/j.bbrc.2018.03.087. url: https://www.sciencedirect.com/science/article/pii/S0006291X18305795.

[296] H. Yin et al. “Effect of ammonium and nitrate supplies on nitrogen and sucrose metabolism of Cabernet Sauvignon (Vitis vinifera cv.)” In: *Journal of the Science of Food and Agriculture* 100.14 (2020-11), pp. 5239–5250. issn: 1097-0010. doi: 10.1002/JSFA.10574.

[297] M. J. Considine and C. H. Foyer. “Metabolic responses to sulfur dioxide in grapevine (Vitis vinifera L.): photosynthetic tissues and berries”. In: *Frontiers in Plant Science* 6 (2015). issn: 1664-462X. doi: 10.3389/fpls.2015.00060. url: https://www.frontiersin.org/articles/10.3389/fpls.2015.00060.

[298] S. M. Lundberg and S.-I. Lee. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. url: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[299] M. Wink. *Functions and Biotechnology of Plant Secondary Metabolites*. Second. 2010, pp. 1–410. isbn: 9781405185288. doi: 10.1002/9781444318876.

I

Supplementary Material

All supplementary files are described below and available at: https://bit.ly/phd_supplementary_material

I.1 Supplementary File 1

List of studies using ML for analysing plant omics data.

I.2 Supplementary File 2

List of class metabolite identifiers from PlantCyc and MetaCyc that were replaced in the model by the identifier of an instance metabolite of that class.

I.3 Supplementary File 3

List of reactions removed from the model despite the associated enzyme being identified in the *V. vinifera* genome. Most of these reactions represent generic unbalanced reactions replaced by specific reactions.

I.4 Supplementary File 4

Details on the generic biomass composition of *V. vinifera* and the composition of the tissue-specific biomass reactions as well as the equations for the biosynthesis of macromolecules that are biomass components.

I.5 Supplementary File 5

List of secondary metabolites analysed and included in the *V. vinifera* model.

I.6 Supplementary File 6

List of blocked reactions in the *V. vinifera* model.

I.7 Supplementary File 7

List of reactions added to the *V. vinifera* model during manual curation that were not automatically included in the model from the annotation results.

I.8 Supplementary File 8

List of reactions in the model whose reversibility was changed during manual curation.

I.9 Supplementary File 9

Comparisons between the *V. vinifera* model and other plant models. This file includes the statistics of the models (number of reactions, metabolites, genes and compartments) and the number of entities in common between models. It also comprises the list of unique pathways to *V. vinifera* model, the list of secondary pathways in all plant models, and the secondary reactions that are in other plant models and not in *V. vinifera* model.

I.10 Supplementary File 10

The number of reactions from each pathway in each tissue-specific model.

I.11 Supplementary File 11

Phenotype predictions of each tissue-specific model.

I.12 Supplementary File 12

Differential flux analysis with the list of pathways and reactions that carry differential flux between tissue-specific models.

I.13 Supplementary File 13

Phenotype predictions of each diel multi-tissue model.

I.14 Supplementary File 14

Phenotype predictions of each diel multi-tissue model under different levels of nitrate uptake.

I.15 Supplementary File 15

Phenotype predictions of each diel multi-tissue model under different levels of sulfate uptake.

I.16 Supplementary File 16

Correlation values between the variables of the RNA-Seq, metabolomics and predicted fluxomics datasets calculated for the Circos plot.

I.17 Supplementary File 17

List of the most contributing variables for DIABLO classifications.

I.18 Supplementary File 18

Summary of the most contributing features for all analyses with ML and DIABLO models.

I.19 Supplementary File 19

Generic *V. vinifera* GSMM (iMS7199) in SBML format.

I.20 Supplementary File 20

V. vinifera leaf GSMM in SBML format.

I.21 Supplementary File 21

V. vinifera stem GSMM in SBML format.

I.22 Supplementary File 22

V. vinifera green berry GSMM in SBML format.

I.23 Supplementary File 23

V. vinifera mature berry GSMM in SBML format.

I.24 Supplementary File 24

V. vinifera diel multi-tissue GSMM with green grape in SBML format.

I.25 Supplementary File 25

V. vinifera diel multi-tissue GSMM with mature grape in SBML format.