

Integrating machine learning and mechanistic modeling towards adaptation of plants to climate change

Ana Babo

UMinho | 2023

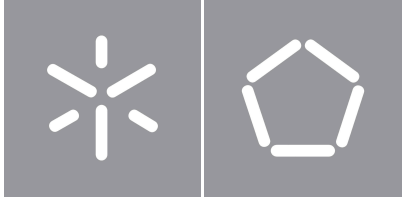


University of Minho
School of Engineering

Ana Camila Ribeiro de Babo

Integrating machine learning and mechanistic modeling towards adaptation of plants to climate change

october 2023



University of Minho
School of Engineering

Ana Camila Ribeiro de Babo

Integrating machine learning and mechanistic modeling towards adaptation of plants to climate change

Masters Dissertation
Master's in Bioinformatics

Dissertation supervised by
Óscar Manuel Lima Dias
Aalt-Jan Van Dijk

october 2023

Copyright and Terms of Use for Third Party Work

This dissertation reports on academic work that can be used by third parties as long as the internationally accepted standards and good practices are respected concerning copyright and related rights.

This work can thereafter be used under the terms established in the license below.

Readers needing authorization conditions not provided for in the indicated licensing should contact the author through the RepositóriUM of the University of Minho.

License granted to users of this work:



CC BY-SA

https://creativecommons.org/licenses/by-sa/4.0/

Acknowledgements

I want to express my gratitude to all those who have contributed, both directly and indirectly, to the successful completion of this project.

First and foremost, I want to express my endless gratitude to my parents. Despite their wish to keep their daughter close to them at all possible times, they wholeheartedly encouraged me to pursue this opportunity and to grow in every possible direction throughout this time. This exchange would have not been possible without their undying support, both emotionally and, with great effort, financially, as well. This support and effort extends far beyond these past months, carrying all of my years of education, and it has never gone unnoticed.

To Aalt-jan, who I am deeply grateful for trusting me with the opportunity to join the Bioinformatics Group and to take a peek into the world of plant science. I would like to sincerely thank you for all of the expertise, guidance, and patience provided throughout this project. For your time and effort in continuously scheduling weekly meetings, for the attention provided at each update presentation, and for every insight shared. I am especially grateful for every bit of feedback I have received that has and will continue to promote my growth on academic and interpersonal fronts.

To Ben, who so many times came to the rescue to the little corner in the student area. For sharing his knowledge, insights, and, most importantly, his work. Thank you for the Red Panda branch and for providing me with the opportunity to use your pipeline, which has largely enriched this thesis. It was a pleasure being your first supervisee!

To Professor Óscar, for the supervision and feedback provided throughout this project.

To everyone in the Bioinformatics Group, who have been so welcoming since day one. But a special thanks to my *parceritos* in crime, Alejandro, Devin, Laurens, Lucina, Maurício, Maurits, Jacky, and Katarina, that made my day-to-day in the office so lively. Thank you for all of the tips, feedback, motivating

words, and laughs shared over lunch. But maybe most unforgettably, thank you for the picnics, for sharing your culture and, most unforgettably, your food!

To all the beautiful people I met outside the Radix building, but most importantly to Niels, Riekje, Sila, and Thyco, who gave me a deep feeling of home during these months. Thank you for all of the weekend trips, the freezing 7 a.m. dips in the Rijn, breakfasts at Ivo's, Maté, for always providing me with a safe space and for always being so caring.

To Dioguinho and Tomasinho, who despite their continuous chaotic presence in my life, have had, quite unknowingly, a deep impact on the person I have grown into these past two years. Thank you for your ever-present company in the white glaring walls of the library and for constantly motivating me to get out of my bubble.

To all of my friends, who despite the distance and my occasional neglect in staying connected, have continued to offer their profound support and gentle reminders to find balance during the most demanding phases of this journey.

This six-month mobility exchange was made possible thanks to the financial support from the Erasmus+ traineeships funding.

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, october 2023

Ana Camila Ribeiro de Babo

Abstract

This dissertation focuses on the pressing issue of climate change and its impact on agricultural yield. Crop species, particularly polyploids, present challenges when investigating these effects. To overcome these challenges, our research centres on *Arabidopsis thaliana*, known for genes homologous to those linked to crop yield, found in various crops like wheat, rice, and tomato, contributing to similar traits and enhancing tolerance to adverse conditions. Furthermore, prior studies have revealed connections between these yield-associated genes and flowering genes in Arabidopsis.

The primary aim of this study is to improve the predictive power of Arabidopsis's flowering time in response to variations in ambient temperature by harnessing the potential of integrated machine learning and mechanistic modelling techniques. Synergizing machine learning's data-driven pattern recognition with mechanistic modelling's knowledge-based causal understanding enhances predictability and deepens our understanding of plant responses to temperature.

To achieve this goal, we employ three methods: Biological Network Integration using Convolutions (BIONIC), Python-based Weighted Gene Correlation Network Analysis (PyWGCNA), and Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN), for the robust detection of gene modules, the inference of relationships between module expression and relevant traits, as well as the mathematical formulation of found regulatory interactions. This endeavour is further strengthened by validating these relationships through Gene Set Enrichment Analysis (GSEA), ensuring a coherent functional grouping of modules within biologically relevant processes.

Our research uncovers 14 modules that exhibit significant associations with temperature. Notably, five of these modules are intricately linked to the flowering process, while two modules exhibit dual associations with cold response and flowering regulation. This work deepens our understanding of how temperature influences the regulatory interactions within Arabidopsis, offering valuable insights into the influence of climate on plant development.

Keywords machine learning, mechanistic modelling, climate change, *Arabidopsis thaliana*, temperature stress, flowering time

Resumo

Esta dissertação foca-se na questão das alterações climáticas e o seu impacto no rendimento agrícola. Espécies de culturas agrícolas, particularmente poliploides, desafiam a investigação desses efeitos. Para superar estas dificuldades, este estudo centra-se na *Arabidopsis thaliana*, conhecida pelos genes homólogos aos associados ao rendimento, encontrados em culturas como o trigo, o arroz e o tomate, que contribuem para características semelhantes e melhoram a tolerância a condições adversas. Estudos anteriores revelam também ligações entre genes associados ao rendimento e à floração na Arabidopsis.

O principal objetivo deste estudo é aprimorar a capacidade preditiva do tempo de floração da Arabidopsis em resposta a variações na temperatura ambiente, explorando o potencial da integração de técnicas de aprendizagem automática e modelação mecanística. A combinação da capacidade de reconhecimento de padrões da aprendizagem automática com a compreensão causal da modelação mecanística melhora a previsibilidade e aprofunda a nossa compreensão das respostas das plantas à temperatura.

Este estudo recorre a três métodos: *Biological Network Integration using Convolutions (BIONIC)*, *Python-based Weighted Gene Correlation Network Analysis (PyWGCNA)*, e *Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN)*, para detetar módulos de genes, inferir associações entre a expressão de módulos e características relevantes, e a formulação matemática de interações regulatórias. Esta abordagem é consolidada através de *Gene Set Enrichment Analysis (GSEA)*, assegurando o agrupamento funcional dos módulos em processos biologicamente relevantes.

Os resultados revelaram 14 módulos de genes que demonstram associações significativas com a temperatura. Cinco destes ligados ao processo de floração, enquanto dois módulos apresentam uma dupla associação com a resposta ao frio e a regulação da floração. Este trabalho aprofunda o nosso conhecimento sobre como a temperatura influencia as interações regulatórias na Arabidopsis, proporcionando perspetivas valiosas em relação à influência do clima no desenvolvimento da planta.

Palavras-chave aprendizagem automática, modelação mecanística, alterações climáticas, *Arabidopsis thaliana*, stress térmico, tempo de floração

Contents

1 Introduction **1**

1.1 Context and Motivation . . . 1

1.2 Research Objectives . . . 4

1.3 Report Overview . . . 5

2 State of the Art **6**

2.1 Plant-climate Interactions . . . 6

2.1.1 Background . . . 6

2.1.2 Temperature-mediated Plant Responses . . . 7

2.1.3 Plant-climate Interaction Data . . . 10

2.2 *Arabidopsis thaliana* . . . 11

2.2.1 Model Plant and Translational Crop Improvement . . . 11

2.2.2 Flowering-time Integration Network . . . 13

2.2.3 Temperature Regulation . . . 15

2.3 Integrated Mechanistic Modelling . . . 18

2.3.1 Background . . . 18

2.3.2 Modelling Plant Regulatory Networks . . . 19

2.3.3 Ordinary Differential Equations-based Modelling . . . 21

2.4 Machine Learning In Mechanistic Modelling . . . 24

2.4.1 Machine Learning Overview . . . 24

2.4.2 Inference of Gene Regulatory Modules . . . 25

2.4.3 Integration of Machine Learning and Mechanistic Modelling . . . 29

3 Methodology **31**

3.1 Data Collection . . . 31

3.2 BIONIC . . . 32

3.2.1 BIONIC's Design and Implementation . . . 34
3.2.2 Data Pre-processing . . . 36
3.2.3 Module Identification and Evaluation . . . 37
3.3 PyWGCNA . . . 38
3.3.1 PyWGCNA's Design and Implementation . . . 39
3.3.2 Data Pre-processing and Initialization of the PyWGCNA Object . . . 40
3.3.3 Module Identification . . . 41
3.3.4 Downstream Analysis and Module Visualization . . . 42
3.3.5 Comparison of Modules . . . 44
3.4 BADDADAN . . . 45
3.4.1 Data Pre-processing . . . 46
3.4.2 Extraction of Gene Modules . . . 47
3.4.3 Inference of GRNs . . . 47
3.4.4 Formulation and Simulation of ODEs . . . 48

4 Results and Discussion **51**
4.1 Extracted Gene Modules . . . 51
4.1.1 BIONIC . . . 51
4.1.2 PyWGCNA . . . 53
4.1.3 BADDADAN . . . 57
4.2 Analysis of Agreement in Extracted Modules . . . 61
4.3 Gene Set Enrichment Analysis . . . 62
4.3.1 PyWGCNA . . . 62
4.3.2 BADDADAN . . . 65

5 Conclusions and future work **67**
5.1 Conclusions . . . 67
5.2 Prospects for Future Work . . . 68

A Details of Results **94**

B Support work **97**

List of Figures

1 Flowering time regulatory pathways associated with temperature in *Arabidopsis thaliana*, adapted from Leijten et al. [2018] 14

2 Regulatory mechanisms controlling the flowering transition according to ambient temperature in *Arabidopsis thaliana*, adapted from Verhage et al. [2014]. 17

3 Representation of a Machine-Learning workflow for the inference of regulatory interactions, adapted from Camacho et al. [2018]. 25

4 Mathematical representations of a regulatory network, adapted from Lucas et al. [2011] . 26

5 Illustration representing diverse input data sources and the contrasting outcomes achieved through the use of data-driven and knowledge-driven approaches in data integration, adapted from Noor et al. [2019]. 30

6 Overview of Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network's unsupervised workflow, adapted from Forster et al. [2022] 34

7 Schematic overview of Python-based Weighted Gene Correlation Network Analysis's established workflow, adapted from Rezaie et al. [2023] 39

8 Schematic overview of Python-based Weighted Gene Correlation Network Analysis's gene clustering workflow, adapted from Rezaie et al. [2023]. 41

9 Visual representation and comparison of a random network and scale-free network. . . . 42

10 Schematic representation of the general workflow of the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach. . . 46

11 Evaluation of Biological Network Integration using Convolutions's module detection performance with the E-MEXP-728 dataset Balasubramanian et al. [2006]. 53

12 Correlation heatmap showing the relationships between extracted modules and data traits within the E-MEXP-728 dataset Balasubramanian et al. [2006]. 55

13 Correlation heatmap showing the relationships between extracted modules and data traits within the E-MEXP-1304 dataset Michael et al. [2008]. 56
14 Visual representation of modules extracted through the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach, alongside the corresponding transcription factors predicted by the TF2Network tool, in a network format. 58
15 Visual representation of modules extracted through the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach in a network format. 59
16 Plots illustrating the temporal expression dynamics of modules obtained through the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach based on experimental data and simulation of the ODE system. 60
17 Plot representing the changes in expression levels over time for all genes attributed to Module 0 with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach and E-MEXP-1304 dataset Michael et al. [2008]. 61
18 Bubble plot representing the Functional Enrichment Analysis for the modules extracted from the E-MEXP-728 dataset Balasubramanian et al. [2006], through the Python-based Weighted Gene Correlation Network Analysis approach 63
19 Bubble plot representing the Functional Enrichment Analysis for the modules extracted from the E-MEXP-1304 dataset Michael et al. [2008], through the Python-based Weighted Gene Correlation Network Analysis approach 64
20 Bubble plot representing the Functional Enrichment Analysis for the modules extracted from the E-MEXP-1304 dataset Michael et al. [2008], through the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach 65

A1 Network visualization of module similarity in the E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] datasets using Python-based Weighted Gene Correlation Network Analysis 95

A2 Network visualization of module similarity in the E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] datasets using Python-based Weighted Gene Correlation Network Analysis and Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network methods. 96

B1 Network visualization for the extracted modules of the E-MEXP-728 Balasubramanian et al. [2006] dataset through the Python-based Weighted Gene Correlation Network Analysis approach. 97

B2 Network visualization for the extracted modules of the E-MEXP-1304 Michael et al. [2008] dataset through the Python-based Weighted Gene Correlation Network Analysis approach.. 98

B3 Plot representing the changes in expression levels over time for all genes attributed to 'MODULE1' with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach. 99

B4 Plot representing the changes in expression levels over time for all genes attributed to 'MODULE2' with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach. 99

B5 Plot representing the changes in expression levels over time for all genes attributed to 'MODULE3' with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approach. 100

B6 Network visualization illustrating module similarity and outcome representations for the modules extracted from the E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] datasets using Python-based Weighted Gene Correlation Network Analysis and Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network approaches. 101

List of Tables

1 Overview of the experimental conditions within the chosen datasets, E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] 33

2 Definition of key parameters employed in the Ordinary Differential Equation's system. . . 59

List of Abbreviations

AGL24 Agamous-Like 24. 15

AI Artificial Intelligence. 3, 23

AMI Adjusted Mutual Information. 38, 52

AP1 Apetala 1. 14, 15

ARACNE Algorithm for the Reconstruction of Accurate Cellular Networks. 27, 28

AraGWAS *Arabidopsis thaliana* database Genome-wide association studies. 2, 11

ARF Auxin Response Factor. 13

ARI Adjusted Rand Index. 38, 52, 53

BADDADAN Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network. v, vi, 31, 45, 46, 47, 48, 49, 50, 51, 57, 58, 59, 60, 61, 62, 65, 66, 67, 69, 96, 99, 100, 101

BCE Binary Cross Entropy. 36

BFGS Broyden–Fletcher–Goldfarb–Shanno algorithm. 49

BIC Bayesian Information Criterion. 50

BIONIC Biological Network Integration using Convolutions. v, vi, 27, 28, 31, 33, 34, 35, 36, 37, 38, 51, 53, 67

CBF C-repeat-Binding Factor. 8, 9, 12, 57

ChIP-seq Chromatin Immunoprecipitation Sequencing. 19, 20, 29

CLR Context Likelihood of Relatedness. 27, 28

CNGCb Cyclic Nucleotide-gated Calcium Channel 2. 9

CO2 Carbon Dioxide. 7

COLDAIR Cold Assisted Intronic Noncoding RNA. 17

COR Cold-responsive. 9

CRPK1 Cold-responsive Protein Kinase 1. 9

CRT C-repeat. 8

CSV Comma-Separated Values. 36, 37, 46

DAP-seq DNA affinity purification sequencing. 20

DL Deep Learning. 33, 51

DNA Deoxyribonucleic acid. 9, 16, 17, 22

DREB Dehydration Responsive Element-Binding. 8, 9, 57

DREs Dehydration Responsive Elements. 8

FAO Food and Agriculture Organization of the United Nations. 1

FEA Functional Enrichment Analysis. 28, 38, 40, 42, 43

FLC Flowering Locus C. 15, 16, 17

FLM Flowering Locus M. 10, 16

FLM-δ Flowering Locus M-δ. 10, 16, 17

FLM-β Flowering Locus M-β. 10, 16, 17

FLOR-ID Flowering Interactive Database. 2, 11, 52

FRI Frigida. 17

FT Flowering Locus T. 14, 15, 16

GAT Graph Attention Network. 35, 36

GCN Graph Convolutional Network. 27, 33, 34, 35

GEO Gene Expression Omnibus. 10, 31

GRN Gene Regulatory Network. 4, 11, 13, 19, 20, 21, 22, 23, 27, 29, 46, 47, 48, 57

GSEA Gene Set Enrichment Analysis. v, vi, 29, 43, 62, 64, 65, 67, 68

H2A Histone 2A. 9

H2A.Z Histone 2A.Z. 9, 10, 16, 17

HS Heat Stress. 3, 6, 8, 21

HSF Heat Shock Factors. 8, 9

HSFA1 Heat Shock Factor A1. 9

IPCC Intergovernmental Panel on Climate Change. 1, 2

JSON JavaScript Object Notation. 36, 37, 38

L-BFGS Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm. 49

LD Long Day. 33

LFY Leafy. 14, 15

MAD Mean Absolute Deviation. 46, 47

MADS Minichrome Maintenance 1, Agamous, Deficiens, Serum Response Factor. 14, 15, 16

MAF MADS Affecting Flowering. 17

MAtDB Munich Information Center for Protein Sequences *Arabidopsis thaliana* database. 2, 11

Mb Megabytes. 11

MBF1c Multiprotein Bridging Factor 1c. 8, 57

miRNA micro RNA. 16

ML Machine Learning. 4, 5, 23, 24, 25, 27, 29, 31

mRNA messenger RNA. 22

MSE Mean Squared Error. 36

NCBI National Center for Biotechnology Information. 10

ODE Ordinary Differential Equation. 3, 4, 5, 18, 21, 22, 29, 46, 47, 48, 49, 57, 58, 60, 68, 69

PANTHER Protein ANalysis THrough Evolutionary Relationships. 43

PCA Principal Component Analysis. 25

PCC Pearson Correlation Coefficient. 43, 47

PDE Partial Differential Equations. 19, 22

PhyB Phytochrome B. 9

PIF Phytochrome-interacting Factor. 9, 10, 16, 17

PNN Probabilistic Neural Networks. 27

PWM Position Weight Matrice. 48

PyWGCNA Python-based Weighted Gene Correlation Network Analysis. v, vi, 28, 31, 38, 39, 40, 41, 42, 43, 44, 45, 51, 53, 54, 55, 61, 62, 63, 64, 65, 66, 67, 69, 95, 96, 97, 98, 101

RFIM Random-Field Ising Model. 27

RNA Ribonucleic acid. 9, 10, 11, 17, 28, 29, 53, 69

SA Shoot Apex. 14

SD Short Day. 33

SOC1 Suppressor of Overexpression of Constans 1. 14, 15, 16

SVP Short Vegetative Phase. 14, 15, 16, 17

TAB Tab-separated value. 47, 48

t-DNA transfer DNA. 12

TAIR The Arabidopsis Information Resource. 2, 11

TF Transcription Factor. 8, 12, 14, 15, 16, 19, 20, 47, 48, 57

TOM Topological Overlap Matrix. 40, 41

TPM Transcripts per Million. 54

VIN3 VERNALIZATION INSENSITIVE3. 17

VRN1 VERNALIZATION1. 17

VRN2 VERNALIZATION2. 17

WGCNA Weighted Gene Coexpression Network Analysis. 28, 38, 39, 40, 53

Chapter 1

Introduction

1.1 Context and Motivation

Plants are multicellular photosynthetic eukaryotic organisms contained within the Plantae kingdom Adl et al. [2005], Lewis and McCourt [2004]. Their domestication is estimated to have begun between 11,000 and 7,600 before the present years, when the earliest signs of human occupation were revealed, with the intent to produce food Piperno [2011]. Fast forward to today, plants are the main provider of human nutrition. High-yielding cereals and other starchy staples produce a large energy source, protecting sensible communities against starvation Mathers [2006], Sands et al. [2009].

A Food and Agriculture Organization of the United Nations (FAO) 2007 report states that only 3.5% of the total global land area is not affected by any environmental constraint Van Velthuizen [2007]. Environmental constraints include a variety of factors such as drought, anoxia, soil salinity, low and high temperatures, and low and high light exposure. When cultivated in any of these adverse environmental conditions, plants are exposed to abiotic stress that disrupts the normal functioning of physiological and biochemical pathways, inhibiting growth and crop yield Cramer et al. [2011], Sreenivasulu et al. [2007]. The effect of these stressors is expected to be aggravated by climate change, amplifying agriculture losses Minhas et al. [2017].

In the last 50 years, food production underwent an accentuated growth yet global food demand continues to increase. This is correlated with the continuous growth of the global population, higher consumption levels, and rising incomes Godfray et al. [2010], Von Braun [2007]. It is estimated that food demand will increase by 50% before 2030 Bruinsma [2017]. This issue combined with the current and predicted state of climate conditions has a high impact on worldwide agricultural production, posing a serious threat to the security of the global food supply Mittler et al. [2010], Wheeler and Von Braun [2013]. A 2013 Intergovernmental Panel on Climate Change (IPCC) report estimated temperature increases in the range of 0.8 to 4.8°C until the end of the twenty-first century Change [2013]. These projections are expected

to strongly impact plant survival, resulting in biodiversity loss Noble et al. [2014], Thuiller et al. [2005], Quint et al. [2016]. There are already reported cases of species loss where plants are unable to adjust the flowering-time response to increased temperatures Willis et al. [2008]. All of this data supports the following 2014 IPCC report that proposes the development of new crop varieties as one of the needs for adaptation to climate change Noble et al. [2014].

Flowering is a crucial stage of development in the plant life cycle. It occurs when a plant transitions from vegetative growth to a reproductive stage and its timing is the primary determinant of successful sexual reproduction, affecting seed number and dispersion. This phenomenon is regulated by genetic and molecular networks that integrate environmental and endogenous signals to induce flowering when conditions are optimal Craufurd and Wheeler [2009], Srikanth and Schmid [2011], Leal Valentim et al. [2015]. Additionally, alterations to growth and development stages are a natural response of plants to daily and seasonal temperature variations, a process designated as thermomorphogenesis. As such, these organisms are highly sensitive to thermal fluctuations Gil and Park [2019], Quint et al. [2016].

Arabidopsis thaliana (Arabidopsis) is a small flowering plant of the mustard family that is described to be an excellent model organism in Systems Biology for researching plant genetic, metabolic, proteinic, and developmental research, due to its small genome composed of five chromosomes that facilitated an extensive molecular comprehension Zachgo et al. [1996], Meinke et al. [1998], Wienkoop et al. [2010]. Over the years, the combined efforts of the research community produced a vast amount of diversified information that has been scattered in different resources and sometimes duplicated, limiting its access Meinke et al. [1998]. However, there has been a recent focus on combining and integrating all relevant data in context-specific databases and catalogues including The Arabidopsis Information Resource (TAIR) Rhee et al. [2003], Munich Information Center for Protein Sequences *Arabidopsis thaliana* database (MAtDB) Schoof et al. [2004], Flowering Interactive Database (FLOR-ID) Bouché et al. [2016], AraPheno Seren et al. [2016] and *Arabidopsis thaliana* database Genome-wide association studies (AraGWAS) Togninalli et al. [2018].

The flower transition process has been extensively studied in model organisms. The genetic analysis of mutant phenotypes of *Arabidopsis thaliana* in different environmental settings Weigel and Glazebrook [2008], Koornneef et al. [1991], Zhang et al. [2003], Michaels et al. [2004], Michaels [2009] combined with molecular biology analysis of genetic and molecular interactions He and Amasino [2005], Bäurle and Dean [2006], Tsuji et al. [2008] has allowed a clear identification and functional characterization of key genes involved in flowering and the regulatory pathways involved in the integration of developmental and environmental signals to induce flowering Jung and Müller [2009], Craufurd and Wheeler [2009]. Despite

these advances, the effect of temperature in these regulatory interactions is still understudied Craufurd and Wheeler [2009]. Recent data availability and new advances in Systems Biology approaches for plant omics have allowed an improvement in the analysis of plant-climate interactions, deriving knowledge on regulatory mechanisms of plants' response to abiotic stress Parmesan and Hanley [2015], Sreenivasulu et al. [2007]. The application of transcriptomics, more specifically gene expression data, enables the construction of biological networks through the identification of key genes and gene-gene interactions in regulatory networks Cramer et al. [2011]. These interactions are reported to have a varying rate of false positives and negatives, hence network integration methods emerged to address data limitations by resorting to algorithms that accurately identify biological modules through pattern recognition in gene interactions Mitra et al. [2013], Forster et al. [2022].

The improvement of Heat Stress (HS) tolerance and climate resistance in plants requires a comprehensive understanding of the stress response and regulatory interactions happening between the different compartments, at different levels of organization, in different spatio-temporal scales, and their impact on the whole-plant system Mittler et al. [2010], Chew et al. [2014]. This knowledge is achieved by Systems Biology approaches that model the functional interactions between the inner components Chew et al. [2014]. These models are implemented quantitatively, most commonly through Ordinary Differential Equation (ODE), that express time-dependent changes to simulate and predict the dynamics of plant behaviour, deriving a meaningful mechanistic and biological understanding of how interacting components regulate response to the surrounding environment Chew et al. [2014], Kitano [2002], Iyengar [2009], Seydel et al. [2022].

There has been an effort to develop quantitative models of the effect of temperature and HS in flowering regulation Hadley et al. [1983, 1984], Brewster [1994], Adams et al. [1997], Kinmonth-Schultz et al. [2018]. However, the reported models are simple, only regarding the effect of temperature on metabolic and reproductive development rate. More recent attempts resort to Neural Networks to support the accurate quantitative prediction of the effect of temperature in plant development Elizondo et al. [1992], Welch et al. [2003]. However, this approach presents a major limitation as the obtained model is trained for the specific initial conditions it was given, requiring a large amount of data to reach the desired level of generalization Legaard et al. [2021]. As such, there is still a need for improved simulations of the diverse temperature responses, particularly to stressful climate conditions, to correctly predict the flowering time for a wide range of environments Craufurd and Wheeler [2009].

With increasingly large data sets and experimental alternatives, integrating classic mathematical modelling approaches with Artificial Intelligence (AI) techniques improves the predictability power of the bio-

logical hypotheses at test, due to their effectiveness in data analysis, clustering, and trend detection. This combination can also limit scalability problems while optimizing context-appropriate parameter estimation. Nonetheless, when modelling within the biology scope it is important to avoid 'black-box' approaches. Despite limiting algorithm choice, it guarantees that both biological and physical laws are not disregarded and that the attained mechanistic insights are biologically meaningful Lopatkin and Collins [2020], Baker et al. [2018].

1.2 Research Objectives

The core objective of this dissertation is to synergize both Machine Learning (ML) and mechanistic modeling techniques, aiming to deepen our understanding of how plants respond to variations in ambient temperature. This process involves expanding upon an existing model for *Arabidopsis thaliana* relevant to the regulation of flowering time Leal Valentim et al. [2015].

To effectively address this research question, we set out to accomplish the following scientific and computational tasks:

- **Compilation of a data inventory:** Aggregating existing data and knowledge concerning *Arabidopsis thaliana*'s response to variations in ambient temperature, with a specific focus on gene expression levels and regulatory interactions.
- **Revision and expansion of Machine-Learning techniques:** Enhancing available ML methods to (1) extract gene modules linked to the plant's response to temperature stress and (2) predict the regulatory interactions among these gene modules.
- **Mathematical formulation of Gene Regulatory Network (GRN)s:** Formulate modules within GRN and their regulatory interactions as ODEs.
- **Integration of derived functions:** Incorporating the derived functions into the existing model for *Arabidopsis thaliana* to enhance its capability to predict flowering time in response to varying ambient temperature conditions.

These tasks collectively contribute to the overarching goal of this research, which is to improve our understanding of how plants, particularly *Arabidopsis thaliana*, adapt to changes in temperature, ultimately facilitating more accurate predictions of flowering time in response to varying environmental conditions.

1.3 Report Overview

The present dissertation is organized into three parts and five chapters, structured as follows:

- **Chapter 1: Introduction**: Explores the background, motivation and relevance of the research in the context of plant-climate interactions and temperature-mediated responses. Followed by the description of the specific research objectives set out to address the research problem. Finally, it provides a concise overview of the entire dissertation's structure and contents.
- **Chapter 2: State of the Art**: Explores various key topics central to this project, including plant-climate interactions, the model plant *Arabidopsis thaliana*, integrated mechanistic modeling, and the integration of ML techniques with mechanistic modeling.
- **Chapter 3: Methodology**: Provides an in-depth examination of data collection methods and sources, offering insights into the selected approaches and workflow for the module-detection and mechanistic modelling methodologies.
- **Chapter 4: Results and Discussion**: Presents key attained results, showcasing the gene modules extracted from different methods, their relevance to relevant biological processes, and a detailed comparison. The chapter introduces inferred regulatory interactions and ODE formulation. It concludes with validation efforts to affirm the biological significance of the findings.
- **Chapter 5: Conclusions and Future Work**: Summarizes the main conclusions derived from the research, highlighting key findings and implications. Discusses potential directions for future research and areas of improvement, building on the dissertation's contributions.
- **Appendix A: Details of Results**: Contains results that are crucial for the primary findings and conclusions but are presented separately to maintain the main text's readability.
- **Appendix B: Support Work**: Comprises supplementary data that, while not integral to the primary findings and conclusions, enhances the reader's comprehension of the research.

Chapter 2

State of the Art

2.1 Plant-climate Interactions

2.1.1 Background

Agriculture sustains the global population by providing food and other essential raw materials through crop production Harris and Fuller [2014]. Crops are a collection of growing plants that interact with each other and the surrounding environment. They are grown with an economic intention that is reliant on their quantitative production, designated as yield. To optimize the supply, farmers often manipulate environmental conditions and plant strain, selecting the more advantageous combination Forbes et al. [1992]. However, climate change presents new challenges beyond farmers' traditional methods to increase crop productivity Singh and Singh [2017].

While natural processes also contribute to the concentration of greenhouse gases in the atmosphere, anthropogenic activities are the primary driver of climate change with most of the global pattern of warming being observed in the last 50 years Solomon et al. [2007]. One of the main ways that climate change affects agriculture is through changes in temperature and precipitation patterns. Higher temperatures can lead to drought and HS, which can reduce crop yields, while extreme weather events such as floods and storms can damage crops and disrupt farming operations Porter et al. [2014], Craufurd and Wheeler [2009]. Previous studies have reported a 5 to 10% decrease in wheat yield per 1 °C increase in ambient temperature Craufurd and Wheeler [2009], Wheeler et al. [1996], Mitchell et al. [1993]. This is especially significant when intense climate variations occur in key developmental and growth stages, in particular the maturing of reproductive organs and sinks, like fruit and seeds. Mittler et al. [2010], Porter and Semenov [2005], Craufurd and Wheeler [2009]. Several studies have shown that global warming is responsible for changes in the timing of plant events such as flowering and leafing, as well as changes in the geographical distribution of plants Bertin [2008], Kelly and Goulden [2008]. Although rising levels

of atmospheric Carbon Dioxide (CO2) are expected to promote an increase in yield, the combined effect of high temperatures could limit or even revert this occurrence. These environmental conditions are also expected to alter the nutritional content of some crops. Crops grown in higher CO2 concentrations are less nutritious and are reported to have a 10-15% reduction in protein content Mittler et al. [2010], Taub et al. [2008].

Besides being responsible for food security, medicinal plants also play an important role in the pharmaceutical and biomedicinal industry due to their bioactive compounds exploited in drug discovery Balunas and Kinghorn [2005], Thatoi and Patra [2011]. Hence, recent research has been concentrated on identifying key sensors and regulators in stress adaptation to better regulate flowering time and improve abiotic stress tolerance Mittler et al. [2010]. Due to the scope of this study, the following section will be entirely focused on the plant's response to temperature among the different environmental factors.

2.1.2 Temperature-mediated Plant Responses

Plants are poikilothermic organisms whose internal temperature varies in accordance with the temperature in their immediate surroundings Körner [2006]. In addition, they are also sessile, meaning that they are unable to move and must endure and adapt to the conditions of their environment Morison and Morecroft [2008], Hatfield and Prueger [2015]. Hence these organisms are highly sensitive to variation in ambient temperature as it greatly impacts their growth, development, and morphology Sriden and Charoensawan [2022], Quint et al. [2016]. When these fluctuations are below the stress-inducing level, the shape and organization of the plant can sustain transformations that allow for better acclimation and adaptation. This process is designated as thermomorphogenisis Casal and Balasubramanian [2019], Quint et al. [2016], Crawford et al. [2012]. For example, when exposed to warmer temperatures, the air space between plant tissues increases, improving airflow and efficient leaf cooling through evaporation. On the contrary, thermal injury is provoked by extreme cold or hot temperatures that are responsible for injuries and even death of plants Forbes et al. [1992].

Every species has its particular temperature range, with a minimum, maximum, and optimal temperature at which it performs best Hatfield and Prueger [2015]. The model plant *Arabidopsis thaliana* is widely distributed with its ecotypes being found in varying growth environments. As such, these have genetic variations that provide them with the ability to tolerate various abiotic stresses and allow for better survival and adaptation to different environmental factors Hasdai et al. [2006]. With regards to temperature, the optimal range for this species is set between 21 and 22 °C, while temperatures between 27 and 28 °C are considered high and stressful. The temperatures between 28 and 37 °C are considered to be heat shock

temperatures, which can be harmful to the plant Li et al. [2018]. Despite these thresholds only differing by a few degrees, the morphological responses can be very contrasting. Effectively, warmer temperatures generally promote plant growth, by accelerating early flowering, increasing root growth, and inhibiting seed production, despite increasing susceptibility to pathogen infection; whilst heat shock drives cellular damages that trigger stress-responsive pathways, leading to changes such as slowed or inhibited root and shoot growth and decreased seed and pollen germination that reduces plant fertility Li et al. [2018], Sriden and Charoensawan [2022]. Colder temperatures also have an adverse effect on plants, hindering their growth and development due to reduced enzymatic activity and biochemical reactions. Temperatures around 14 °C are considered chilling, while temperatures between 6°C and 0 °C are cold. In this ambient, plants have a delayed flowering time, longer inflorescence, reduced concentration of chlorophyll, and increased anthocyanin content. Temperatures below 0 °C are regarded as freezing and are associated with a higher degree of damage caused by ice formation in the intercellular spaces that result in osmotic dehydration Hasdai et al. [2006], Thomashow [1999].

To survive and thrive under different environmental stresses, plants have developed complex transcriptional regulatory networks that control the expression of temperature-dependent genes that lead to plant acclimation Mittler et al. [2010]. This process involves multiple signal transduction pathways that are activated by signalling mechanisms, some of which have been previously described Nakashima et al. [2009], Xiong et al. [2002]. The expression of target genes is controlled by specific trans-acting proteins known as Transcription Factors TF that bind to cis-acting elements. These cis-elements are located in the promoter regions of genes and are considered the main regulators of the rate of transcription Mehrotra et al. [2011], Yamaguchi-Shinozaki and Shinozaki [2006, 2005]. These TFs include the Dehydration Responsive Element-Binding (DREB) 1/C-repeat-Binding Factor (CBF) and the Multiprotein Bridging Factor 1c (MBF1c) protein, involved in the cold and HS response in *Arabidopsis thaliana*, respectively Mittler et al. [2010], Nakashima et al. [2009], Suzuki et al. [2011], Yamaguchi-Shinozaki and Shinozaki [2005]. DREB 1/CBF is encoded by three genes (DREB1B/CBF1, DREB1A/CBF3, and DREB1C/CBF2) and is responsible for controlling the expression of more than 40 transcripts by binding specific cis-acting elements, such as the Dehydration Responsive Elements (DREs) and the C-repeat element, to their promoter regions Nakashima et al. [2009], Yamaguchi-Shinozaki and Shinozaki [2005]. The MBF1c is a transcriptional co-activator also encoded by three genes (MBF1c1a, b, and c) and controls 36 different transcripts that include HSF and the DREB2 protein Suzuki et al. [2011]. While it was only associated with drought and salt stress, it was recently reported that the DREB2A protein is an active regulator of the HS response, inducing the expression of HSF when overexpressed Sakuma et al. [2006], Schramm et al. [2008].

These regulatory systems are activated through thermosensing mechanisms and modulated by temperature-sensitive genes in signalling cascades that activate multiple responsive pathways. The products of these genes employ (1) tolerance mechanisms that increase the expression of osmoprotectants, ion transporters, regulatory proteins, and antioxidants; and (2) signal transduction pathways that control gene expression Ambrosino et al. [2020], Hasanuzzaman et al. [2013], Dai Vu et al. [2019], Nakashima et al. [2009], Yamaguchi-Shinozaki and Shinozaki [2006]. The alteration of gene expression patterns is achieved through the action of chromatin-remodelling enzymes in regulatory regions that modify the state of chromatin to either an "open" state, activating transcription, or a "closed" state, repressing transcription Mlynárová et al. [2007].

Thermosensing mechanisms and modulators are still understudied in plants. However, it is known that the red light Phytochrome B (PhyB), among other light sensors, plays a crucial role in perceiving temperature changes and its intrinsic role in light signalling Jung et al. [2016]. The processes governed by thermomorphogenisis happen mostly at night and are driven by the Phytochrome-interacting Factor (PIF)4 Jung et al. [2016], Sriden and Charoensawan [2022]. The activity of this transcription factor is regulated by phytochromes, namely, PhyB Delker et al. [2022]. An increase in ambient temperature stimulates the transformation of active PhyB into its inactive Pr form, decreasing the repression of the transcription factor PIF4. On the contrary, the active PhyB form suppresses the activity of PIF4 by promoting its breakdown through phosphorylation Delker et al. [2022]. Calcium channels, like the Cyclic Nucleotide-gated Calcium Channel 2 in Arabidopsis, also act as a primary thermosensor that combined with phytochrome signalling, activates cascade signalling and promotes the production of HSF such as the Heat Shock Factor A1 Nishad and Nandi [2021], Sriden and Charoensawan [2022]. Similar mechanisms occur in chilling and cold temperatures. The transcription of the CBF or the DREB1 is induced in these temperature ranges, both being responsible for the response to cold stress and tolerance to dehydration through the activation of cold-responsive Cold-responsive (COR) genes like COR15a and COR78 Stockinger et al. [1997], Sriden and Charoensawan [2022]. In freezing conditions, the Cold-responsive Protein Kinase 1 phosphorylates 14-3-3 proteins, destabilizing the CBF transcriptor factor to avoid an exaggerated cold-stress response Guo et al. [2018].

In addition to membranes and membrane proteins, recent research has also recognized other potential thermosensory mechanisms such as the remodelling of chromatin structures and the Ribonucleic acid (RNA) secondary structures Dai Vu et al. [2019]. Histone 2A.Z (H2A.Z) is a variant of the Histone 2A histone that when present at higher quantities near the initiation site decreases that gene's expression level. This phenomenon occurs due to this histone's ability to condense Deoxyribonucleic acid (DNA), re-

sulting in difficulted access to RNA polymerase II, inhibiting transcription. H2A.Z-containing nucleosomes demonstrate thermosensitivity as it is reported that they have different reactions to temperature variation, thermally regulating the transcriptome Sakamoto and Kimura [2018], Kumar and Wigge [2010]. Increases in temperatures decrease H2A.Z content in nucleosomes, promoting the transcription of thermosensitive genes Lorenzo et al. [2016]. Within the PIF family, PIF7 has been identified as an RNA-mediated thermosensor. When exposed to warmer environments, the conformation of a hairpin structure found in the 5' region of its transcript is altered, resulting in improved translation efficacy and, thus, increased protein levels. As PIF7 forms a heterodimer with PIF4, it collaborates in the regulation of genes associated with thermomorphogenisis Chung et al. [2020], Delker et al. [2022], Dai Vu et al. [2019]. Alternative splicing is also reported to have a crucial role in the thermosensory flowering response Dai Vu et al. [2019]. The alternative splicing of Flowering Locus M (FLM), a known flowering inhibitor, is a temperature-dependent that produces two antagonistic isoforms, Flowering Locus M-β (FLM-β) and Flowering Locus M-δ (FLM-δ), responsible for triggering different temperature responses for the flowering transition Scortecci et al. [2003], Sureshkumar et al. [2016], Pose et al. [2013], Zhang et al. [2003]. The integration of these signalling mechanisms in the regulation of flowering time will be better described later in section 2.2.3.

2.1.3 Plant-climate Interaction Data

Microarray and RNA sequencing technology allow for the simultaneous analysis of the expression levels of thousands of genes, enabling the identification of stress-induced genes associated with plant acclimation and abiotic stress tolerance Nakashima et al. [2009], Mittler et al. [2010], Seki et al. [2002], Kreps et al. [2002]. The availability of high-throughput omic data obtained from next-generation technologies has facilitated the comprehensive evaluation and comparison of gene expression datasets, elucidating the complex regulatory systems of the transcriptional response to abiotic stressors Sriden and Charoensawan [2022], Yamaguchi-Shinozaki and Shinozaki [2005]. However, it's worth noting that the majority of available datasets are focused on the analysis of paired differential gene expression (stressful versus non-stressful condition) with limited time points, highlighting the need for greater efforts to attain high-resolution temporal data that can better capture the dynamic regulation of genes in response to environmental factors Gehan et al. [2015].

Plant transcriptomic information can be obtained from a variety of databases such as the National Center for Biotechnology Information (NCBI), the Gene Expression Omnibus (GEO) database Barrett et al. [2010] and the ArrayExpress Athar et al. [2019], that function as archives for a high volume of genomic data retrieved from several plant species. These and similar databases have integrated a multitude of

profiling experiments performed under different environmental conditions that allow for the detection of similar expression patterns and gene-gene interactions.

Several public databases are solely composed of information regarding Arabidopsis, concentrating research efforts and limiting data duplication. These include The Arabidopsis Information Resource (TAIR) Rhee et al. [2003], Munich Information Center for Protein Sequences *Arabidopsis thaliana* database (MAtDB) Schoof et al. [2004], Flowering Interactive Database (FLOR-ID) Bouché et al. [2016], AraPheno Seren et al. [2016] and *Arabidopsis thaliana* database Genome-wide association studies Togninalli et al. [2018], among others. As of 2020, there were over 20 000 RNA-seq libraries solely performed on *Arabidopsis thaliana* Zhang et al. [2020]. Single-cell RNA sequencing has the potential to improve understanding of GRNs in plants as it allows for the characterization of the molecular regulatory mechanisms associated with plant resistance in different cell types and their states Tripathi and Wilkins [2021], Rhee et al. [2019]. Despite this, the majority of the performed studies until now have been focused on developmental processes with a few focused on Arabidopsis root cells Zhang et al. [2019], Ryu et al. [2019].

For this particular study, it is crucial to collect gene expression data under different ambient temperatures, especially regarding stress-inducing ones, with information on the flowering time, to infer gene-gene interactions associated with the regulatory response to temperature changes.

2.2 *Arabidopsis thaliana*

2.2.1 Model Plant and Translational Crop Improvement

Arabidopsis thaliana is a small flowering weed from the Brassicaceae family Zachgo et al. [1996], Sivasubramanian et al. [2015]. Since its proposal as a model organism, it has proven its significance in the study of plant biology research by contributing to the understanding of basic molecular mechanisms and the identification of key genes in fundamental processes Zachgo et al. [1996], Nelissen et al. [2014]. Due to its small genome size of approximately 125 Megabytes and a high proportion of repetitive sequences, this plant was selected as the organism of choice for the first plant genome sequencing project Page and Grossniklaus [2002], Sivasubramanian et al. [2015]. Arabidopsis is also self-pollinating and of compact size, with a short life cycle, a small number of chromosomes, and a well-annotated genome Korwin Krukowski et al. [2020], Sivasubramanian et al. [2015]. When compared to important crops, usually polyploids, its genome size can be three to 100 times smaller Sivasubramanian et al. [2015]. Despite not having any practical economic value, its characteristics have made it an ideal research subject, as the attained knowledge can be translated to more economically important crops, improving food security Korwin Krukowski

et al. [2020], Nelissen et al. [2014].

This process is known as plant translational research and can be described as the application of fundamental molecular understanding to improve agronomic crops Jacob et al. [2018]. As previously described, plants can adapt to alternations in their surrounding environment through physiological alterations that result in a decrease in yield. The majority of studies performed on the regulatory mechanisms associated with the abiotic stress response have been applied to *Arabidopsis thaliana* Dresselhaus and Hückelhoven [2018], Feuillet et al. [2011], Ferrier et al. [2011]. As such, key agronomic genes identified in *Arabidopsis thaliana* can either be directly used in commercial crops or be presumed as candidate genes for the discovery of ortholog genes Feuillet et al. [2011]. Despite efforts to domesticate and select crop strains over the years, which have affected both gene and network functions, key regulatory mechanisms responsible for controlling crucial processes such as plant growth, development, and stress tolerance can be conserved across different plant species, especially among eukaryotes Goldman et al. [2010], Gan et al. [2021]. It is commonly agreed upon by the scientific community that there should be a bigger focus on narrowing the gap between research on plant models and important crops to improve crop performance in real-world conditions Nelissen et al. [2014]. However, this progress has been pushed back by inherent variations in plant structure and behaviour Korwin Krukowski et al. [2020].

The direct comparative analysis of crop genomes with model plants, like Arabidopsis, accelerates the identification process of the different and conserved mechanisms between species and their ortholog genes Feuillet et al. [2011], Andrés and Coupland [2012]. For instance, the genome sequencing techniques enhanced for Arabidopsis have been streamlined for *Oryza sativa* (rice), speeding up the development of crop yield and stress tolerance Rensink and Buell [2004]. Translational biology has been known for its most successful advances in improving yield and stress tolerance in crops following the identification of candidate genes from Arabidopsis. In particular, CBFs identified in the model plant are also expressed in major economic crops such as *Triticum aestivum* (wheat), *Secale cereale* (rice), *Glycine max* (soybean), *Zea mays* (corn), *Solanum lycopersicum* (tomato) and *Solanum tuberosum* (potato), improving tolerance for freezing, salt, drought stressors Century et al. [2008], Zhang et al. [2004]. Additionally, the identification of the ortholog ZmNF-YB2 TF in maize through the comparison of the AtNF-YB1 TF in *Arabidopsis thaliana* also improved drought tolerance in the crop, after application of whole-plant transformation methods Nelson et al. [2007]. These transformations are made possible by *Agrobacterium tumefaciens* mediated approaches that resort to transfer DNA to insert mutations in cells Chang et al. [1994], Ferrier et al. [2011], Meinke et al. [1998].

The flowering time regulatory network has been thoroughly studied in Arabidopsis Leal Valentim et al.

[2015], Leijten et al. [2018]. This research is also relevant in providing target genes for increasing crop yield, as genes that influence this trait are homologs of flowering-controlling genes in Arabidopsis Korwin Krukowski et al. [2020]. For example, the identification of SlARF7 in tomato, an Auxin Response Factor (ARF) associated with the regulation of fruit formation, through the Arabidopsis homolog ARF7, associated with flowering regulation, allowed the transformation of the crop fruit with a mutated allele of ARF8 that resulted in increased fruit yield Ariizumi et al. [2013], Goetz et al. [2007]. However, the translation of this knowledge is still restricted for the majority of crops Leijten et al. [2018].

2.2.2 Flowering-time Integration Network

Flowering in a plant marks the transition from a vegetative state to a reproductive state, a key developmental stage in its life cycle Chailakhyan [1968], Hill and Li [2016]. The timing of this phenomenon is determinant for its reproductive success as it affects seed number Leal Valentim et al. [2015], Craufurd and Wheeler [2009]. Being a crucial trait for plant survival, it is highly sensitive to environmental cues. This process is reported to be more sensitive to environmental conditions than any other relevant agricultural trait Hill and Li [2016]. The regulation of the onset of flowering also coordinates with endogenous signalling and internal developmental programming to ensure it occurs in the most favourable conditions Hill and Li [2016], Leijten et al. [2018]. Among all environmental conditions, the temperature is the most critical one, especially when considering the current state of global warming Craufurd and Wheeler [2009], Leijten et al. [2018].

Previous studies have revealed that warmer temperatures are associated with early flowering and plant maturing, accompanying a reduced yield Wheeler et al. [1996], Mitchell et al. [1993]. In the context of agriculture, the production of engineered crops with improved abiotic stress tolerance and predictability in flowering time would assure high-yield production throughout the year and in varied environmental settings, increasing profits and security in food demand Jung and Müller [2009], Andrés and Coupland [2012]. The ability to control flowering time would facilitate production, harvest, and commercialization management, as this process also affects the quality and taste of the produce Leijten et al. [2018].

To effectively control this process it is necessary to take into account that every plant species has different varying optimal conditions for flowering due to their differently evolved regulation and response mechanisms Andrés and Coupland [2012], Srikanth and Schmid [2011]. For *Arabidopsis thaliana* it has been asserted that six core GRNs converge to the floral integrator genes, responsible for integrating the endogenous and exogenous cues to trigger flowering: the vernalization pathway, the autonomous pathway, the photoperiod/circadian clock pathway, gibberellin pathway, the ambient temperature pathway and the

age pathway (Figure 1) Srikanth and Schmid [2011], Fornara et al. [2010], Leijten et al. [2018], Leal Valentim et al. [2015]. Due to this project's scope, we will focus on the floral integration network, the ambient temperature, and vernalization regulatory pathways.

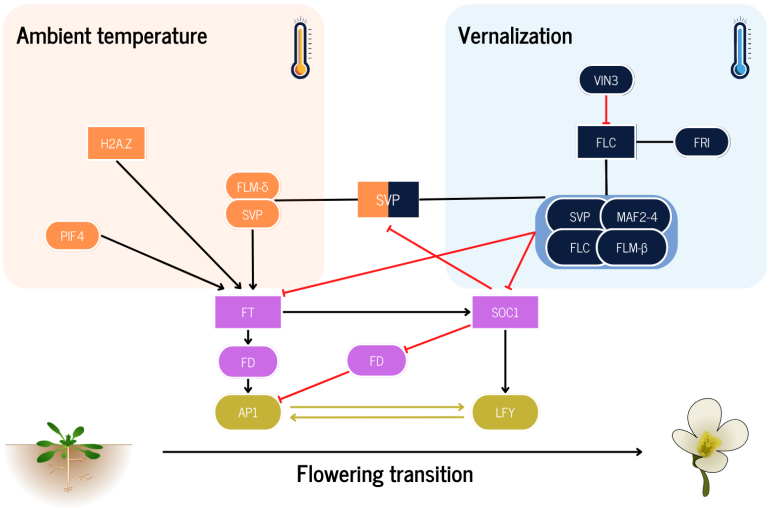


Figure 1: Flowering time regulatory pathways associated with temperature in *Arabidopsis thaliana*, adapted from Leijten et al. [2018]. The ambient temperature pathway is represented in orange and the vernalization pathway is in blue. The colour purple identifies the main floral integrators, Flowering Locus T (FT) and Suppressor of Overexpression of Constans 1 (SOC1); whereas the colour yellow portrays the inflorescence meristem identity genes, Apetala 1 (AP1) and Leafy (LFY). Genes crucial to the pathway are depicted as squared boxes, while boxes with rounded corners represent multiple genes/complexes. Arrows denote positive regulation (black) or negative regulation (red T-ends).

Three primary genes are involved in the integration of signalling pathways to regulate the flowering onset. These are the Suppressor of Overexpression of Constans 1 (SOC1), and Leafy (LFY) Moon et al. [2005]. Flowering Locus T (FT) is widely conserved among plant species and is reported to be the main regulator of flowering time, integrating environmental cues. It is formed at the leaves, being transported through the phloem to promote flowering at the Shoot Apex (SA) Abe et al. [2005], Wigge et al. [2005], Liu et al. [2012]. As such, FT is designated as a mobile signal, due to its spatial mobility Jang et al. [2009]. The transcription factor FD is expressed at the SA and integrates temporal and spatial information, mediating the FT transportation Wigge et al. [2005]. FT and FD are also interdependent antagonists where FD promotes the initiation of flowering while positively regulating SOC1 and Apetala 1 (AP1); whereas FT delays flowering Yoo et al. [2005].

Short Vegetative Phase (SVP) and SOC1 are TFs that belong to the Minichrome Maintenance 1, Aga-

mous, Deficiens, Serum Response Factor (MADS) family, a group of four genes involved in the regulation of the developmental processes of flowering Becker and Theißen [2003]. The SVP TF represses the activity of FT and SOC1 through its interaction with Flowering Locus C (FLC), another TF of the MADS-box family. Both SVP and FLC are shown to have a repressing role in the flowering of Arabidopsis Li et al. [2008], Helliwell et al. [2006], Jang et al. [2009], Hartmann et al. [2000]. The SOC1 TF has a crucial role in transmitting the outcome of multiple flowering-time signal transduction pathways to LFY through its direct binding to the gene's regulatory sequences Lee et al. [2008], Immink et al. [2012]. When overexpressed, SOC1 and Agamous-Like 24 (AGL24), another MADS-box TF with a reported role in the promotion of flowering, upregulate each other's expression. This evidence suggests that SOC1 and AGL24 operate as a protein complex Michaels et al. [2003], Lee et al. [2008].

The first step involved in the induction of the flowering process is the activation of floral meristem identity genes where the shoot apical meristem is transformed into an inflorescence meristem. However, in Arabidopsis, the use of this term is inaccurate as the flower primordia are formed from a non-specialized bract, designated as a rudimentary bract Kwiatkowska [2008], Long and Barton [2000]. The transformation to a floral meristem is induced by AP1, whose expression is upregulated through a positive feedback loop interaction with LFY. The initial expression of AP1 is responsible for the initiation of the floral primordia and is later on involved in the development of sepals and petals Weigel et al. [1992], Kaufmann et al. [2010], Alejandra Mandel et al. [1992].

2.2.3 Temperature Regulation

As reiterated throughout this report, plants display a sensitive responsiveness to the environmental temperature that is particular to each species' adaptation process and majorly influences its flowering time. These organisms understand seasonal changes through the accumulation of signalling molecules Sakamoto and Kimura [2018], He et al. [2020], Ding et al. [2020], such as the ones mentioned in the section 2.1.2. The flowering time response is widely varying for the different *Arabidopsis thaliana* accessions. For some varieties, the natural variation of ambient temperature they experience while in vegetative growth is enough to trigger the flowering process; whereas, for others, it is necessary to be exposed to a prolonged period of cold temperatures to be able to flower Leijten et al. [2018], Alonso-Blanco and Koornneef [2000]. This last phenomenon is designated as vernalization and it is commonly described as a long-term seasonal consequence of low temperatures Sakamoto and Kimura [2018].

Ambient Temperature Pathway

The ambient temperature pathway regulates flowering time according to variations in the plant's surrounding temperature (Figure 2) Leijten et al. [2018]. In comparison with other regulatory networks, the molecular mechanisms associated with the ambient temperature response and their respective interactions are still poorly understood Verhage et al. [2014], Srikanth and Schmid [2011], Leijten et al. [2018]. Despite this, two reported key regulators in this pathway are the SVP and the FLC gene Leijten et al. [2018]. These two floral repressors constantly interact during the plant's vegetative growth, acting co-dependently in its flowering regulation Li et al. [2008]. In lower temperature variations, SVP and FLC bind to the CArG motifs of floral activator genes FT and SOC1's promoter region, decreasing their expression levels and repressing early flowering Lee et al. [2007]. This regulation mechanism also involves two isoform regulators mentioned in section 2.1.2, the FLM-β and FLM-δ. These isoforms interact with the SVP protein forming the SVP–FLM-β and the SVP–FLM-δ complexes. At lower temperatures the SVP–FLM-β complex binds to DNA to repress flowering initiation; whereas the SVP–FLM-δ complex prevents flowering at higher temperatures Pose et al. [2013], Zhao et al. [2022]. Additionally, the miR156, a micro RNA (miRNA), aids in the repression of FT and delay of flowering time Capovilla et al. [2015], Kim et al. [2012].

The rise of ambient temperature results in the degradation of the SVP protein. In this scenario, SVP–FLM-δ, described as a dominant negative mutation of FLM due to its role in the disruption of MADS-box TF's functions, is responsible for the increase in FT and SOC1 transcription levels Capovilla et al. [2015]. The miR172, a miRNA with an inversely-correlated expression relation with miR156, is also an activator of FT and promoter of flowering time Verhage et al. [2014], Capovilla et al. [2015]. A gradual increase in temperature is also associated with a decrease in H2A.Z occupancy in nucleosomes of FT's promoter regions. This modification results in increased accessibility of FT's promoter binding site to PIF4, activating its transcription Kumar and Wigge [2010], Capovilla et al. [2015], Verhage et al. [2014].

Vernalization Pathway

Vernalization allows plants to flower when prolonged cold exposure occurs during the winter period, enabling them to blossom once favourable conditions arise Chouard [1960]. It also reinforces cold acclimation, improving the plant's tolerance to freezing temperatures Lee et al. [2007]. While being mainly observed in winter accessions of Arabidopsis, it is not an exclusive trait, occurring facultatively in other accessions with varying degrees of cold sensitivity Leijten et al. [2018]. This regulatory network (Figure 1) has a central mechanism associated with FLC, where flowering onset is promoted after a period of cold temperature through constant repression of its transcription Lutz et al. [2015], Sheldon et al. [2000].

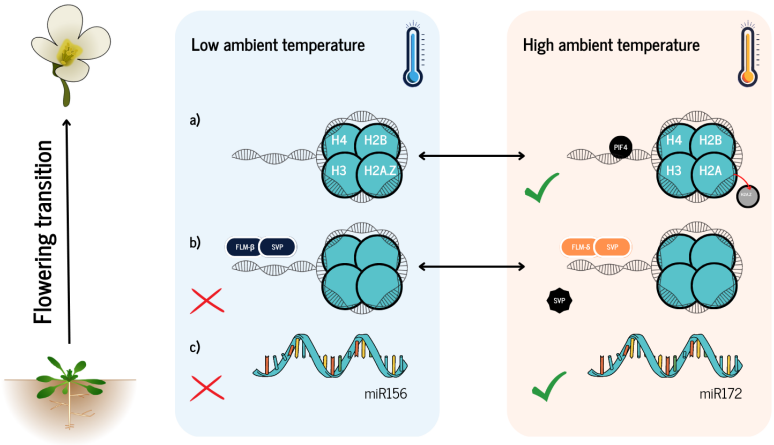


Figure 2: Regulatory mechanisms controlling the flowering transition according to ambient temperature in *Arabidopsis thaliana*. The figure displays the suppression mechanisms on the left and activation mechanisms on the right. a) At high temperatures, the Histone 2A.Z (H2A.Z) is removed from the nucleosome allowing the activator Phytochrome-interacting Factor (PIF)4 to bind to the genes that trigger flowering. b) At low temperatures, the Short Vegetative Phase (SVP)–Flowering Locus M-β (FLM-β) complex binds to Deoxyribonucleic acid (DNA) to suppress flowering. This repression is overcome at higher temperatures through the formation of SVP–FLM-δ complexes and degradation of the SVP protein, resulting in the activation of flowering. c) Representation of the miR172's suppressing effect at low temperatures and the miR156's activating effect at high temperatures. Adapted from Verhage et al. [2014]

This process is assured by chromatin level repression that includes histone modification by the activity of tree genes: VERNALIZATION1, VERNALIZATION2, and VERNALIZATION INSENSITIVE3 Lee et al. [2007], Leijten et al. [2018]. Additionally, Cold Assisted Intronic Noncoding RNA, a noncoding RNA from FLC, is also involved in his repression Sakamoto and Kimura [2018], Heo and Sung [2011].

In winter plant accessions, FLC expression levels are maintained at a high level due to Frigida positive regulation, ensuring that these plants do not flower prematurely Balasubramanian et al. [2006], Michaels and Amasino [1999]. The repression of flowering is also supported by the previously described activity of SVP–FLM-β and MADS Affecting Flowering2-4. MADS Affecting Flowering2 is said to prevent the activation of the vernalization pathway when plants are exposed to a short interval of cold temperatures, acting as a floral repressor Ratcliffe et al. [2003].

2.3 Integrated Mechanistic Modelling

2.3.1 Background

In the last few years, the progression of experimental methodologies used in the field of molecular biology has resulted in an immense volume of high-throughput quantitative data Lopatkin and Collins [2020], Kitano [2002]. However, the translation of this wealth of data into meaningful biological mechanistic knowledge is a challenging task as it requires extensive quantitative analysis Lopatkin and Collins [2020]. The emergence of the Bioinformatics field has facilitated these analyses through the use of computational techniques Luscombe et al. [2001]. This is especially true when regarding the study of complex regulatory networks where Systems Biology techniques have aided in the identification of new regulators, target genes, and interaction pathways to better understand the behaviour of plants and their functional relationships established between individual components in response to environmental cues Chew et al. [2014], Linde et al. [2015].

This research field is known as network inference, and it relays on mathematical models to test new hypotheses and predict the dynamics and structure of biological systems, including the structure of the interactions between elements and their directionality Lopatkin and Collins [2020], Linde et al. [2015], Kitano [2002]. To accomplish this, these models consider the molecular mechanisms and interactions underlying plant's responses through their mathematical formulation, allowing for a comprehensive representation of the system's behaviour Legaard et al. [2021], Baker et al. [2018], Lopatkin and Collins [2020]. This technique allows for the precise representation of plant systems in a quantitative form that best describes its naturally observed dynamics Lopatkin and Collins [2020]. To validate if the hypothesis being studied is coherent with real-life observations, the range of values that the predicted input-output variables can have to accurately capture the system is evaluated with analytical tools Baker et al. [2018].

The dynamical models used to describe the behaviour of a system can be discrete or continuous. For discrete models, the variables representing the time and/or molecular components of the system can only assume specific values, meaning that their state is fixed between defined points in time; whereas for continuous models, the time and/or state variables can vary continuously and, as such, assume any value within a range for the defined time window Lopatkin and Collins [2020]. The mathematical method chosen to employ these models can vary substantially according to the purpose of the model. However, the time-dependent variation of the concentration of the biomolecules is predominantly represented through a system of ODEs Coen et al. [2004]. In instances where the model requires the consideration of not only

temporal variation but also a spatial variation to describe space distribution, designated as a compartmental model, its state changes can be described as a set of Partial Differential Equations (PDE) Chew et al. [2014].

2.3.2 Modelling Plant Regulatory Networks

A comprehensive understanding of plants' response to the surrounding environment requires the identification and further characterization of GRNs that govern the physiological response in stress adaptation Sun and Dinneny [2018], Moreno-Risueno et al. [2010]. The majority of the research efforts in this scope have been applied in the model plant *Arabidopsis thalina* and have often been used as a framework to study similar processes in crop species Ferrier et al. [2011], Provart et al. [2016]. These advancements have been facilitated by increasing plant genome sequencing projects that expand transcriptome datasets and enable network reconstruction Moreno-Risueno et al. [2010]. Modelling these networks entails establishing hierarchical relationships between molecular regulators that further deepen our knowledge of the temporal and/or spatial mechanisms regulating complex agronomical traits and ultimately drive the improvement of stress tolerance in plants Li et al. [2015], Sun and Dinneny [2018], Mejia-Guerra et al. [2012], Ferrier et al. [2011].

The nature of the high-throughput data used to model GRNs can vary greatly Moreno-Risueno et al. [2010], Nachman et al. [2004]. However, building dynamic models requires temporal transcriptional data to depict the temporal behaviour of genes of interest. As mentioned in section 2.1.3, this data includes time-course expression data obtained from experimental techniques such as microarrays and next-generation sequencing. Assimilating gene functionality is often performed experimentally through phenotype examination of loss-of-function mutants. Such mutants are induced through insertional knockout mutations that when performed under various environmental settings, provide valuable insights into the gene roles underlying the plant's response to environmental cues Tran et al. [2007]. Interacting genes and regulators can also be deducted through the analysis of correlated expression levels present in the data. Genes exhibiting similar expression patterns, described as co-expressed, are likely triggered by identical regulatory mechanisms and, as such, may be functionally related Jen et al. [2006], Li et al. [2015]. These networks are known as coexpression networks and despite supporting the identification of direct or indirect gene interactions they have limited potential in predicting biological functions. This limitation can be overcome by the inclusion of additional data sources that convey casual relationships by taking into account the directionality of the interacting components. This data includes cis-regulatory elements, gene ontology annotation, and TF-binding sites Vandepoele et al. [2009], Li et al. [2015]. For instance, Chromatin Immunoprecipi-

tation Sequencing (ChIP-seq) and DNA affinity purification sequencing and other related techniques can identify TF-binding sites on a genome-wide scale, allowing for a better understanding of GRNs' dynamics Bartlett et al. [2017].

Despite the immense availability of experimental data, the inference of GRNs in *Arabidopsis thalina* still presents itself as a challenging task. One of the reasons associated with this is its genome complexity, with an estimated 28,000 genes, where around 2,000 are responsible to encode TFs with many others encoding various regulators Needham et al. [2009], Ocone et al. [2015], Ferrier et al. [2011]. This genetic abundance makes it difficult to identify interactions, limiting the understanding of their roles in the plant's development and response to environmental changes. Additionally, to improve the accuracy of the predictions made by the models, a large amount of data is required. To achieve this, several different datasets must be integrated to reach the desired level of dimensionality, However, this process can also introduce added noise and technical variability, which can further complicate the analysis Li et al. [2015].

Regardless of these challenges, the identification of key regulatory genes involved in the developmental process in Arabidopsis has been accomplished through the reconstruction of GRNs. Some of these processes include the circadian clock Locke et al. [2005], photoperiod regulatory network Salazar et al. [2009], and floral organ specification Alvarez-Buylla et al. [2010], Sanchez-Corrales et al. [2010], Barrio et al. [2010], with the control of flowering time becoming an active area of research in recent years. Efforts to model the regulatory network involved in this crucial reproductive process have been met with success, resulting in the identification of more than 180 genes that play a role in this process Leal Valentim et al. [2015], Fornara et al. [2010], Higgins et al. [2010], Wellmer and Riechmann [2010].

Introducing plants' responses to exogenous stimuli in already established developmental GRNs models can provide valuable insights into how environmental conditions influence the transcriptional state of cells in specific developmental stages. This can be achieved by integrating time course data from experiments conducted under a variety of environmental conditions. This additional data not only has the power to improve the models' performance but also deepen the understanding of inherent gene expression dynamics Dinneny et al. [2008], Leal Valentim et al. [2015]. However, despite current global warming trends, ambient growth temperature has been widely neglected Blázquez et al. [2003].

Previous studies have tried to predict the seasonal timing of various developmental processes in plants Chew et al. [2012]. Earlier phenology models introduced the temperature variable as thermal units Granier et al. [2002], Gobin [2012]. This idea stemmed from the conception that the developmental rate of plants was directly correlated with the accumulated temperature that was needed to complete developmental processes (thermal history) Monteith [1981]. This notion was later dismissed after discovering that the

temperature threshold that triggered flowering depended on day length, leading to the introduction of modified photothermal units Williams [1974]. However, these models did not consider underlying genetic variability until later. The integration of genetic and molecular information greatly improved models' predictions Wilczek et al. [2009], Kinmonth-Schultz et al. [2018], Brachi et al. [2010].

To fully understand the timing of developmental stages in plants, it becomes crucial to consider the dynamics of the genetic interactions involved in regulatory pathways to accurately predict the timing of these processes. Despite being limited, more recent studies have started to incorporate this knowledge. Satake et. al. (2013) constructed a predictive ODE-based model that predicted that a 4.5–5.3°C increase in temperature would halt flowering in *Arabidopsis halleri* Satake et al. [2013]. However, the model had some limitations. Not only did it not consider HS temperatures, but it also integrated a rather small number of genes. Additionally, the use of *Arabidopsis halleri* can limit the application to other crop species as it is not commonly used as a model plant. Ultimately, developing predictive models that factor in ambient temperature regulatory networks remains challenging due to the complexity of temperature-dependent interactions and the limited understanding of the underlying molecular mechanisms that regulate this response in plants.

2.3.3 Ordinary Differential Equations-based Modelling

The complex genetic and mechanistic interactions that underlie cellular regulation can only be thoroughly elucidated through the application of formal analysis methods Stelling and Gilles [2004]. Combining mathematical and computational representations that reliably portray real-life systems enables *in silico* experimentation of test hypotheses and behaviour prediction that would be difficult or impossible to employ experimentally Minorsky [2003]. To construct a mathematical model of GRNs it is necessary to consider the model's structure to accurately describe the dynamics of the system. This process involves the definition of the rates of change over time (kinetics) and the use of experimental data for validation Stelling and Gilles [2004]. To reduce the inherent complexity of GRNs it is possible to take into account a modular structure where different modules act independently Wang et al. [2008]. This modularity allows for the development of separate mathematical models using reusable functional units, improving the coherency of network representation Lauffenburger [2000].

GRNs can be rendered through a wide variety of techniques covering graphical representations, logic relationships, and mathematical formulation Alvarez-Buylla et al. [2007]. Among these, mathematical modelling is one of the most popular approaches and can be accomplished by resorting to commonly used methods such as ODEs, Boolean networks, and probabilistic graphical models including Bayesian

networks Kotiang and Eslami [2020], Friedman et al. [2000], Lovrics et al. [2014], Ay and Arnosti [2011]. Of these, ODEs are considered to be the most widespread approach in GRNs' modelling, due to their ability to provide a continuous description of molecular interactions and to quantify the impact of disruptions to gene expression levels in the system's output Ocone et al. [2015], Polynikis et al. [2009], Leal Valentim et al. [2015]. While the formulation for ODEs is challenging to solve analytically, numerical simulations can be approximated with readily available computational tools. On the other hand, PDE models, which involve multiple variables such as time and space, are theoretically complex and more computationally demanding Ay and Arnosti [2011].

In differential equations, the interactions between the molecules of interest are described through rate equations where the levels of each component are defined over time and/or space based on their concentrations, production, and degradation rates Ay and Arnosti [2011]. Hill functions are a type of mathematical function that can be used in the formulation of differential equations to model biological systems and describe the transcriptional regulation Hill [1910]. To accurately represent intrinsic interactions, these functions are highly non-linear, and with limited availability of analytical tools to employ, simplification methods, such as piecewise-linear approximations or discrete-time systems, have been proposed to overcome these limitations Polynikis et al. [2009].

To illustrate the process of modelling GRNs as a system of differential equations, we can consider a simple scenario of the production of a protein through transcription and translation Lopatkin and Collins [2020]. This process involves the transcription of DNA into messenger RNA (mRNA) and the subsequent translation of mRNA into protein. The rate of mRNA transcription is described by the equation:

$$\frac{d[mRNA]}{dt} = k_r - d_r[mRNA] \tag{2.1}$$

where k_r represents the rate of mRNA transcription, $[DNA]$ is the concentration of DNA, and d_r is the rate of mRNA degradation. The rate of protein production is described by the equation:

$$\frac{d[protein]}{dt} = k_p[mRNA] - d_p[protein] \tag{2.2}$$

where k_p is the rate of protein translation and d_p is the rate of protein degradation. These equations describe how mRNA and protein levels change over time Lopatkin and Collins [2020].

In this example, both mRNA and protein levels will eventually maintain a constant concentration, reaching a steady state. Since transcription typically occurs faster than translation, in case the rate of mRNA transcription is faster than the rate of protein translation, the mRNA level can be assumed to be at a steady state, which is described by:

$$[mRNA] = \frac{k_r}{d_r} \tag{2.3}$$

By substituting this representation into equation (2), we reach a simpler equation that describes the constitutive production of a protein:

$$\frac{d[protein]}{dt} = r_0 - d_p[protein] \tag{2.4}$$

where r_0 represents the lumped reaction rate for both transcription and translation, making it easier to understand and analyze the GRN. This can be represented by a simpler reaction network, where protein is produced at a constant rate r_0 and degraded at a rate d_p Lopatkin and Collins [2020].

Similarly, more complex molecular networks can be transformed into a set of differential equations, by considering different types of transcription regulation, such as activation or inhibition, and combinations thereof. In cases where transcription is proportional to protein levels, the concentration of a protein over time can be described by the activation and repression equation, respectively:

$$\frac{d[protein_X]}{dt} = r_0(\frac{[activator]^n}{K_A^n + [activator]^n}) - d[protein_X] \tag{2.5}$$

$$\frac{d[protein_X]}{dt} = r_0(\frac{K_R^n}{K_R^n + [repressor]^n}) - d[protein_X] \tag{2.6}$$

where r_0 is the lumped production rate, n is the Hill coefficient, $[activator]$ and $[repressor]$ are the activator and repressor concentrations, respectively, K_A and K_R are the half-maximal concentration constants, and d is the degradation rate of protein X. This equation describes how the concentration of protein X changes over time as a result of transcriptional regulation, which occurs either through constitutive production, activation, or inhibition Lopatkin and Collins [2020].

Accurate model parametrization is crucial in the discovery of biological relationships, as it enhances the interpretability of the model Lema-Perez et al. [2019]. The integration of Artificial Intelligence and dynamical modelling approaches can help improve the reliability of parameter estimation and simplify modelling processes through the extraction of biologically relevant parameters from large data sets. This technology provides an algorithmic understanding of the best approach to reduce the number of parameters necessary and improve the accuracy of estimates Lopatkin and Collins [2020], Thorn [2022]. As such, combining ML with mechanistic modelling techniques can improve the accuracy of context-appropriate parameter estimation by incorporating information obtained from both raw data and derived parameters into the modelling process.

2.4 Machine Learning In Mechanistic Modelling

2.4.1 Machine Learning Overview

ML is a technology that enables the development of predictive models through a training experience where useful patterns and insights are retrieved from data Camacho et al. [2018]. This field has expanded through a wide variety of industries in a short amount of time, due to its ability to predict real-world behaviour through the exposure of relevant input-output examples Jordan and Mitchell [2015]. To accurately perform this task, these algorithms require a substantial volume of information whose access has been made easier with recent data availability Baker et al. [2018].

Next-generation and high-throughput technologies have produced a large amount of biological data that seeks to quantify molecular variables and interactions Camacho et al. [2018]. However, the complexity of this data makes it difficult to differentiate between statistical significance and random variability in biological relationships Lopatkin and Collins [2020]. To address these challenges, the integration of ML approaches has become critical in the process of uncovering biologically relevant information through mechanistic modelling problems Camacho et al. [2018], Lopatkin and Collins [2020]. Moreover, ML provides a solution to balance accuracy and computational efficiency in the construction of data-driven models for complex physical systems Legaard et al. [2021]. Nevertheless, the majority of biological data is unlabeled as the functionality of many genes is still largely uncharacterized, making it challenging to solely rely on ML for further discovery.

The process underlying these algorithms (Figure 3) is started through the pre-processing of the input data. The given dataset is comprised of features, the measurements taken across the chosen set of samples, and labels, which are the model's output and what it aims to predict. Then the model starts the learning process of input-output behaviour by setting learning rules established through mathematical formulas and statistical assumptions. This trained model is used to make predictions on previously unseen data. The approach chosen to make new predictions depends on the prediction problem's nature and intrinsic data properties, which can either be labelled or unlabeled. If labelled, a supervised approach can be applied for regression or classification problems; whereas an unsupervised approach is required if unlabeled. To validate the predictions, the model needs to be exposed to new data to improve its performance and develop new hypotheses. The learning process optimizes the parameters of the model through a series of iterations, where parameters are estimated, model performance evaluated, and errors corrected. The process continues until the model's error is minimized, and performance cannot be improved further

Camacho et al. [2018].

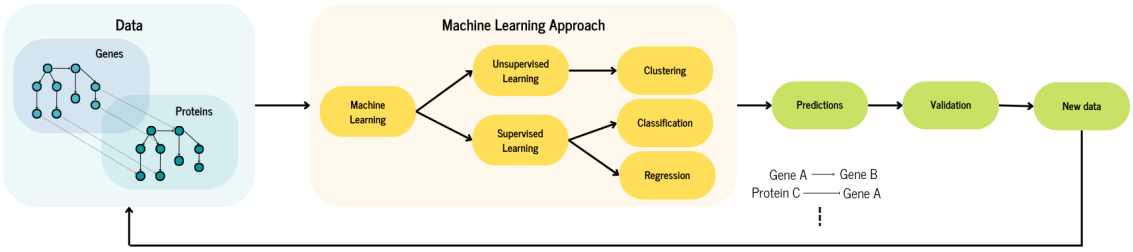


Figure 3: Machine-Learning workflow for the inference of regulatory interactions, adapted from Camacho et al. [2018]. A typical machine-learning workflow involves three key steps: data pre-processing, model training, and prediction. The input data comprises various features from multiple samples that can include gene expression levels, protein measurements, metabolite data, and gene interaction information. The choice of machine-learning approach depends on aspects such as the prediction task, data properties, and whether the data is labeled or unlabeled. Unlabeled data requires unsupervised methods like Principal Component Analysis (PCA) or hierarchical clustering, while labeled data allows for supervised approaches, enabling the creation of predictive models for regression or classification tasks. Post-training, validation and model refinement are essential steps to enhance prediction accuracy and generate novel biological insights.

Combining the strengths of both ML and mechanistic modelling holds great potential for improving our understanding of biological systems. While mechanistic models offer a comprehensive understanding of cause and effect, ML provides the capability to make predictions based on large, complex datasets. By combining the two, we can take advantage of the in-depth understanding of mechanistic models and ML's predictive power, leading to a more extensive understanding of biological processes Presnell and Alper [2019]. This synergistic approach can be integrated at various steps, including inference of gene regulatory networks, determination of context-specific parameters, and analysis of correlations to effectively combine highly correlated data components.

2.4.2 Inference of Gene Regulatory Modules

Network biology is a subfield of Systems Biology that aims to improve the understanding of living cells' underlying structures and functions by identifying and characterizing all biomolecules and their complex interactions. The nature of these interactions varies according to the involved cell components, including protein-protein interaction, metabolic, signalling, signal transduction, transcription, and gene regulatory networks Barabasi and Oltvai [2004], Camacho et al. [2018]. The basic elements of these networks can

be mathematically represented through graph theory where integral elements are represented by nodes that are connected by edges, indicating an existing interaction De Jong [2002]. The structure formed by the established edges is designated as the network's topology. This representation can then be transcribed to a matrix that reflects the presence (1) or absence (0) of an interaction, as represented in Figure 4Lucas et al. [2011].

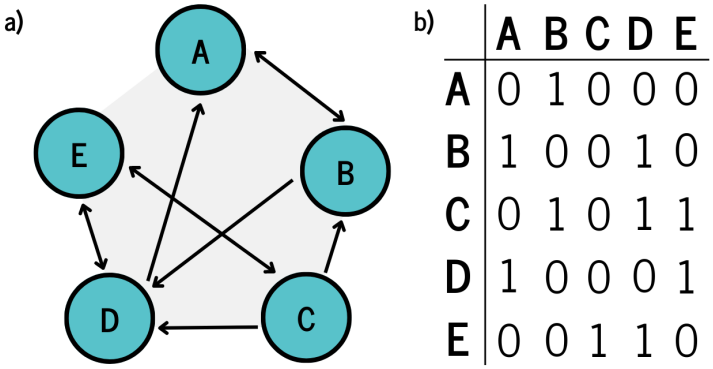


Figure 4: Mathematical representations of a regulatory network, through a) graph theory and b) matrix, adapted from Lucas et al. [2011]. A network (a) can be constructed to represent genes as nodes and their interactions as directed edges. This network can also be mathematically represented as a binary matrix (b), where '1' implies the presence of an interaction and '0' implies its absence, reflecting a gene-gene interaction topology. In this context, interactions represent the possibility of a direct influence from one gene to another, whether bidirectional (for mutual influence) or unidirectional (for one-way influence).

A major challenge associated with studying these networks is the quantitative understanding of their intrinsic topology and dynamics responsible for regulating cell behaviour under specific stimuli. Systems Biology approaches attempt to extract these constitutive elements and fit them into hypothetical functional networks that can characterize the dynamics of the regulatory interactions Cramer et al. [2011], Lucas et al. [2011], Barabasi and Oltvai [2004]. This process is defined as network inference and in the model plant *Arabidopsis thaliana* it is commonly performed through the analysis of co-expression data and comparison of gene expression profiles Wei et al. [2006], Lucas et al. [2011]. However, inferring biological networks is challenging due to the high number of possible interactions between elements, making it difficult to accurately predict these complex systems.

The process of inferring genetic networks can be accomplished through statistical methods. Weston et al. [2008] resorted to weighted co-expression analysis to successfully identify six genetic modules associated with the abiotic stress response in *Arabidopsis thaliana*. This method involves determining a similarity matrix of gene coexpression through Pearson correlation, transforming coexpression similarities

into connection strengths, and identifying network modules through clustering Zhang and Horvath [2005]. In contrast, Wang et al. [2022] developed a Random-Field Ising Model physics-based approach that identified disease modules through the score optimization of entire networks, instead of subgraphs, resorting to a graph theory optimization method, the max-flow algorithm. However, while statistical methods such as these have proven useful in constructing gene coexpression networks, they have limitations in inferring regulatory interactions and directionality, meaning that they cannot accurately simulate the dynamics of biological systems.

To overcome this, ML has become a crucial tool in uncovering underlying regulatory mechanisms, allowing for extensive data analysis to extract features of interest Hiscock [2019], De Smet and Marchal [2010], Lopatkin and Collins [2020]. In the last few years, various ML systems have been established for the inference of GRNs Camacho et al. [2018], Hiscock [2019]. For instance, the development of general-use platforms, such as GenePattern Reich et al. [2006], Kuehn et al. [2008] and CellNet Cahan et al. [2014], has provided researchers with access to a wide range of algorithmic techniques for the analysis of genomic data. For GRN inference, GenePattern Reich et al. [2006], Kuehn et al. [2008] offers two different algorithms, Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) Margolin et al. [2006] and Probabilistic Neural Networks (PNN). ARACNE is an unsupervised algorithm based on information theory that allows for the reverse engineering of transcriptional networks from microarray data. This method reported low prediction error regarding the identification of regulatory interaction in mammalian cells through the use of Hill kinetics Margolin et al. [2006]. Whereas PNNs are neural networks that estimate the probability of an unknown sample belonging to a set of known classes Specht [1990]. On the other hand, CellNet Cahan et al. [2014], is a platform focused on the network biology of human stem cells. For the reconstruction of GRNs, this platform developed a modified approach based on the Context Likelihood of Relatedness (CLR) algorithm Faith et al. [2007]. This unsupervised algorithm uses gene expression data of *Escherichia coli* measured across different conditions to identify transcriptional regulatory interactions with a 60% true positive rate Faith et al. [2007]. In addition to these platforms, there are also specific tools for GRN inference, such as Biological Network Integration using Convolutions (BIONIC) Forster et al. [2022] and Beacon GRN Ni et al. [2016]. BIONIC employs both deep learning and Graph Convolutional Network to extract features from the network structure. This approach has both unsupervised and semisupervised learning modes, the last one focusing on available gene function annotations when possible Forster et al. [2022]. In contrast, Beacon GRN is a supervised ML approach specifically designed for predicting GRNs for the model plant *Arabidopsis thaliana* in seed development, allowing for more accurate predictions in GRN inference for Arabidopsis Ni et al. [2016].

Weighted Gene Coexpression Network Analysis (WGCNA), on the other hand, is a systems biology approach that makes use of high-throughput microarray or RNA-seq data to provide insights on genes' functionality at a system's level. For this reason, this method is commonly used to identify and group genes that are co-expressed across samples into modules Zhang and Horvath [2005], Pei et al. [2017]. WGCNA R was originally developed in R environment as a software package comprising a comprehensive set of functions that enable its users to perform different steps of the WGCNA, such as network construction, gene selection, gene module detection and inference of topological characteristics Langfelder and Horvath [2008]. However, the R implementation has been reported to have downfalls associated with a slow performance and non-scalability Rezaie et al. [2023]. These setbacks become increasingly difficult to ignore as biological datasets become larger and more complex. Additionally, the package does not allow for the comparison of co-expression modules extracted from different WGCNA networks. PyWGCNA was developed with the aim to aid in the identification of co-expressed gene modules from large RNA-seq data natively in Python. The new and reportedly improved implementation performs similar tasks to WGCNA R with a added benefit of having a faster and more scalable implementation and new functions for improved downstream analysis associated with Functional Enrichment Analysis (FEA), inter-module analysis of protein-protein interactions and comparison of co-expression modules extracted from different WGCNA objects Rezaie et al. [2023]. The addiction of the latter, facilitates the evaluation of module reproducibility.

Despite the wide tool availability, the direct integration in network inference problems can be challenging as these methods are often tailored to their context, for instance, organism or condition-specific, with no clear guidelines for the choice of the most fitting model De Smet and Marchal [2010]. However, the suitability of some of these techniques can still be highlighted. ARACNE Margolin et al. [2006], for example, is an unsupervised method with low error rates but it cannot infer the directionality in gene regulation Sokolovska et al. [2018]. On the other hand, CLR has been reported to outperform ARACNE and other methodologies, but its unsupervised nature does not allow for the integration of labelled data Faith et al. [2007]. BIONIC, however, has a semi-supervised approach that enables the integration of labelled data when available Forster et al. [2022]. This method also comprises several features that increase the accuracy of network inference, including automatic input weighing, the consideration of residual connections, and the ability to learn gene features from different input networks based on multiple neighbourhood sizes, which reduces vanishing gradients in the training process Forster et al. [2022].

Selecting the most appropriate model involves requires a multi-step approach, involving training several relevant approaches, comparing their performance, and evaluating their scores to identify the best-performing method. Once identified, its parameters are fine-tuned to ensure that it fits the data accurately

and predicts the underlying biological subsystems. The obtained modules are then analyzed to assess their plausibility. For this validation step, the inferred GRNs' modules are measured up to observed data, available mechanistic information, gene ontology, and annotation Hiscock [2019]. The extracted modules can also be validated experimentally (1) *in silico*, through ODE modelling, (2) *in vivo*, by ChIP-seq assays, and (3) *in planta*, though the analysis of mutant phenotypes Haque et al. [2019].

The challenge in genome-wide RNA expression analysis has shifted from acquiring expression profiles to extracting meaningful biological insights from these analyses Subramanian et al. [2005]. Following a modular evaluation approach, which focuses on understanding biological systems and networks through groups of genes rather than individual genes, reduces the complexity of the statistical problem and facilitates the classification of genes into biologically relevant functional clusters Ackermann and Strimmer [2009], Wang et al. [2008]. Enter Gene Set Enrichment Analysis (GSEA), a technique developed to assess microarray data at a modular level Subramanian et al. [2005]. This method empowers the identification of gene clusters associated with diseases or, in the context of this project, their activation under particular conditions Wang et al. [2008], Ackermann and Strimmer [2009], Hung et al. [2012]. Consequently, enrichment analysis plays a crucial role in the identification of biological pathways contributing to specific traits based on their functionality Ackermann and Strimmer [2009].

2.4.3 Integration of Machine Learning and Mechanistic Modelling

The integration of gene network information is critical for understanding the underlying mechanisms that regulate cellular behaviour. As the amount of biological data being collected and made publicly available grows at a rapid rate, the application of ML in data integration for mechanistic modelling is becoming increasingly critical for effectively extracting meaningful insights. This can be achieved through processes such as predicting the association between gene modules and model components and estimating context-specific parameters. However, while data-driven approaches, such as statistical analysis and unsupervised methods, play an important role in facilitating the analysis of complex biological data, it is important to also consider incorporating knowledge about regulatory interactions through knowledge-based approaches. The integration of both approaches allows for a more comprehensive understanding of the system and an increase in the accuracy and robustness of the predicted networks Noor et al. [2019], Volk et al. [2020]. Some key highlights of this synergy are visually depicted in Figure 5.

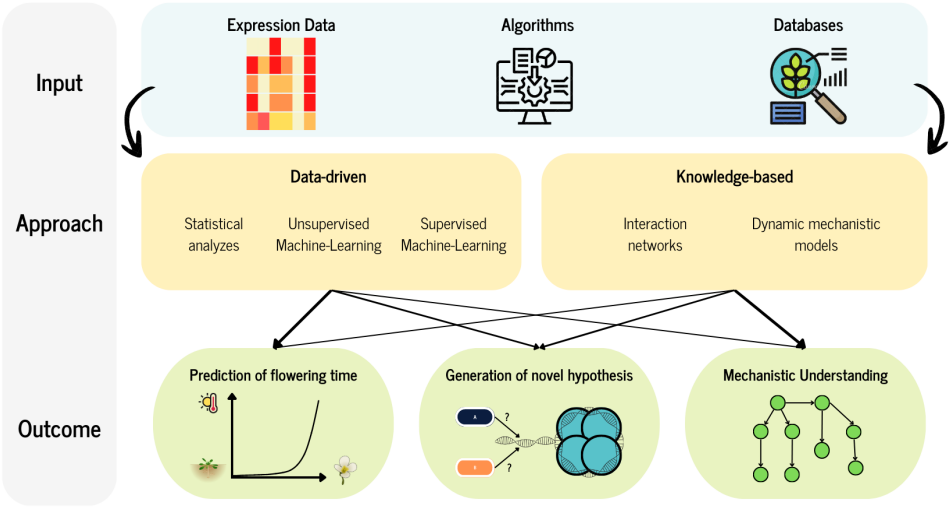


Figure 5: Illustration representing diverse input data sources and the contrasting outcomes achieved through the use of data-driven and knowledge-driven approaches in data integration, adapted from Noor et al. [2019]. This scheme captures how employing distinct approaches can lead to different outcomes, ranging from prediction of relevant variables to the generation of novel hypotheses and mechanistic insights. Overall, this depiction highlights the significance of integrating both approaches to deepen our understanding of biological systems.

Chapter 3

Methodology

Chapter 3 provides a comprehensive overview of the methodology employed throughout the project. It is structured into four distinct subsections that explore the data collection process and employed algorithms: Biological Network Integration using Convolutions (BIONIC), Python-based Weighted Gene Correlation Network Analysis (PyWGCNA), and Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN).

The chapter starts by exploring the data collection step, elucidating the systematic approach undertaken to search public databases on datasets that capture the dynamic expression levels of *Arabidopsis thaliana* in response to varying ambient temperatures. This subsection will also delve into the key characteristics of the datasets selected for further analysis and the selection process itself.

The core focus of this chapter centers on the concept, design, and implementation of the ML-based approaches employed to identify gene modules associated with the induction of flowering time in response to different temperatures. This section offers a thorough breakdown of the algorithms applied, providing a clear understanding of how these approaches were used to extract valuable insights. Furthermore, all additional changes and scripts written for this task will be covered.

3.1 Data Collection

The data collection step plays a crucial role in the extraction of gene modules that accurately describe the gene expression dynamics across varying temperature conditions. This importance arises from the fact that the environmental conditions and variables under which gene expression levels were measured significantly influence the final gene modules and the biological insights derived from the analysis.

For this project, two main public databases were explored: ArrayExpress EMBL-EBI [2023] and GEO NCBI [2023]. To reduce the research space and increase the effectiveness of the search process, several desired data characteristics were defined and data filters were applied. For the GEO database, the

following search query was used: "arabidopsis thaliana"[Organism] AND ("expression profiling by high throughput sequencing"[DataSet Type] OR "expression profiling by array"[DataSet Type]) AND ("temperature"[All fields] AND "time"[All fields]) AND "Flowering time"[All fields]". Whereas, for the ArrayExpress database, *Arabidospsis thaliana* was selected as our desired organism and the keywords "flowering time temperature" were used.

A meticulous review of these searches yielded around 80 studies that held varying levels of relevance to the research objectives. These studies complied in an online `xlsx` file that organized data according to attributes such as 'Study name', 'Link to paper', 'Link to data', 'Expression analysis', 'Genotype', 'Description', 'Age', 'Tissue', 'Light', 'Time series', 'Temperature' and 'Ambient'. The file can be accessed through the URL: `https://1drv.ms/x/s!AmiW-IFDRrFdtgTEn5O6pj9DWMnu?e=XZYxfi`. Upon careful consideration, a strategic decision was made to work with two specific datasets, the E-MEXP-728 Balasubramanian et al. [2006] and the E-MEXP-1304 Michael et al. [2008]. The contents of Table 1 provide a comprehensive overview of the specific experimental conditions underlying the datasets selected for this study. This choice was driven by the fact that neither dataset possessed all the envisioned qualities for the analysis. Both datasets, however, shared key attributes such as non-stressful temperature conditions and a high-resolution collection of time points. Nevertheless, one of the datasets (E-MEXP-1304) lacked crucial information on flowering time, while the other (E-MEXP-728) introduced an additional variable, photoperiod changes. Consequently, the utilization of both datasets was deemed essential to facilitate a comprehensive comparison of the gene modules extracted throughout the project.

3.2 BIONIC

Network integration plays a critical role on deepening our understanding of biological systems by elucidating how an organism forms an integrated response to its dynamic surroundings. This response translates to the organism phenotype and it is built upon various interacting networks that govern responses, such as signalling, regulation and metabolic adaptation. Thus, the integration of multiple biological networks can provide a more comprehensive understanding of how organisms adapt and respond to environmental cues, supporting the mapping of cellular functions Tenazinha and Vinga [2010], Forster et al. [2022]. These networks can be constructed from various data types, such as gene expression data, metabolomic data, protein-protein interactions, genetic interactions, or other relevant biological data. However, conventional network integration methods face constraints, such as the inability to handle diverse networks and network overlap, or they present limitations in scalability.

Table 1: Overview of the experimental conditions within the chosen datasets, E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008]

	E-MEXP-728	E-MEXP-1304
Ecotype	Col-0	Col-0
Age	35 days	7 days
Tissue	Shoot Apex	Seedling
Light	SD, LD	Constant Light
Temperature	16°C, 25°C	12°C, 22°C
Time Series	0, 2, 5 and 9 days	0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40 and 44 hours

The experimental conditions for both the selected datasets, E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008], are summarized in this table. Short Day (SD) corresponds to an 8-hour light and 16-hour dark photoperiod cycle, while Long Day (LD) represents a 16-hour light and 8-hour dark cycle. The E-MEXP-218 experiment involved transitions from SD, 16°C, to LD, 25°C, alongside temperature variations; whereas E-MEXP-1304 was conducted under constant light, with temperature changes occurring in 12-hour cycles (12°C/12 hours, 22°C/12 hours). Both experiments utilized the Affymetrix ATH1 Genechip as a microarray technology.

BIONIC addresses these existing issues in network integration through its general and scalable Deep Learning framework for network integration. The algorithm resorts to a GCN architeture to learn gene features across multiple input networks that reportedly contain more functional information. As a matter of fact, its main claim is that the features extracted from the algorithm are able to more accurately represent the topology of the original network and, as such, have an improved performance for downstream tasks, including the prediction of gene modules. BIONIC displays the flexibility of working with both unsupervised and semi-supervised approaches, allowing the user more adaptability based on the availability and nature of the data Forster et al. [2022]. All scripts developed and the results produced using this approach can be found in a Github repository, accessible through the following url: `https://github.com/camilababo/module-dection`.

3.2.1 BIONIC's Design and Implementation

BIONIC is a comprehensive package, organized into several key components, each serving a unique role in network integration. These components include configuration and input files, test functions, utility functions, command-line interface usage, outputs, the model's build, and the training process. This section will briefly detail the main characteristics of the model's build and its training process. The code is available in the developer's Github repository, accessible through the following url: `https://github.com/bowang-lab/BIONIC`. In Figure 6, it is possible to observe a general schematic overview of BIONIC's unsupervised algorithm.

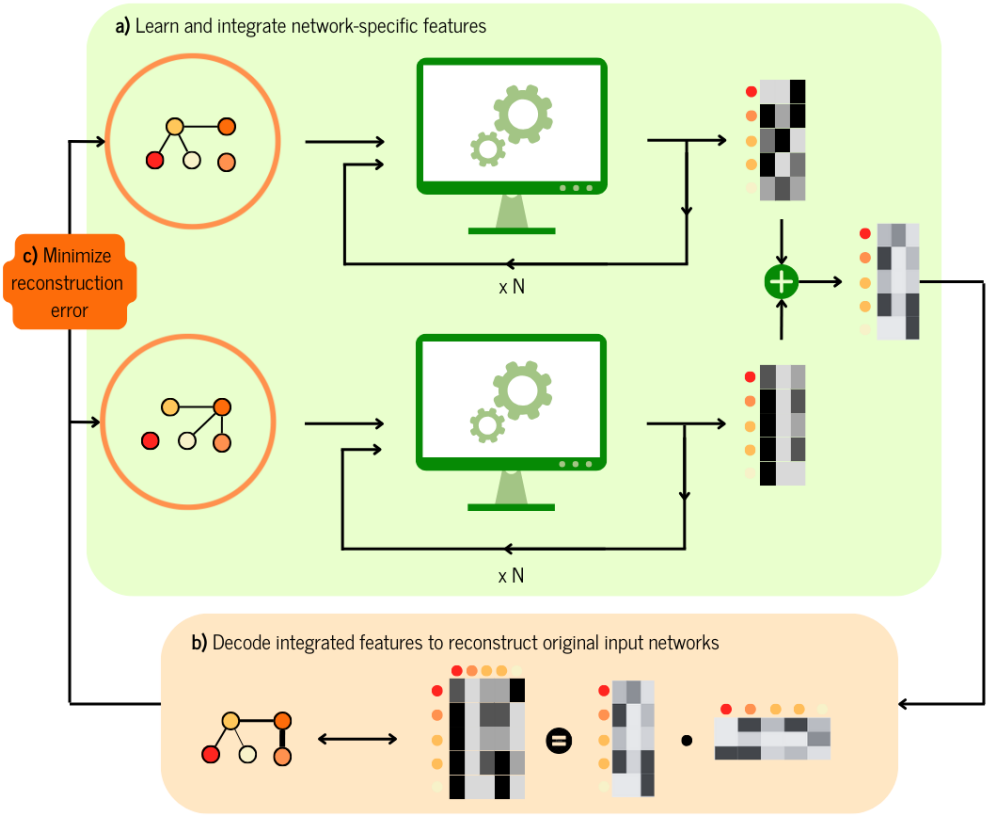


Figure 6: Overview of Biological Network Integration using Convolutions (BIONIC)'s unsupervised workflow, adapted from Forster et al. [2022]. The workflow comprehends three main procedures: (a) learning and integrating the network-specific features, (b) decoding the integrated features to assess the reconstruction back to the original input networks, and (c) the minimization of the reconstruction error.

BIONIC's GCN architecture is developed on a PyTorch-based neural network framework and it was implemented to work with graph data, enabling feature learning and integration across multiple input networks. Within the model component, there are three modules for each main element, the model, the

layers and the loss functions. The `model` module contains three classes `BionicEnconder`, `Bionic` and `BionicParallel`. The `BionicEnconder` class enables feature extraction for the individual networks, taking a network as input and returning a feature matrix. One single encoder takes on multiple parameter-sharing layers of a GCN variant, named Graph Attention Network (GAT). While the original GAT framework operates under the assumption of binary network inputs, in the context of BIONIC, a modification was introduced to account for a priori network edge weights. This approach was chosen as the GAT architecture repeatedly showcases a good performance for node classification tasks due to its ability to learn alternative edge weights, enabling the model to increase or decrease the weight of an edge based on its importance in the reconstruction task Forbes et al. [1992]. Each layers represents the number of directly connect genes to consider, meaning that one layer represents the immediate neighbourhood, a second layer the second-order neighborhood and so on, facilitating the capture of local and global network properties.

The `Bionic` class serves as the core component of the model, playing a central role in handling the integration of multiple encoder modules, each corresponding to a distinct input network. The integration process incorporates an embedding step that captures vital information while simultaneously addressing computational efficiency. It achieves this by mapping the integrated features into a lower-dimensional space, employing advanced dimensionality reduction techniques to optimize the computational complexity of the model. Parallel processing was not employed, however, the `BionicParallel` efficiently distributes encoder modules across GPUs, enabling faster training and paralleled execution.

The `layers` module contains two classes, `WGATConv` and `Interp`. The `WGATConv` class builds upon the foundation of the original GAT from the PyTorch Geometric framework. By introducing the ability to accommodate edge weights, it profoundly influences the message-passing process. In this modified approach, attention coefficients take into account these edge weights, allowing for a more detailed information flow within the network. This improved consideration of edge weights leads to a thoughtful aggregation of results from each GAT pass, ultimately giving rise to the final network-specific features. This process is orchestrated by the `Interp` class, which employs a stochastic summation layer. This layer takes on the role of introducing randomness by calculating mask vectors. These vectors, in turn, determine the selection of random nodes contributing to the combined features. This strategic introduction of randomness, embodied by the notion of "dropout", serves as a crucial mechanism for encouraging the model to assimilate cross-network dependencies. By design, it forces the model to rely on feature vectors from other networks, strategically filling the gaps created by the dropped vectors. This module also contains a stand-alone function called `weighted_softmax`. This function extends the PyTorch Geometric softmax

functionality to integrate edge weights into the optimization process. These features are central to the optimization process, collectively ensuring an efficient training process. Together, they enhance the model's capacity to assimilate and learn the intricate relationships that span across multiple input networks.

Within the `lossmodule`, there are two custom loss functions: `masked_scaled_mse` and `classification_loss`, each with distinct roles. The former is harnessed during unsupervised tasks, where it calculates the Mean Squared Error (MSE) between the reconstructed network and the original input network. This measure assesses the accuracy of the model's reconstruction to the actual network structure, a key aspect of the learning process. In contrast, the latter emerges during semi-supervised tasks, where it tackles the computation of a multi-label classification loss. To accomplish this, it harnesses the PyTorch-native `BCEWithLogitsLoss` function, combining a *Sigmoid* layer with the `BCELoss` function (Binary Cross Entropy (BCE)).

Lastly, the `train` module is responsible for the entire training process and and is designed to receive input in the form of a configuration file in JavaScript Object Notation (JSON) format. This file empowers users to tailor the training process to their specific needs, providing the ability to specify the network files that should be considered, along with a range of training-specific parameters, including number of training epochs, batch size, learning rate, and the shape of the GAT layers and embedding vector. Upon initialization, the `train` module orchestrates the entire training pipeline, which comprises data preprocessing, model setup, training, and the underlying forward pass logic. During the training phase, the forward pass is executed using stochastic optimization techniques. It operates under the guidance of the Adam optimizer, dedicated to minimizing the model's reconstruction loss. The scalability of the algorithm is enhanced through mini-batching processing of genes. This technique entails sampling and integrating subsets of networks at each training step. Not only does this approach optimize memory usage, but it also enhances the overall runtime, boosting the algorithm's efficiency.

3.2.2 Data Pre-processing

For this project, the gene expression data was divided in two separate Comma-Separated Values (CSV) files, each corresponding to different temperature conditions. Thus, one file comprised expression profiles measured under low temperature conditions, and the other encompassed those recorded under high temperature. This division was performed so that two separate networks could be used as input for the BIONIC model. This approach aimed to enhance the algorithm's capacity to potentially discern and encode temperature-specific patterns and features in the data. As previously stated, BIONIC takes in gene interaction networks in edgelist format as input. This implies that gene expression data needs to be

fittingly converted to the correct format. For an edgelist file, this means that each line contains two gene identifiers and, optionally, the weight of the edge that connects both. BIONIC accepts gene interaction networks without edge weights, considering them binary, that is, a pair of gene is or is not connected. However, the inclusion of edge weights allows us to more accurately inform the model on the strength of the functional relationship between gene pairs Forster et al. [2022].

A script was developed for the purpose of pre-processing and transforming the gene expression data into a network-ready edgelist format. It comprises two functions, `pre_process_data` and `data_to_edgelist`, that offer flexible parameterization for a streamlined command-line execution. The `pre_process_data` function removes irrelevant probes, such as control and organ-specific probes, entries with missing values, and applies a log2 transformation. The pre-processed data is then saved as a new CSV file. The `data_to_edgelist` function takes the pre-processed data and constructs a network using a radius-based approach from the `sklearn` library, the `neighbors.radius_neighbors_graph`. This method creates a sparse, weighted graph by considering the nearest neighbors of data points within a specified radius. By default, it calculates distances using the Euclidean metric and within a 0.5 radius. This approach efficiently captures potential gene interactions from high-dimensional data while limiting computational complexity due to its sparse graph nature. It then resorts to NetworkX, a Python library implemented to sustain data structures and methods for storing graphs, to generate a network object. This function also provides the option to include all nodes found in a user-specified node list file. This inclusion ensures that nodes specifically relevant to the project, which might be filtered out during the radius-based method, are preserved and accounted for in the resulting network. The final network is saved as a weighted edgelist file.

3.2.3 Module Identification and Evaluation

The developer of the BIONIC algorithm also implemented a second package named BIONIC-evals, with the intent to facilitate the reproduction of the evaluation of co-annotation prediction, module detection, and gene function prediction tasks, as detailed in the original research article Forster et al. [2022]. The library evaluates the performance of the topological feature matrix extracted from BIONIC alongside the original input networks in the specified evaluation tasks, using established standards as benchmarks. This enables users to accurately determine the model's effectiveness by assessing whether the extracted features genuinely outperform the original network representations. The standards used for evaluation of the module detection task comprise a JSON file where each key correspond to a unique cluster identifier, and the corresponding values for each key are a list of gene identifiers that make up that specific cluster.

Similarly to the BIONIC functions, the different evaluations scenarios can be executed by passing in a JSON configuration file. The file allows the user to specify of the path to input networks, extracted feature matrix, evaluation standards and task-specific parameters such as number of samples, evaluation metrics and linkage methods. For the evaluation of the module detection task, the methods starts by importing the predefined module standards from the input JSON file and reducing them to include only the genes that are common to both the standards and the data at hand, to prevent overlap. This step is essential for calculating evaluation metrics, such as Adjusted Mutual Information (AMI) or Adjusted Rand Index (ARI). These metrics measure the quality of clustering results, with ARI, an adjusted variant of Mutual Information, assessing the similarity between predicted clusters and true labels; and ARI, an adjusted variant of the Rand Index, adjusting the score for agreement occurred by chance between predicted clusters and true labels when evaluating clustering results. ARI is preferred when the ground truth consists of large, equally-sized clusters, while AMI is more suitable for unbalanced scenarios with small clusters Romano et al. [2016].

The evaluation process, for both features and networks, iteratively explores the various combinations of specified clustering methods and distance metrics outlined in the configuration file. For each combination, it calculates the pairwise distances between features or edges in the network using the chosen metric and proceeds with hierarchical clustering, determined by the specified method and threshold values. This hierarchical clustering results in cluster labels, which are then compared to the ground truth labels using the chosen evaluation metric. The goal is to identify the combination of parameters that yields the highest score, indicating the optimal clustering solution for the given data. Additionally, BIONIC-eval incorporates a visualization component to facilitate an intuitive interpretation of the evaluation results. The package is available in the developer's Github repository, accessible through the following url: `https://github.com/duncster94/BIONIC-evals`.

3.3 PyWGCNA

As described in previous sections, PyWGCNA stands out as a powerful tool for the identification of co-expressed gene modules from large RNA-seq datasets. In contrast to the original WGCNA implementation in R, PyWGCNA offers a more efficient and scalable solution, addressing previous issues of slow performance and non-scalability. Moreover, PyWGCNA has enhanced functions for downstream analysis, introducing features like FEA, inter-module analysis of protein-protein interactions, and the ability to compare co-expression modules across different WGCNA networks, thus facilitating module reproducibility

evaluation. All scripts developed and the results produced using this approach can be found in a Github repository, accessible through the following url: https://github.com/camilababo/pywgcna.

In Figure 7 it is possible to observe a schematic overview of the workflow used with the PyWGCNA package. The workflow can be divided in three different main operations regarding the identification, analysis and comparison of co-expressed modules. To carry out these operations, a Jupyter Notebook was created and structured to include dedicated sections for each analysis step corresponding to the chosen datasets.

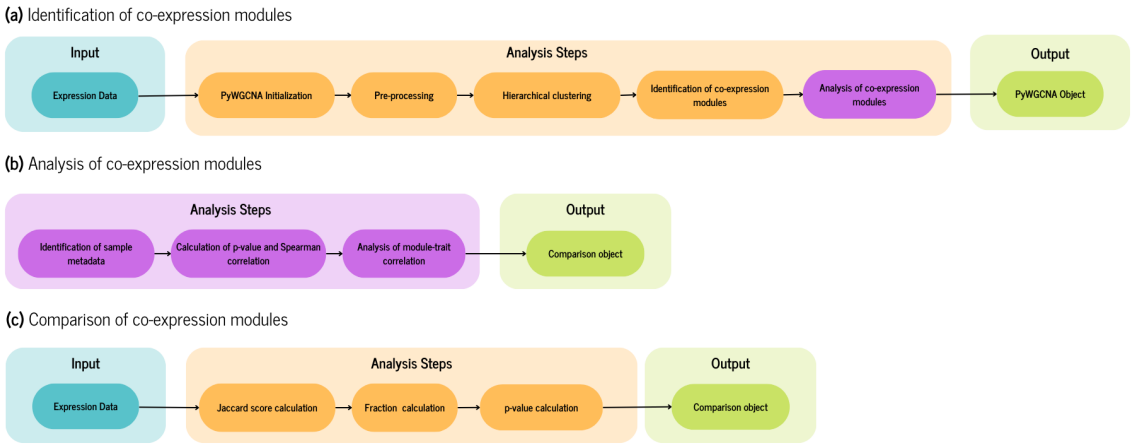


Figure 7: Schematic overview of Python-based Weighted Gene Correlation Network Analysis (PyWGCNA)'s established workflow, adapted from Rezaie et al. [2023]. The workflow comprehends three main procedures: (a) the identification of the co-expression modules, (b) analysis of the predicted modules, and (c) the comparison of the extracted co-expression modules.

3.3.1 PyWGCNA's Design and Implementation

The package is organized in four modules: `PyWGCNA.geneExp`, `PyWGCNA.wgcna`, `PyWGCNA.comparison`, and `PyWGCNA.utils`. The code is available in the author's Github repository, accessible through the following url: https://github.com/mortazavilab/PyWGCNA.

The module `PyWGCNA.geneExp` was implemented to facilitate the manipulation of gene expression data in the WGCNA framework. Its core component, a class designated as `GeneExp`, comprehends various methods that allow for the preparation, creation and modification of gene expression data stored in Anndata objects, described later in the sub-section 3.3.2. The module is structured to provide users with a clear and structured approach to tailor their data for subsequent downstream analysis. It enables users to specify input data sources, formats, species, and data levels (gene or transcript, with 'gene' as

the default) when initializing the class. Moreover, it offers methods for updating both gene and sample metadata within the Anndata object.

The module `PyWGCNA.wgcna` comprehends the entire framework for conducting the WGCNA analysis. It was implemented to simplify the entire process, offering a wide-range of usefull customizable functions that facilitate the execution of the co-expression network analysis. The class automates critical steps like pre-processing of the PyWGCNA object, hierarchical clustering, Topological Overlap Matrix (TOM) (further detailed in Subsection 3.3.3) and adjacency matrix calculation, making WGCNA accessible to researchers with varying levels of expertise. Users can specify parameters such as module size, value of the soft power threshold, network and TOM type ("unsigned", "signed" and "signed hybrid", with the last being the default option), tailoring the analysis to their specific research needs. Additionally, the class provides methods for analyzing module-trait relationships, plotting module eigengene heatmaps, performing FEA, and retrieving gene lists associated with specific modules.

The `PyWGCNA.comparison` was developed to enable the comparison of gene modules across multiple PyWGCNA objects in a versatile manner. It offers functions to calculate Jaccard similarity, p-values, and common fractions, further detailed in the section 3.3.5, that enhance the assessment of gene module relationships efficiently by providing quantitative insights into module relationships. This class also includes visualization functions for an eased interpretation of the comparison metrics, such as bubble plots, heatmaps, and network representations.

At last, the `PyWGCNA.utils` module offers a set of helper functions to enhance the functionality of the package. It includes methods for comparing WGCNA results with gene markers from single-cell experiments, retrieving gene mapping information from Biomart datasets, and reading saved PyWGCNA objects. These utilities streamline steps of data integration, annotation, and analysis process.

3.3.2 Data Pre-processing and Initialization of the PyWGCNA Object

For the PyWGCNA package, gene expression input data is expected to be organized in a tabular format, with rows representing specific samples and columns corresponding to unique genes. It is then stored in AnnData format, where it is also possible to integrate gene and sample metadata. Anndata is a Python package created to serve as a bridge between pandas and xarray to improve efficiency in managing annotated data matrices in memory. Its ability to organize and store learned annotations in an on-disk format also simplifies the sharing of data and analysis results Virshup et al. [2021]. From this, a PyWGCNA object is initiated in order to store user-defined parameters crucial for the network analysis, such as minimum module size, cut-off for gene expression level and network type. The use of the `.preprocess()` function

allows for data cleaning and pre-processing by removing low-expression genes based on a user-defined threshold and outliers based on hierarchical clustering Rezaie et al. [2023].

3.3.3 Module Identification

In Figure 8, we can observe a schematic summary of the gene clustering workflow in PyWGCNA that leads to module identification.

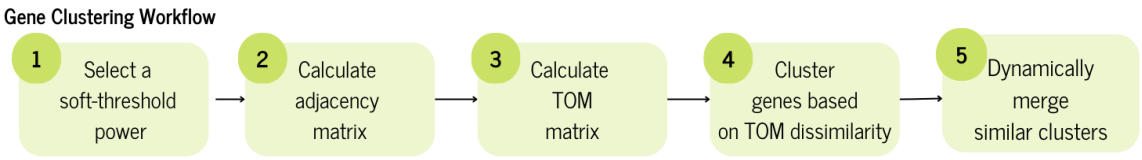


Figure 8: Schematic overview of Python-based Weighted Gene Correlation Network Analysis (PyWGCNA)'s gene clustering workflow, adapted from Rezaie et al. [2023]. The identification of gene modules requires several steps, such as the selection of an appropriate power for soft-thresholding the analysis of the network topology, computation of the adjacency matrix and calculation of the Topological Overlap Matrix (TOM). Genes are then clustered based on TOM dissimilarity and dynamically merged with similar clusters.

PyWGCNA first takes the preprocessed expression data and constructs an adjacency matrix by calculating the correlation between each pair of genes that represents the connection strength between them. This adjacency matrix is constructed by selecting a soft-thresholding power Rezaie et al. [2023]. More specifically, the approach performs an analysis regarding the scale free topology of the network for multiple soft-thresholding powers. A network is characterized as scale free when the distribution of links to nodes follows a power law. By other words, this means that the growth of the network entails the addition of new nodes to already established ones, which results in the emergence of few highly connected nodes, assigned as hubs. As such, the majority of nodes has small amount of edges, while few nodes have a large amount of connections Zhang and Horvath [2005], Barabási [2009]. In Figure 9 it is possible to observe the visual difference between a random and a scale-free network.

The final network is produced by computing the TOM. The TOM is commonly used in biological networks to provide a measure of similarity by reflecting the relative interconnectedness of two nodes Ravasz et al. [2002], Ye and Godzik [2004]. The calculation of the topological overlap between a pair of genes is based in the comparison of the connections that both genes have with all other genes in the networks Yip and Horvath [2007]. Previous reports declare that with regards to clustering gene expression pro-

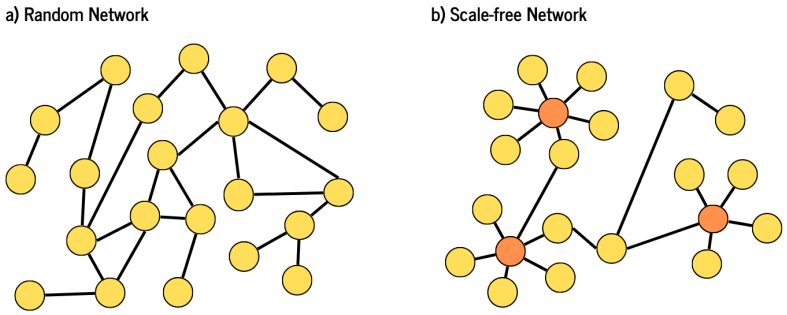


Figure 9: Visual representation and comparison of a (a) random network and (b) scale-free network. In the latter (b), it is possible to observe the presence of highly connected nodes, commonly designated as hubs.

files, the use of topological overlap as a dissimilarity measure contributes to more distinct and biologically meaningful modules Ravasz et al. [2002], Zhang and Horvath [2005].

As previously stated, modules are defined as groups of genes with highly-correlated gene expression profiles across multiple samples. In the PyWGCNA package, these modules correspond to branches of a hierarchical clustering tree that has been cut through a variant of the dynamic tree cut approach, 'Dynamic Hybrid'. According to the paper that reported its original implementation in R, the dynamic hybrid-tree cut algorithm enhances the detection of outliers within each cluster by employing a bottom-up approach. The algorithm can be described in two steps: firstly, initial clusters are assigned based on specific criteria such as, minimum number of objects (genes), distinctness when compared to neighboring clusters, and tightly connect cored; secondly, unassigned objects are assessed for their proximity to these initial clusters and subsequently assigned if deemed sufficiently close Langfelder et al. [2008].

3.3.4 Downstream Analysis and Module Visualization

Following the identification of the co-expression modules it becomes crucial to better understand the associations between the grouped genes and their biological significance. To achieve this, the PyWGCNA package includes several different downstream analysis options, such as: (1) find top hub genes within a module; (2) compute modules' eigengene and summarize its expression across sample traits; (3) calculate intramodular genes' connectivity; (4) infer the relationship between a module and a data trait; (5) retrieve known and predicted protein-protein interactions by assessing the STRING database; and (6) perform FEA according to databases as GO, KEGG and REACTOME via GSEApy and BioMart. For analysis of this character, the module eigengene serves as an effective representation of modules' expression profile and it

can be defined as the first principal component of a particular module Langfelder and Horvath [2008].

To better assess if the extracted modules were biological meaningful, two main downstream analysis were performed for this project: (1) the inference of the module-trait relationship and (2) GSEA through the use of gene ontology. For the module-trait relationship, the PyWGCNA package calculates the Pearson Correlation Coefficient and the p-value between the modules' eigengene vector and the traits that were identified from the dataset metadata. The values can then be plotted as a heatmap that allows the user to easily infer if the correlation between a given module and a specific trait is strong or weak. This is an especially fitting analysis to the project's scope as it allows to assess if the extracted modules have a high or low correlation with relevant traits, such as temperature and flowering time. The heatmaps representing the module-trait relationship can be observed in Section 4.1.2 of Chapter 4.

As previously mentioned, the PyWGCNA package includes a built-in function to perform FEA, however several issues arose with the employment of functions needed to map gene IDs to gene names across the different databases, as code implementation was focused on vertebrate genomes. Despite the initial effort to modify the variables regarding species, dataset and server domain information in the source code for the appropriate plant species, no enrichment terms were found for a variety of p-values when using PyWGCNA's internal methods. On the contrary, enrichment terms were found when opting for manual input into the 'Term Enrichment Service' of the AmiGO2 website, which is powered by Protein ANalysis THrough Evolutionary Relationships (PANTHER) Carbon et al. [2009], Mi et al. [2021].

PANTHER is a tool designed to facilitate high-throughput analysis, performing functional classification by combining information on gene function, gene ontology and pathways. For this project, the 'Statistical Overrepresentation Test' was employed. This analysis requires users to submit a list of genes to be compared against a reference gene list with the goal to determine whether a specific class (in this case, 'biological process') is statistically overrepresented or underrepresented. This evaluation relies on a statistical analysis, with Fisher's exact test selected as the method of choice Mi et al. [2013].

To facilitate the visualization of the functional enrichment terms, a small function was implemented to produce a bubble plot to represent the different biological processes that were present in each module, where the bubble size represents the fold enrichment value and the bubble color is based on the p-value. When available, fold enrichment values and p-values were extracted for the following biological processes: 'response to abiotic stimulus', 'response to stimulus', 'cellular response to stimulus', 'cellular response to abiotic stimulus', 'cellular response to environmental stimulus', 'response to cold', 'response to temperature stimulus', 'cold acclimation', 'cellular response to heat', 'response to heat', 'regulation of photoperiodism, flowering', and 'regulation of timing of transition from vegetative to reproductive phase'.

This extraction was carried out to assess the modules' representation of these processes, which align with the project's scope. Specifically, fold enrichment values were collected for relevant biological processes if they exceeded a value of one, indicating an overrepresentation of these processes in the gene set. For biological processes with no direct relevance to this project, fold enrichment values above 50 were selectively extracted. The results of this analysis are documented on Section 4.3, present in Chapter 4.

3.3.5 Comparison of Modules

Another critical analysis involves the comparison of extracted modules across multiple different PyWGCNA objects. This allows users to extract modules from different expression datasets and evaluate the level of agreement in these modules. Typically, data is collected under various environmental conditions, such as different temperature ranges, light conditions, flowering time, and time points. By assessing if the algorithm groups similar groups of genes together into similar modules, we can determine if there are consistent expression patterns across diverse conditions.

This analysis is performed by computing similarity measures, such as the Jaccard similarity coefficient and the common fraction of genes. For this project, the similarity between modules was mainly evaluated by resorting to the Jaccard coefficient. This measure is commonly used in the comparison of similarity or dissimilarity regarding the distance of two data points and its calculation is based on the division of the number of objects in both clusters by the number of objects in either cluster Niwattanakul et al. [2013]. Furthermore, to determine statistical significance, p-values were calculated using Fisher's exact test Rezaie et al. [2023]. Similar to previous analysis, the PyWGCNA package also includes plotting functions to ease interpretation. The comparison matrices can be plotted as a network or bubble map, with the first option being chosen to display module similarity results for this project. This representation showcases each module as a node, colored according to the object it belongs. If a Jaccard coefficient has a value above a chosen threshold, an edge is established between the nodes that represent the two given modules, with its Jaccard index printed on top. The network representations of these comparisons are available in Appendix A and will be later detailed in Chapter 4.

Originally, the implementation of PyWGCNA allows for the comparison of PyWGCNA object and an external gene list. While in theory this method would be greatly convenient to compare the modules extracted with the PyWGCNA algorithm and the modules extracted from other tested algorithms, in practice this method showed itself impractical as it did not allow to filter out non-overlapping genes in both modules. The ability to perform the analysis with common genes only can increase interpretability of the modules' similarity with a focus on shared characteristics. As such, a script was created to allow the comparison of

modules identified by the different algorithms that were tested in this project.

The analytical functions, mainly the calculation of the Jaccard similarity score and common fraction, computation of the Jaccard similarity matrix and the respective plotting, were largely derived from the original PyWGCNA implementation. The function that allows for the comparison of multiple objects is named `create_combined_module_file` and it was designed to generate a comprehensive dictionary containing gene modules by processing data from multiple module text files. The module text files can be extracted from the module extraction algorithm in effect and they should contain all genes of the network space-separed by the respective module/cluster they were assigned. This implies that any number of files could be analyzed. In this project, only one other approach was fruitful in the extraction of gene modules, however, should the need arise to compare more than one approach, the PyWGCNA package would present itself as a limitation. As previously mentioned, the function also provides an option to filter-out non-overlapping genes to ensure that only those present in all files are included. This option can be used by setting the `remove_non_overlapping_genes` parameter as 'True'.

The function begins by initializing data structures to store gene information and module data. It then iterates through the input module files, extracting gene names and their assigned module. The assumption being that each line in the files contains a gene name followed by its corresponding module identifier. If the `remove_non_overlapping_genes` parameter is set to 'True', the function identifies common genes across all provided files, ensuring that only genes shared by all input files are considered. Next off, the function creates DataFrame structures for the common genes, associating them with their respective module colors. These DataFrames are organized within the dictionary, with keys representing network or module names. At last, the function returns the compiled dictionary (`self.geneModule`), which provides a consolidated view of gene modules across multiple input files and, as such, serves as a valuable resource for further analysis and comparisons of gene modules derived from various sources.

3.4 BADDADAN

Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN), as the name suggests, is a method designed to analyse the dynamics of expression profiles within *Arabidopsis thaliana*'s networks. This novel approach is currently being developed by Ben Noordijk, at the Bioinformatics Group of Wageningen University & Research. It is worth noting that this method is still in the developmental phase and has not been officially released. Consequently, the following section will provide a brief overview of the implementation steps, recognizing that they may not entirely align with

the final release.

Compared to previous approaches, BADDADAN offers a more comprehensive solution for analyzing gene expression dynamics. It not only extracts gene modules from expression data but also allows for the inference of regulatory interactions between predicted modules. This additional functionality enhances our understanding of the underlying regulatory networks by providing insights into how genes within these modules influence each other. Additionally, BADDADAN formulates ODEs that aim to simulate the dynamic patterns of gene module expression levels over time, enabling users to predict how these modules will respond under varying environmental conditions. This capability facilitates the generation of more precise hypotheses and experimental design in plant breeding techniques. A scheme of the general workflow for BADDADAN can be observed in Figure 10.

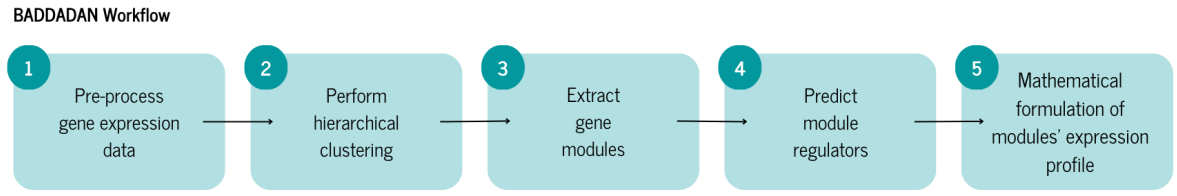


Figure 10: Schematic representation of the general workflow of the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach. The algorithm starts by pre-processing the gene expression data. It then performs hierarchical clustering to predict gene modules within the input data. Next, potential regulators are predicted for the extracted modules, allowing the user to infer a Gene Regulatory Network (GRN). Lastly, the dynamics of the expression profile of each module are mathematically formulated as Ordinary Differential Equation (ODE)s.

3.4.1 Data Pre-processing

The BADDADAN algorithm processes input data in the form of a CSV file, converting it into a pandas Dataframe that represents a time series expression matrix. By default, the algorithm performs Log2 transformation as part of its data pre-processing procedure. The Log2 transformation is particularly beneficial for Affymetrix data, where intensities represent transcript numbers. It widens the range of genes with low-intensity values and compresses the scale of highly intense ones, improving the algorithm's adaptability to diverse data patterns Archer et al. [2004].

After data normalization, the algorithm removes non-differentially expressed genes from the input dataset by determining the variation between samples based on the user-specified method. For this project, the chosen method was the Mean Absolute Deviation (MAD) score. As stated by previous reports, the

MAD method has been shown to significantly enhance cluster separation by effectively representing gene variability, a crucial aspect for clustering analysis. Rafique and Mir [2020]. The algorithm proceeds by selecting the user-specified number of genes ('n_max') with the highest variation scores, representing differentially expressed genes. For this project, it was chosen to perform the analysis with only the top 2000 most deviating genes. It then subsets the dataset to retain only these highly variable genes, providing an option to save the resulting filtered dataset.

3.4.2 Extraction of Gene Modules

The BADDADAN algorithm employs hierarchical clustering to identify genes that share similar expression patterns and, subsequently, extract gene modules from the filtered time series expression matrix. This key step is initiated by computing Pearson Correlation Coefficient among pairs of genes within the input expression matrix, effectively measuring the degree of similarity in their expression profiles. Next, it calculates the distances between genes based on these correlation values. Hierarchical clustering is then applied to this distance matrix, generating a hierarchical tree that visually represents the relationships among genes.

It's worth emphasizing that BADDADAN, compared to the other approaches employed for this project, is the only one that allows users to specify their desired number of modules, enhancing user customization. For this particular project, a choice of four modules was made to simplify the formulation and simulation of ODEs. The algorithm then proceeds to cut the hierarchical tree into discrete clusters, employing the maximum linkage method. Each gene receives a cluster label, referred to as 'cluster_id', depending on its inclusion in one of the identified clusters. This feature enables users to visualize the clustering outcomes using a clustermap, that facilitates a visual interpretation of the relationships among the different clusters.

3.4.3 Inference of GRNs

In the subsequent algorithmic step, the BADDADAN framework proceeds to generate a file that maps each probe to its corresponding cluster ID. This file serves as a TF2Network input file, a crucial part of constructing GRNs. As previously stated in Section 2.2, GRNs govern various biological processes through complex regulatory interactions between TFs and their respective target genes. Despite the substantial efforts to experimentally map *Arabidopsis thaliana* GRNs, the sheer number of TFs and limited throughput has left regulatory gaps for many genes. TF2Network has emerged as a valuable tool that takes advantage of extensive TF binding site information to elucidate GRNs by identifying potential regulators for co-expressed genes. With a prediction accuracy of 72-92% rate, this tool effectively addresses this challenge. The TF2Network tool generates a Tab-separated value (TAB) file, containing a comprehensive list of

regulators and corresponding genes. This file provides valuable information on the co-expression between input genes and predicted regulator, as well as enrichment statistics for each motif Kulkarni et al. [2018].

The workflow for the inference for GRN starts with a check to ensure the existence of the TF2Network output file. The following step includes transforming the TAB file into a Module Regulatory Network object. This object is designed to integrate information about TFs and modules, crucial for further downstream analysis. The function starts by reading the TF2Network output file and then adding a module prefix to the target gene column and a TF prefix to the predicted regulators column. This step aids in the accurate identification of modules and TFs. The function specifies the regulatory relationship as 'BINDS_TO', signifying when a TF binds to a specific gene module. Notably, users can define a quality cutoff with the 'nr_top_hits' parameter, enabling the retention of a specific number of high-scoring Position Weight Matrice (PWM). PWMs gauge the probability of a specific TF binding to a particular DNA sequence, with higher scores indicating a stronger binding likelihood.

Additionally, the framework also includes functions to: (1) include edges that indicate what TF is transcribed by which module by asserting its edge as 'TRANSCRIBED_BY'; (2) remove TFs and modules that are not connected to clean up and reduce noise within the network; (3) check if a TF is regulated by a module by verifying if the module's mean expression has a positive correlation with the TF; and (4) assert if each TF up or downregulates its target module. The network plotting method computes the network by resorting to `Matplotlib` library, identifying TFs as blue nodes and modules as orange nodes. To reflect the different relationships, the edges that connect the different nodes are also color-coded. The following relationships can be identified: 'REGULATES' as 'green', 'BINDS_TO as 'brown', 'TRANSCRIBED_BY' as 'black', 'UPREGULATES' as 'green', 'DOWNREGULATES' as'red', and 'UP_OR_DOWN' as 'grey'. Lastly, it is also possible to plot a module-module network. In this representation, the intermediaries of gene regulation, TFs, are removed, allowing for an eased identification of the regulatory relationship between modules. An example of these networks can be observed in Figure 14 of Subsection 4.1.3 of Chapter 4.

3.4.4 Formulation and Simulation of ODEs

Before moving to the construction of the ODE model based on the regulatory networks, the BADDADAN approach first asserts if the expression data has been clustered. The next step involves the calculation of the modules' expression profiles over time. These profiles are structured as tuples, which consist of time points and corresponding module expressions. This process simplifies the ODE fitting process, by grouping similar expression profiles of clustered genes as a single equation that represents the module's behaviour over time. All relevant information to the ODE equations is stored under the same class that

implements the various needed functions for ODE formulation.

One significant consideration in the modelling process is the handling of temperature variations and their influence on gene expression. The implementation of a 'heat_end_time' parameter was recently introduced by Ben to improve the fitting process according to the project's aim. This parameter can be specified by the user and it is employed to control the transition between warm and cold conditions, mirroring the temperature cycles observed in experimental conditions. The modelling process further entails extracting relevant information from the module-module network. This step encompasses specifying regulators, and target modules, and delineating the nature of their interactions, such as activation or inhibition.

The ODE models employed in this process primarily rely on Hill equations, sigmoid functions that describe transcriptional switch mechanisms with cooperative responses. These equations effectively illustrate how regulators impact the gene expression of the target module, capturing both activation and inhibition effects Radivoyevitch [2009], Hoops et al. [2016]. Furthermore, these ODE models are categorized into two main types: linear and non-linear formulas. In the context of the BADDADAN implementation, the key distinction lies in their formulation, where non-linear formulas account for directionality, accommodating both activation and inhibition effects. In contrast, linear formulas primarily consider the regulator's influence on the target module

Following the formulation of the ODE model, the BADDADAN framework moves forward to the ODE fitting. During this step, model parameters are estimated using real data, specifically, the average module expression levels across all time points. The approach allows the user to specify values for important aspects, such as the number of iterations, the maximal number of fittings and the interval of possible parameter values. The last one holds a higher significance, as it imposes constraints on the possible parameter values of the equation to ensure a realistic biological representation. For this project, a number of five fits, five hundred iterations and a parameter limit of 100 was chosen. A loss function is defined to quantify the disparity between predicted and observed gene expression profiles at specific time points, guiding the fitting process. The algorithm also allows the user to select the most fitting optimization methods. While Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) is selected as the default option, it also allows for the use of Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS), Differential Evolution, Basin-Hopping, and Simplicial Homology Global Optimization for different optimization purposes. These methods are employed to determine the best-fitting parameters that minimize the loss value.

Once all aspects of the fitting process are established, the parameters are initialized with randomly

assigned initial values within predefined ranges. These initial values are strategically chosen to closely align with the measured data, aiming to facilitate convergence during optimization and promote the discovery of meaningful solutions. BADDADAN also encompasses model evaluation, which involves identifying the best and worst-performing modules, plotting the Hill equation, and evaluating the quality of the model fit. To enhance the robustness of parameter estimation, multiple fits are conducted simultaneously using various starting conditions. Throughout this process, a diverse array of models is explored, allowing for an extensive exploration of various parameter combinations. To select the most suitable model, the Bayesian Information Criterion (BIC) is employed. The BIC enables a quantitative evaluation of various models' performance. The goal is to identify the model that not only demonstrates a strong fit to the data but also strikes a balance between goodness of fit and model complexity. The best-fit model is determined based on the lowest BIC value, indicating an optimal compromise between model quality and simplicity Anderson and Burnham [2004].

Chapter 4

Results and Discussion

Chapter 4 of this report offers an in-depth analysis of the results achieved throughout this research, delving into their interpretation, significance and implications for the project's aim. This chapter is structured into distinct sections, each dedicated to a critical aspect of the analysis. The first Section, 'Extracted Gene Modules' explores the modules that were extracted by each tested methodology, BIONIC, PyWGCNA, and BADDADAN, and their association with key traits, such as flowering time and temperature to ascertain valuable insights. Within this Section, a Subsection on 'ODE System and Simulation' explores the utilization of some of these modules in mechanistic modelling, simulating the modules' expression dynamics over time. The subsequent Section delves into the 'Analysis of Agreement in Extracted Modules', where similarities shared among a set of modules extracted through different methodologies are assessed. The Chapter then progresses to a 'Gene Set Enrichment Analysis' where we assess if the extracted modules have an association with relevant traits.

4.1 Extracted Gene Modules

This subsection will present the modules extracted using the selected module extraction algorithms and explore the relevant characteristics.

4.1.1 BIONIC

Let us recall that, as previously stated in Section 3.2, BIONIC is a DL-based framework designed to integrate diverse networks. It achieves this by learning gene features that more accurately represent the original network topology, resulting in enhanced performance in downstream tasks, such as module detection. This approach entails a two-step procedure for our desired task, first, it involves training the DL algorithm to obtain the integrated feature matrix, and, secondly, the employment of the BIONIC-evals package to compare the performance of the module detection task of the original input networks and the integrated

features by accessing the cluster agreement to the evaluation standard, according to either AMI or ARI scores. The difference between the use of these two scores can be found in Section 3.2.3.

One of the aspects that could greatly impact this task is the chosen evaluation standards. In fact, for the specific task at hand, predicting gene modules associated with the ambient temperature response in the flowering time transition, there are no reported gene ontology annotations for gene complexes. This inherently increases the difficulty and interpretability of obtained results as there are no reports of scientifically verified complexes relevant to this task.

To allow for this analysis, two evaluation standards were created. The first one represented flowering time clusters. Within this standard, cluster IDs correspond to specific genes associated with the flowering time response as described in Flowering Interactive Database (FLOR-ID) Bouché et al. [2016]. The constituent genes within each cluster were the genes that form the interaction network of the gene representing that cluster-ID. The second standard involved creating clusters of annotated genes related to keywords relevant to the project's aim. In this context, the AmiGO2 website's gene ontology search engine Carbon et al. [2009] was utilized to identify genes and gene products associated with relevant gene ontology terms, including 'temperature', 'cold', 'heat', 'flowering', and 'flowering time'. These associations are established through either manual curation, using published experimental or phylogenetic results, or automated annotation tools developed by experienced curators Consortium [2015]. Here, each cluster ID was indicative of a specific term, and the constituent genes within that cluster were genes linked to the respective term.

The module detection task was initially tested with the E-MEXP-728 Balasubramanian et al. [2006] dataset. The ARI scores obtained from evaluating the cluster agreement can be observed in Figure 11. Whereas the original paper reported AMI scores from 0.3 to 0.8, depending on the used benchmark, the ARI score attained from both evaluation standards never reached values above 0.1.

ARI is a symmetric measure, ranging from -1 to 1. A score of -1 indicates a discording community, while a score of 1 indicates a community in complete accordance Patgiri et al. [2023]. As such, a score of 0.1 indicates a substantial discrepancy with the evaluation standard, suggesting that the clustered genes were likely assigned randomly, rather than reflecting meaningful patterns or associations. Given that the quality of the integrated features relies heavily on the model's training performance, the parameters for this procedure were adjusted. This step aimed to troubleshoot and assess whether altering these parameters could lead to an improvement in the integrated features' performance for the module detection task. Initially, the process followed the default parameters outlined in the example configuration files from the original repository. However, the retrieval of recurring low scores was a clear indicator of a persistent issue

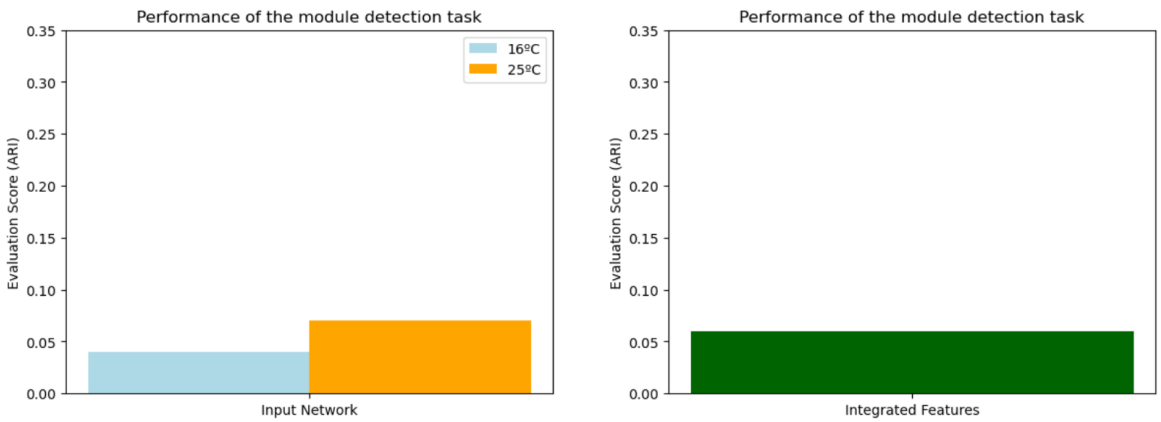


Figure 11: Evaluation of BIONIC's module detection performance with the E-MEXP-728 dataset Balasubramanian et al. [2006], utilizing the Adjusted Rand Index (ARI) score to measure the cluster agreement. The figure comprises two bar plots. The left plot illustrates the similarity between modules predicted from the original input networks for 16°C and 25°C, compared to the standard. The right plot demonstrates the similarity between modules predicted from the integrated feature matrix and the standard. It is possible to observe that the scores for both tasks do not surpass the 0.1 value.

in the clustering process. To delve deeper into the problem, the predicted modules were extracted and analyzed. This process revealed a consistent pattern where the vast majority of genes were consistently assigned to the same cluster.

All of the issues faced throughout the use of this approach and detailed in this Section were reported to the developer of the tool via e-mail. This contact promoted an active discussion over produced results. However, the provided feedback and suggested alterations did not improve any of the attained results. As such, a final decision was made to abandon this method.

4.1.2 PyWGCNA

To revisit, the implementation of PyWGCNA intends to replicate the original R WGCNA library in a Python environment to offer a more efficient and scalable implementation. It did so while introducing additional functions for enhanced downstream analysis of extracted modules, including comparing module objects and conducting functional enrichment analysis. It is important to emphasize that the PyWGCNA approach was designed for the analysis of RNA-sequencing data, whereas the data used for this project was generated through GeneChip technology. However, the data has been effectively normalized, and thus, any potential variances caused by variations in the data source are anticipated to have a minimal impact on the interpretation of the insights gained.

For the initialization of both PyWGCNA objects, a Transcripts per Million (TPM) cutoff value was determined. Initially, the default value of one was employed, however, this led to the generation of a large number of modules for both datasets. Subsequently, we iteratively adjusted these values to obtain a more manageable number of modules for each dataset. In the case of E-MEXP-728 Balasubramanian et al. [2006], a TPM value of six was selected, while a cutoff of seven was applied to the E-MEXP-1304 dataset Michael et al. [2008]. The final number of extracted modules was 19 for the E-MEXP-728 and 12 for the E-MEXP-1304 dataset. In Figure B1 and B2, present in Appendice B, it is possible to visualize a capture of the PyWGCNA E-MEXP-728 and E-MEXP-1304 objects in a network format, respectively.

As stated in Section 3.3.4, one of the most relevant downstream analysis procedures after the module detection step in the PyWGCNA approach is the analysis of the relationship between modules and data traits. Figure 12 and 13 display the resulting heatmap for the E-MEXP-728 and the E-MEXP-1304 datasets, respectively. The E-MEXP-728 possesses seven data traits: four corresponding to the organism's age (35, 37, 40 and 44 days), two linked to environmental conditions (temperature and photoperiod) and one representing flowering time (leaf number). In contrast, the E-MEXP-1304 contains only one data trait, primarily due to the consistent environmental conditions maintained throughout the experiment. This consistency, except for cyclical temperature alterations, aligns with the project's focus on extracting modules associated with the temperature response, minimizing unwanted biases introduced by additional changes in light conditions. However, it is important to note this uniformity, with the absence of data traits for different conditions, limits the presence of distinct traits. Consequently, this hinders the algorithm's ability to attribute the different detected expression profiles to their corresponding traits. On the contrary, when variations in conditions are introduced within a single run, as in the case of the E-MEXP-728 dataset, it diminishes uniformity. This variation carries the potential to impose limitations on the algorithm's ability to effectively detect specific expression profiles and accurately correlate them with distinct traits. As such, an ideal dataset would exhibit variations in the same characteristics within a single run.

In Figure 12, it is possible to make the conjecture that the expression profiles of modules associated with the traits 'Temperature', 'Photoperiod', and 'LeafNumber' exhibit a high correlation. This is evidenced by their nearly identical Pearson correlation values. However, this observation is also indicative of the previously discussed limitation that arises from introducing varying conditions within a single run. Such variability can make it challenging to detect differences in expression profiles. With a significance level of $p < 0.05$, it is evident that most of the correlations discovered are statistically significant. Specifically, the 'Gold', 'Sienna', 'Chocolate', 'Oldlace', 'Darkgoldenrod', 'Firebrick', 'Floralwhite', 'Brown', 'Black', 'Coral', and 'Darkred' modules exhibit statistically significant correlations. This indicates that the 'Gold', 'Sienna',

'Chocolate', 'Oldlace', and 'Darkred' modules have negative correlations with temperature, while the 'Darkgoldenrod', 'Firebrick', 'Floralwhite', 'Brown', 'Black', and 'Coral' modules display positive correlations with temperature.

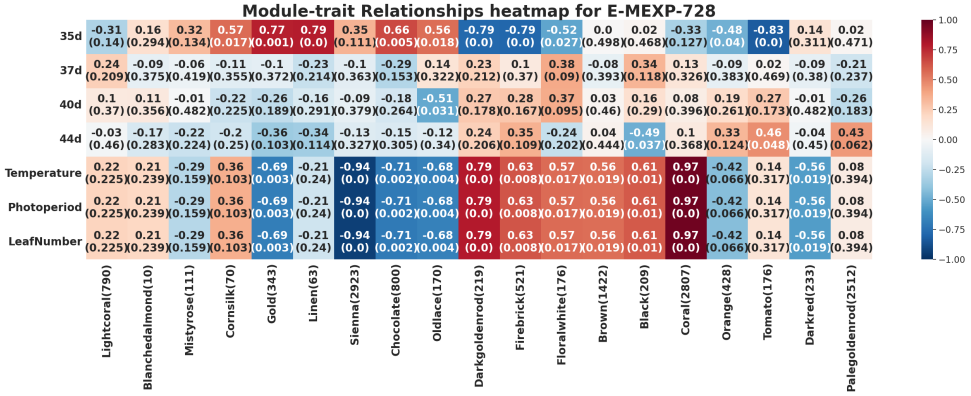


Figure 12: Correlation heatmap showing the relationships between extracted modules and data traits within the E-MEXP-728 dataset Balasubramanian et al. [2006]. The heatmap displays Pearson correlation values, indicating the strength of the association between a module and a specific trait. Positive correlations are depicted in red, while negative correlations are shown in blue. Additionally, the plot represents p-values below the Pearson correlation values, providing insights into statistical significance. It is also possible to detect seven data traits, these are related to age (35, 37, 40 and 44 days), environmental conditions (temperature and photoperiod) and flowering time (leaf number).

In contrast, the E-MEXP-1304 heatmap, depicted in Figure 13, highlights the impact of the limitation mentioned earlier, the reduced number of data traits. For this dataset, with a significance level of $p < 0.05$, only three of the identified modules can be considered statistically significant: 'Sienna', 'Rosybrown', and 'Lightsalmon'. Remarkably, each of these modules exhibits a negative correlation with temperature.

The subsequent step in the PyWGCNA pipeline involved conducting a comparative analysis to assess the agreement of genes associated with the extracted modules from both the E-MEXP-728 and E-MEXP-1304 datasets. As previously explored in Section 3.1, this comparison is of particular significance as it enables us to assess PyWGCNA's ability to capture coherent expression patterns associated with a specific condition and consistently assign genes to corresponding clusters. The visual representation of this comparison is depicted in a network format in Figure A1. Due to its size, the figure has been relocated to Appendix A. Notably, when adopting a Jaccard similarity index of 0.2 as a threshold for meaningful module similarity, the analysis revealed eight pairs of modules demonstrating agreement:

- E-MEXP-1304 'lightcoral' and E-MEXP-728 'sienna';
- E-MEXP-728 'firebrick' and E-MEXP-1304 'rosybrown';

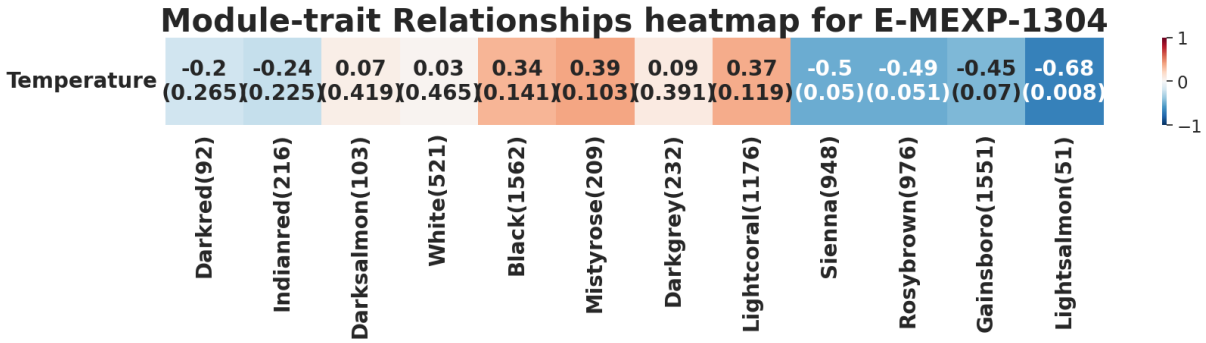


Figure 13: Correlation heatmap showing the relationships between extracted modules and data traits within the E-MEXP-728 dataset Michael et al. [2008]. The heatmap displays Pearson correlation values, indicating the strength of the association between a module and a specific trait. Positive correlations are depicted in red, while negative correlations are shown in blue. Additionally, the plot represents p-values below the Pearson correlation values, providing insights into statistical significance. In this heatmap, it is possible to detect only one trait, associated with the temperature.

- E-MEXP-1304 'rosybrown' and E-MEXP-728 'brown';
- E-MEXP-1304 'black' and E-MEXP-728 'lightcoral';
- E-MEXP-728 'lightcoral' and E-MEXP-1304 'white';
- E-MEXP-1304 'gainsboro' and E-MEXP-728 'coral';
- E-MEXP-1304 'sienna' and E-MEXP-728 'chocolate';
- E-MEXP-1304 'sienna' and E-MEXP-728 'darkgoldenrod'.

When further looked into it is possible to infer that some of the meaningful relationships found between modules are between modules with disagreeing correlations with temperature. For example, E-MEXP-1304 'lightcoral' and E-MEXP-728 'sienna', have a high agreement, however, denote a contrasting module-temperature profile, with the E-MEXP-1304 'lightcoral' having a positive correlation with temperature, and the E-MEXP-728 'sienna', a negative correlation with temperature. Some of these cases, where a contrasting module-temperature profile is found, like the E-MEXP-1304 'lightcoral' and E-MEXP-728 'sienna', E-MEXP-1304 'gainsboro' and E-MEXP-728 'coral', E-MEXP-1304 'black' and E-MEXP-728 'lightcoral', and E-MEXP-728 'lightcoral' and E-MEXP-1304 'white', the modules' correlation with temperature is not statistically significant, meaning that although there is a disagreement between temperature profiles, it would still be necessary to analyse more expression data to infer the clustered genes are in fact associated with

a specific temperature profile. For other cases, like the E-MEXP-728 'firebrick' and E-MEXP-1304 'rosybrown', E-MEXP-1304 'rosybrown' and E-MEXP-728 'brown', and E-MEXP-1304 'sienna' and E-MEXP-728 'darkgoldenrod', the module-temperature profiles are contradictory but statistically significant, meaning that the clustered genes in both clusters could be involved in both the hot and cold temperature response.

The extracted modules were also searched for the presence of TFs associated with the cold and heat response, previously mentioned in Section 2.1.2: DREB 1/CBF (AT4G25470), MBF1c (AT3G24500), and DREB2A (AT5G05410). These were only found in the modules extracted from the E-MEXP-728 dataset, with two hits for the DREB 1/CBF TF, one in the 'coral' module and another in the 'lightcoral' module; one hit for the MBF1c TF in the 'firebrick' module; and no hits for the DREB2A TF.

4.1.3 BADDADAN

As a reminder, the BADDADAN algorithm was designed to analyse expression profile dynamics within *Arabidopsis thaliana*'s networks. Besides its capacity to extract gene modules, this algorithm enables the inference of regulatory interactions among these modules by computing a GRN. Additionally, it offers functionalities for formulating these interactions as ODE and simulating the expression dynamics of modules over time. It is worthy to note that this approach was employed only to the E-MEXP-1304 Michael et al. [2008] dataset, due to its consistent experimental conditions regarding temperature variations.

Among the top 2000 most deviating genes, this approach extracted a total of four modules. To infer the regulatory relationships between these modules, potential TFs were predicted. Figure 14 provides a visual representation of the GRN inferred using the TF2Network tool.

For the sake of simplifying and easing the interpretation of regulatory relationships, the algorithm permits the removal of intermediary components, such as the TFs. A network that exclusively portrays the regulation of modules is presented in Figure 15.

From this image, it is possible to affirm that all established relationships are of downregulation nature. As a matter of fact, the following relationships could be stipulated:

- 'MODULE0' and 'MODULE1' exhibit a bidirectional downregulation relationship;
- 'MODULE0' downregulates both 'MODULE2' and 'MODULE3' and is downregulated by 'MODULE1';
- 'MODULE1' downregulates both 'MODULE0' and 'MODULE2' and is downregulated by 'MODULE0';
- 'MODULE2' is downregualted by all other extracted modules;
- 'MOUDLE3' downregulated 'MODULE2' and is downregualted by 'MODULE0'.

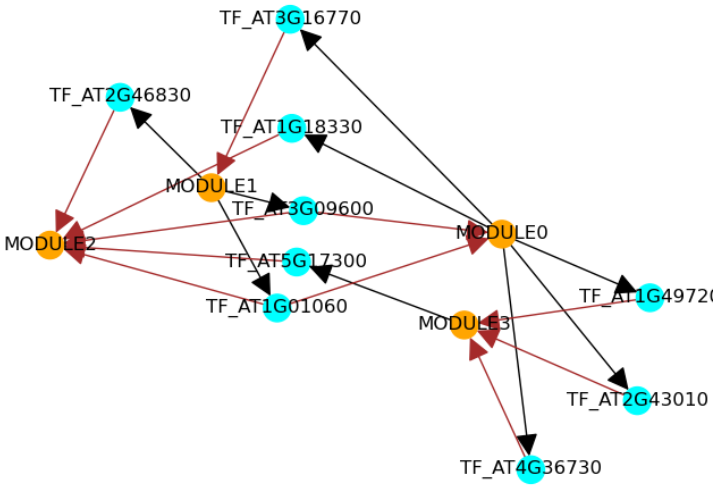


Figure 14: Visual representation of modules extracted through the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach and their corresponding transcription factors predicted by the TF2Network tool, presented in a network format. In this visualization, modules are depicted as orange nodes, while transcription factors are represented as blue nodes. The connecting edges are color-coded, with red denoting downregulation relationships and black indicating transcriptional associations. The presence of arrows in the network highlights the directionality of regulatory interactions between transcription factors and modules.

ODE System and Simulation

The following equations constitute the system of ODEs that describe the expression dynamics of the extracted modules and the relationships previously described:

$$\frac{dy_0}{dt} = -\delta_0 \cdot y[0] + \gamma_0 \cdot \text{temp} \cdot y[0] + \frac{\beta_1 \cdot k_1}{k_1 + y[1]} \tag{4.1}$$

$$\frac{dy_1}{dt} = -\delta_1 \cdot y[1] + \gamma_1 \cdot \text{temp} \cdot y[1] + \frac{\beta_0 \cdot k_0}{k_0 + y[0]} \tag{4.2}$$

$$\frac{dy_2}{dt} = -\delta_2 \cdot y[2] + \gamma_2 \cdot \text{temp} \cdot y[2] + \frac{\beta_1 \cdot k_1}{k_1 + y[1]} + \frac{\beta_0 \cdot k_0}{k_0 + y[0]} + \frac{\beta_3 \cdot k_3}{k_3 + y[3_1]} \tag{4.3}$$

$$\frac{dy_3}{dt} = -\delta_3 \cdot y[3] + \gamma_3 \cdot \text{temp} \cdot y[3] + \frac{\beta_0 \cdot k_0}{k_0 + y[0]} \tag{4.4}$$

This mathematical formulation not only aids in understanding how expression dynamics evolve over time but also leverages ODEs for their ability to offer a continuous depiction of molecular dynamics and to

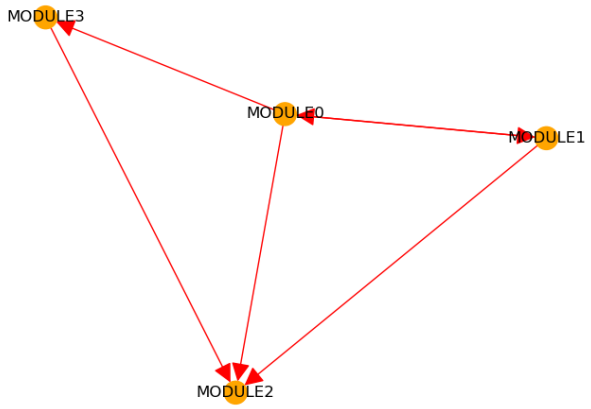


Figure 15: Visual representation of modules extracted using the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach, presented in a network format. Each module is depicted as an orange node. The edges connecting them are colored red, indicating downregulation relationships. This representation also elucidates the directionality of regulatory interactions between modules, highlighted by the presence of arrows.

quantify the influence of gene expression disturbances on a system's output. To enhance interpretability, Table 2 provides insights into the significance of each parameter.

Table 2: Definition of key parameters employed in the Ordinary Differential Equation's system.

Parameter	Meaning
δ	Degradation rate
y	Input variable
γ	System's responsiveness to temperature variations
temp	Specific temperature at a given point
β	Maximum transcription rate
k	Abundance of a product at a half-maximum transcription rate

Therefore, each equation can be divided into three factors, each representing distinct aspects that govern significant changes in the system's dynamics: the decay factor, temperature factor, and regulatory factor. Taking Formula 4.1 of 'MODULE0' as an example:

- **Decay factor**, describes the rate of degradation: $-\delta_0 \cdot y[0]$;
- **Temperature factor**, describes the temperature's effect on module regulation: $\gamma_0 \cdot \text{temp} \cdot y[0]$;
- **Regulatory factor**, describes the rate of transcription: $\frac{\beta_1 \cdot k_1}{k_1 + y[1]}$.

The algorithm's fitting process successfully simulated the expression dynamics representing the behavior of the formulated system, as shown in the right plot of Figure 16. A comparative analysis with the left plot of the same figure, which illustrates the system's real-life behavior based on experimental data, reveals that there is still room for improvement in the mathematical model. Notably, both the experimental and simulated data capture the cyclical dynamics, with fluctuations in the expression levels occurring at corresponding time points. However, it is evident that these dynamics exhibit distinct patterns.

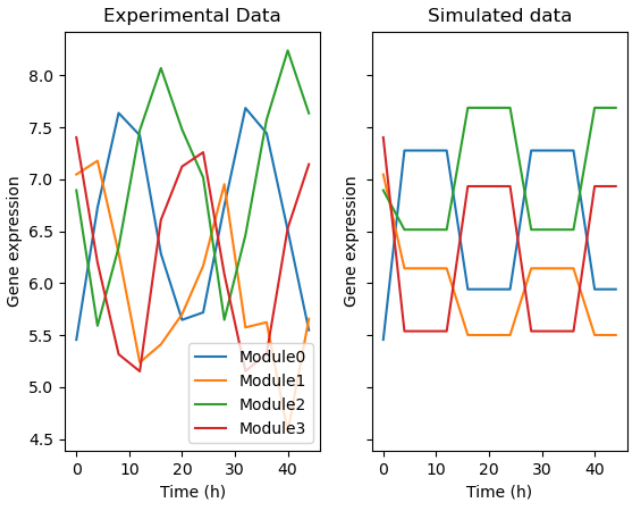


Figure 16: Plots illustrating the temporal expression dynamics of modules obtained through the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach based on experimental (left) and simulation of the Ordinary Differential Equation (ODE) system (right). Each module is represented with a different color as indicated in the legend.

Figure 17 depicts the expression levels of all 63 genes assigned to 'MODULE0' across all measured time points using this approach. Similar plots for 'MODULE1', 'MODULE2', and 'MODULE3' can be found in Figure B3, Figure B4, and Figure B5, respectively, in Appendice B. This image highlights the advantages of adopting a modular approach to group genes with similar expression patterns. In fact, in a non-modular approach, the higher dimensionality would exponentially escalate the complexity of the ODE system, making the parameter fitting process significantly more challenging. Moreover, the modular approach enhances biological interpretability by clustering genes that likely participate in the same biological processes or pathways, simplifying the inference of functional relationships within modules. This, in turn, provides improved insights into the coordinated regulation of genes and the underlying mechanisms that govern their behavior, ultimately contributing to a more comprehensible analysis Ackermann and Strimmer [2009].

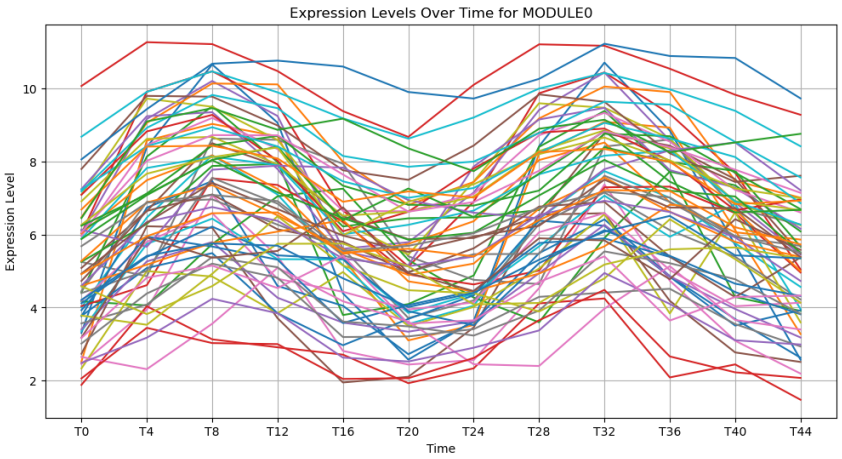


Figure 17: Plot representing the changes in expression levels over time for all 63 genes attributed to Module 0 with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach and E-MEXP-1304 dataset Michael et al. [2008]. Each line and color represent a unique genes.

4.2 Analysis of Agreement in Extracted Modules

Consistent with the analysis conducted in Subsection 4.1.2 of this chapter, this subsection delves into the comparison of modules extracted using both approaches, PyWGCNA and BADDADAN. The aim is to evaluate aspects such as, the extent to which both algorithms assign the same genes to identical clusters, their ability to detect similar expression patterns, and the degree of overlap between the extracted modules. A visual representation of the comparison between overlapping genes is provided in Figure A2. Like the previous comparison figure, this one has also been relocated to Appendix A.

Similarly, when applying a Jaccard similarity index threshold of 0.2 for meaningful module similarity, this analysis unveiled four pairs of modules that exhibited agreement between the PyWGCNA and BADDADAN approaches:

- E-MEXP-1304 'lightcoral' and E-MEXP-1304 'MODULE0';
- E-MEXP-1304 'MODULE0' and E-MEXP-728 'lightcoral';
- E-MEXP-1304 'black' and E-MEXP-1304 'MODULE3';
- E-MEXP-1304 'white' and E-MEXP-1304 'MODULE1';

It's worth noting that the modules extracted using the BADDADAN approach exhibited a higher frequency of agreement with the modules obtained from the PyWGCNA approach in the case of the E-MEXP-

1304 dataset. This observation is primarily due to the fact that the BADDADAN approach was applied to the E-MEXP-1304 dataset.

Given that the BADDADAN approach does not provide information for inferring module-trait relationships, it's challenging to determine whether the agreement between modules corresponds to similar module-temperature profiles. However, we can still observe some interesting patterns. For instance, 'MODULE0' displayed substantial agreement with two modules related to a positive temperature correlation and a third module, E-MEXP-1304 'black', which came close to the Jaccard cutoff with a similarity of 0.19 and also had a positive temperature correlation. On the other hand, 'MODULE1' only exhibited a high agreement with the E-MEXP-1304 'white' module that had no temperature correlation. 'MODULE2' did not demonstrate any noteworthy agreements with other modules but it is possible to observe a agreement of 0.18 with the E-MEXP-1304 'gainsboro' module that has a negative correlation with temperature. Finally, 'MODULE3' had one relevant agreement with the E-MEXP-1304 'black' module, which had a positive temperature correlation. Nonetheless, none of these modules exhibited a statistically significant temperature correlation at the 0.05 level.

4.3 Gene Set Enrichment Analysis

As previously discussed in Section 2.4.2, GSEA provides an effective means of identifying gene sets linked to phenotypic changes. This subsection will explore the outcomes of functional enrichment for the modules extracted from the applied approaches. If a module is not mentioned, it can be inferred that it either did not yield statistically significant results for any biological process or did not exhibit relevant functional enrichment within the project's scope. Additionally, modules showing a major overrepresentation of biological processes, albeit not directly associated with the project, will still be discussed.

4.3.1 PyWGCNA

E-MEXP-728

The visual representation of the GSEA of modules extracted from the E-MEXP-728 dataset Balasubramanian et al. [2006] through the PyWGCNA approach can be observed in Figure 18.

Before delving deeper into the interpretation of the plot, it is possible to observe that most of the modules present in the analysis for agreement between extracted modules were found to be functional enriched, showcasing the algorithm's ability to detect biologically relevant gene modules. Three modules emerged as overrepresented in either the regulation of flowering or the transition to flowering: the

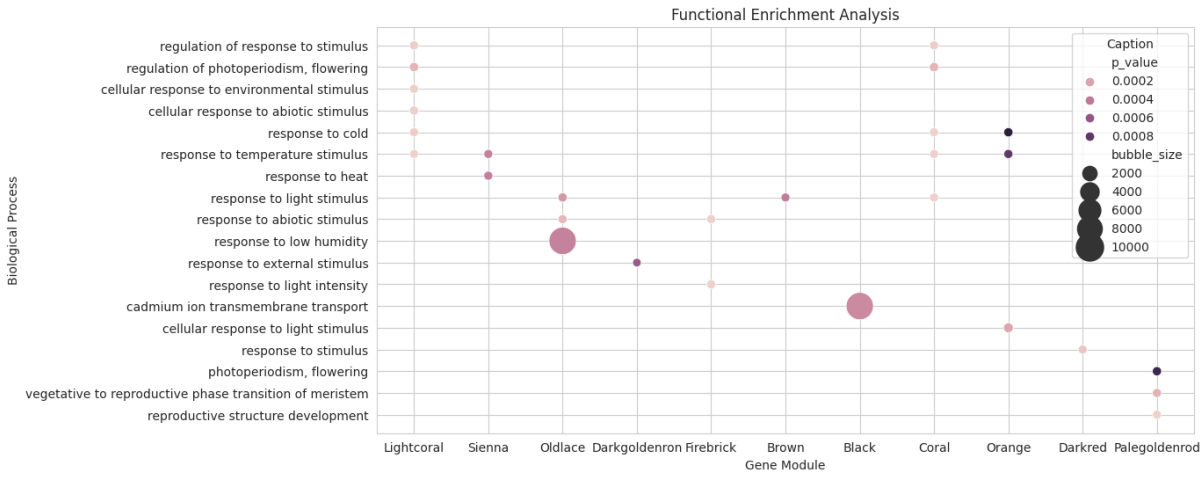


Figure 18: Bubble plot representing the Functional Enrichment Analysis for the modules extracted from the E-MEXP-728 dataset Balasubramanian et al. [2006], through the Python-based Weighted Gene Correlation Network Analysis (PyWGCNA) approach. The bubble size corresponds to the fold enrichment, while the color of the bubble signifies the p-value.

'lightcoral', 'coral', and 'palegoldenrod' modules. With the 'lightcoral' and 'coral' module found to be overrepresented both for 'regulation of photoperiodism, flowering' and 'response to cold'.

Notably, both of these modules displayed a positive correlation with temperature in the module-trait analysis, implying lower gene expression as temperature increases. This observation is consistent with the genes' involvement in the 'response to cold' since their response might involve a decrease in expression. This downregulation could potentially activate downstream genes, denoting a cascade effect in the cold response pathway. The 'sienna' module, in contrast, displayed a statistically significant negative correlation with temperature and was found to be overrepresented in terms of 'response to heat'. These observations suggest a complex regulatory network where genes respond to environmental temperature by downregulating their expression, potentially triggering downstream effects that drive the a response to both heat and cold. Furthermore, the 'coral' module was identified as overrepresented in 'response to light stimulus', highlighting the limitations posed by the presence of inconsistent environmental variables within the E-MEXP-728 dataset, as previously highlighted.

The 'oldlace' module, with no relevant overlap with any other module, was found to be highly overepresented (fold enrichment value above 100) with genes associated with a 'response to low humidity'. Although not directly related to this project, as the focus is on the organism's temperature-response, this could prove to be a key module in the expansion of the mechanistic modelling for other environmental conditions, such as humidity. Another module that was found to have a highly overepresentation (fold enrichment value above 100) of a biological process was the 'black' module, found to be associated with

the 'cadmium ion transmembrane transport'.

E-MEXP-1304

The visual representation of the GSEA of modules extracted from the E-MEXP-1304 dataset Michael et al. [2008] through the PyWGCNA approach can be observed in Figure 19.

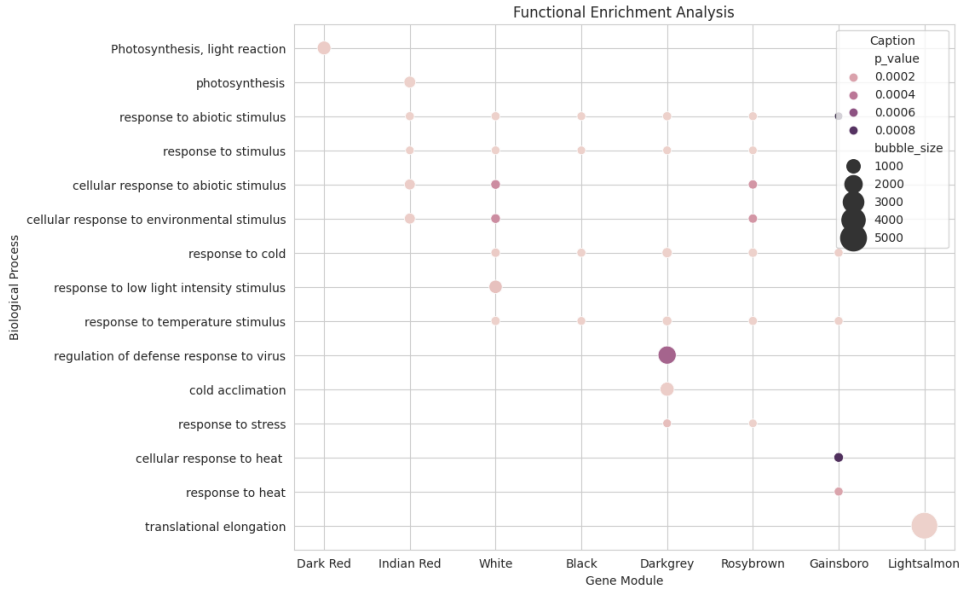


Figure 19: Bubble plot representing the Functional Enrichment Analysis for the modules extracted from the E-MEXP-1304 dataset Michael et al. [2008], through the Python-based Weighted Gene Correlation Network Analysis (PyWGCNA) approach. The bubble size corresponds to the fold enrichment, while the color of the bubble signifies the p-value.

Two modules emerged with an overall overrepresentation in biological processes associated with the response to abiotic and environmental stimuli and 'response to cold': the 'white' and 'rosybrown' modules. The 'white' module exhibited no correlation with temperature in the module-trait relationship. In contrast, the 'rosybrown' module revealed a statistically significant negative correlation with temperature, supporting the hypothesis that these modules are involved in temperature-related responses to environmental stimuli. Another module whose association with temperature could be further reinforced, is the 'darkgrey' module. This module displayed no correlation with temperature in the module-trait analysis, yet it was notably overrepresented in terms of 'response to cold' and 'cold acclimation' during the enrichment analysis. The 'black' module exhibited a positive correlation with temperature and an overrepresentation in 'response to cold', whereas the 'gainsboro' module showed a negative correlation with temperature and was overrepresented in terms related to 'response to heat' and 'cellular response to heat'. These observa-

tions indicate that genes within these modules may be involved in responses to heat and cold, with their responses potentially involving both increased and decreased expression levels, influencing downstream genes accordingly. None of the modules extracted from the E-MEXP-1304 dataset using the PyWGCNA displayed overrepresentation in terms of the flowering process.

Similarly to the E-MEXP-728 dataset, albeit with fold enrichment values hovering around the 50 range, two modules stood out for their substantial overrepresentation in non-relevant biological processes. Specifically, the 'darkgrey' module was linked to 'regulation of defense response to virus', while the 'lightsalmon' module was associated with 'translational elongation'.

4.3.2 BADDADAN

E-MEXP-1304

A visual representation of the GSEA of modules extracted from the E-MEXP-1304 dataset Michael et al. [2008] through the BADDADAN approach can be observed in Figure 20.



Figure 20: Bubble plot representing the Functional Enrichment Analysis for the modules extracted from the E-MEXP-1304 dataset Michael et al. [2008], through the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach. The bubble size corresponds to the fold enrichment, while the color of the bubble signifies the p-value.

A brief review of the presented enrichment analysis suggests that the modules extracted through the BADDADAN approach exhibited notably higher functional enrichment values. This discrepancy can be attributed to the size of the extracted modules. For instance, the largest module in BADDADAN comprises

63 genes, while some of the larger modules extracted by the PyWGCNA approach contain around 1500 genes. Take, for instance, the 'gainsboro' module extracted from the E-MEXP-1304 dataset, which offers a comparative example.

It's worth noting that the modules extracted using this approach displayed distinct characteristics, notably featuring more focused and coherent functional gene groups. For example, 'MODULE0' exhibited significant overrepresentation in biological processes related to environmental stimuli, particularly those linked to light. 'MODULE1' stood out for its high overrepresentation in processes such as the negative regulation of the circadian rhythm, flowering, photoperiodism, and the regulation of timing of the transition to the reproductive phase. 'MODULE2' was notably overrepresented for plant development and growth. Meanwhile, 'MODULE3', much like 'MODULE1', demonstrated overrepresentation in processes associated with the regulation of photoperiodism, flowering, reproductive processes, and flower development. Additionally, although not directly relevant to temperature response, 'MODULE3' also exhibited overrepresentation in 'response to water deprivation', making it another potential candidate for future expansion of mechanistic modeling to accommodate other environmental conditions.

Since the module-trait analysis is not available in the BADDADAN approach, direct comparisons with other analyses are not feasible. Nonetheless, it's evident that no module extracted using the BADDADAN approach displayed overrepresentation in 'response to cold' or 'response to heat'. In contrast, it is of significance to point out that both modules showing overrepresentation in flowering processes, 'MODULE1' and 'MODULE3', exhibited a high agreement with the 'lightcoral' module extracted from the E-MEXP-728 dataset using the PyWGCNA approach.

Figure B6, located in Appendice B, offers a network visualization illustrating module similarity between modules extracted from the E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] datasets using PyWGCNA and BADDADAN approaches. This visualization features symbolic representations denoting the results obtained from both the module-trait and functional enrichment analyses, specifically focusing on temperature and flowering-response.

Chapter 5

Conclusions and future work

5.1 Conclusions

The primary objective of this research project was to synergize machine learning and mechanistic modelling for enhancing plants' adaptability to climate change. The proposed approach, at its core, aimed to detect gene modules linked to the *Arabidopsis thaliana* response to environmental temperature variations. Subsequently, this knowledge would be integrated into an existing mechanistic model through a mathematical formulation of the regulatory interactions governing these modules. In theory, this approach held the potential to enhance the model's ability to predict flowering time in response to shifts in environmental conditions.

The employment of module detection tools, including BIONIC, PyWGCNA, and BADDADAN, allowed for the extraction of gene modules from datasets relevant to the research question. To infer the modules' relationships with the temperature response, intrinsic functions of these methods, such as PyWGCNA module-trait analysis, were employed. To validate the biological significance of these inferred relationships and the functional affinity of assigned genes, GSEA was conducted. These analyses yielded valuable insights, with several gene modules demonstrating statistically significant associations with temperature and notable overrepresentation in biological processes pertinent to the project's objectives. Indeed, numerous modules were found to be linked to heat and cold responses, regulation of flowering time, and the transition to the reproductive phase. While these findings enriched our understanding of the relationships between these gene clusters and the phenotypical changes induced by shifts in environmental conditions, they also prompted questions about the realism of certain associations. This concern stemmed from conflicting results obtained from the module trait and functional enrichment analyses.

One significant disappointment in this project was the inability to integrate biological networks and deduce gene modules through the BIONIC approach. In reality, the BIONIC method exhibited flaws in extracting relevant topological features and predicting gene modules accurately. Furthermore, the challenges

encountered during implementation were directly reported to the developer, although this discussion did not impact the results previously obtained. Given the method's claims in the original research paper, a substantial portion of the project's original timeline was dedicated to troubleshooting, making modifications, and implementing suggestions to align with those claims. Unfortunately, this poor time management ultimately hindered the achievement of one of the proposed goals for the project. The delay in progressing to the next project phases directly impacted the integration of the derived ODE into the pre-existing mechanistic model. This step was critical for enhancing the model's ability to predict flowering time in response to temperature changes.

As mentioned in the Introduction, this research encountered several barriers hindering the inference of plant-climate interactions. These challenges included the fragmentation of relevant data on expression level measurements across varying ambient temperature conditions, a limited understanding of the impact of ambient temperature on flowering regulation, and the translation of data into meaningful biological knowledge. By constructing an inventory of datasets relevant to this project's scope, we've taken a significant step towards consolidating context-specific data into an aggregated space, enhancing accessibility and combining research efforts.

Furthermore, this project successfully identified 14 modules with statistically significant relationships to temperature. Subsequent GSEA analysis revealed five gene modules associated with the regulation of flowering processes, two of which displayed dual associations with both the response to cold and the regulation of flowering. These discoveries translate into valuable biological insights extracted from the collected data.

By addressing these challenges and contributing to a clearer understanding of the relationship between environmental temperature and plant responses, this research helps bridge gaps in knowledge. These findings pave the way for potential enhancements in mechanistic modelling, particularly regarding predicting flowering time in response to variations in ambient temperature. As a final point, this research improves our understanding of plant-climate interactions and offers a foundation for further improvements in the field.

5.2 Prospects for Future Work

To enhance the insights obtained in this project, future analyses should involve comparing the extracted modules with those from a more cohesive dataset. This additional step could play a crucial role in validating some of the hypotheses presented in this project, particularly those with contrasting results. It would offer

a more precise assessment of the modules' responses to heat, cold, and the transition to flowering. As reiterated throughout this report, finding a singular suitable dataset proved challenging. Although using two datasets yielded valuable insights, an ideal dataset for such a study would maintain consistent environmental conditions, thus presenting an opportunity for future experimental designs.

Additionally, the predicted gene modules and described temperature associations can be validated or refuted through experimental methodologies. This work would involve performing gene expression profiling, which entails measuring the expression levels of genes chosen from the predicted modules in response to varying ambient temperatures while ensuring environmental factors remain constant. Additionally, transcriptomic profiling via RNA-seq experiments could be carried out on plants exposed to different temperature conditions while keeping all other environmental conditions constant. These approaches collectively offer a reliable means to further investigate the responsiveness of the gene modules.

Furthermore, considering the mathematical formulation of the modules extracted from the E-MEXP-728 dataset using the PyWGCNA approach as ODEs could provide a valuable strategy for making comprehensive comparisons with BADDADAN's modelling and fitting process. As reported, there is room for improvement in the fitting process, and this step could potentially yield visible enhancements.

In conclusion, addressing the research question posed in this project would require integrating the derived equations into the available mechanistic model. This integration would significantly enhance the model's predictive capabilities for flowering time in response to changes in environmental temperature. Moreover, it could pave the way for enhancing plants' adaptation to climate change, making it a promising endeavour for future research.

Bibliography

Mitsutomo Abe, Yasushi Kobayashi, Sumiko Yamamoto, Yasufumi Daimon, Ayako Yamaguchi, Yoko Ikeda, Harutaka Ichinoki, Michitaka Notaguchi, Koji Goto, and Takashi Araki. Fd, a bzip protein mediating signals from the floral pathway integrator ft at the shoot apex. *Science*, 309(5737):1052–1056, 2005.

Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC bioinformatics*, 10:1–20, 2009.

SR Adams, S Pearson, and P Hadley. The effects of temperature, photoperiod and light integral on the time to flowering of pansy cv. universal violet (viola× wittrockianagams.). *Annals of Botany*, 80(1):107–112, 1997.

Sina M Adl, Alastair GB Simpson, Mark A Farmer, Robert A Andersen, O Roger Anderson, John R Barta, Samuel S Bowser, GUY Brugerolle, Robert A Fensome, Suzanne Fredericq, et al. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *Journal of Eukaryotic Microbiology*, 52(5):399–451, 2005.

M Alejandra Mandel, Cindy Gustafson-Brown, Beth Savidge, and Martin F Yanofsky. Molecular characterization of the arabidopsis floral homeotic gene apetala1. *Nature*, 360(6401):273–277, 1992.

Carlos Alonso-Blanco and Maarten Koornneef. Naturally occurring variation in arabidopsis: an underexploited resource for plant genetics. *Trends in plant science*, 5(1):22–29, 2000.

Elena R Alvarez-Buylla, Mariana Benitéz, Enrique Balleza Dávila, Alvaro Chaos, Carlos Espinosa-Soto, and Pablo Padilla-Longoria. Gene regulatory network models for plant development. *Current Opinion in Plant Biology*, 10(1):83–91, 2007.

Elena R Alvarez-Buylla, Eugenio Azpeitia, Rafael Barrio, Mariana Benítez, and Pablo Padilla-Longoria. From abc genes to regulatory networks, epigenetic landscapes and flower morphogenesis: making biological sense of theoretical approaches. In *Seminars in cell & developmental biology*, volume 21, pages 108–117. Elsevier, 2010.

Luca Ambrosino, Chiara Colantuono, Gianfranco Diretto, Alessia Fiore, and Maria Luisa Chiusano. Bioinformatics resources for plant abiotic stress responses: state of the art and opportunities in the fast evolving-omics era. *Plants*, 9(5):591, 2020.

D Anderson and K Burnham. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63 (2020):10, 2004.

Fernando Andrés and George Coupland. The genetic basis of flowering responses to seasonal cues. *Nature Reviews Genetics*, 13(9):627–639, 2012.

Kellie J Archer, Catherine I Dumur, and Viswanathan Ramakrishnan. Graphical technique for identifying a monotonic variance stabilizing transformation for absolute gene intensity signals. *BMC bioinformatics*, 5(1):1–9, 2004.

Tohru Ariizumi, Yoshihito Shinozaki, and Hiroshi Ezura. Genes that influence yield in tomato. *Breeding Science*, 63(1):3–13, 2013.

Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A Fonseca, Robert Petryszak, Irene Papatheodorou, et al. Arrayexpress update–from bulk to single-cell expression data. *Nucleic acids research*, 47(D1):D711–D715, 2019.

Ahmet Ay and David N Arnosti. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical reviews in biochemistry and molecular biology*, 46(2):137–151, 2011.

Ruth E Baker, Jose-Maria Pena, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, 14(5): 20170660, 2018.

Sureshkumar Balasubramanian, Sridevi Sureshkumar, Janne Lempe, and Detlef Weigel. Potent induction of arabidopsis thaliana flowering by elevated growth temperature. *PLoS genetics*, 2(7):e106, 2006.

Marcy J Balunas and A Douglas Kinghorn. Drug discovery from medicinal plants. *Life sciences*, 78(5): 431–441, 2005.

Albert-László Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.

Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, et al. Ncbi geo: archive for functional genomics data sets—10 years on. *Nucleic acids research*, 39(suppl_1):D1005–D1010, 2010.

Rafael Angel Barrio, Aurora Hernandez-Machado, C Varea, Jose Roberto Romero-Arias, and Elena Alvarez-Buylla. Flower development as an interplay between dynamical physical fields and genetic networks. *PloS one*, 5(10):e13523, 2010.

Anna Bartlett, Ronan C O'Malley, Shao-shan Carol Huang, Mary Galli, Joseph R Nery, Andrea Gallavotti, and Joseph R Ecker. Mapping genome-wide transcription-factor binding sites using dap-seq. *Nature protocols*, 12(8):1659–1672, 2017.

Isabel Bäurle and Caroline Dean. The timing of developmental transitions in plants. *Cell*, 125(4):655–664, 2006.

Annette Becker and Günter Theißen. The major clades of mads-box genes and their role in the development and evolution of flowering plants. *Molecular phylogenetics and evolution*, 29(3):464–489, 2003.

Robert I Bertin. Plant phenology and distribution in relation to recent climate change. *The Journal of the Torrey Botanical Society*, 135(1):126–146, 2008.

Miguel A Blázquez, Ji Hoon Ahn, and Detlef Weigel. A thermosensory pathway controlling flowering time in arabidopsis thaliana. *Nature genetics*, 33(2):168–171, 2003.

Frédéric Bouché, Guillaume Lobet, Pierre Tocquin, and Claire Périlleux. Flor-id: an interactive database of flowering-time gene networks in arabidopsis thaliana. *Nucleic Acids Research*, 44(D1):D1167–D1171, 2016.

Benjamin Brachi, Nathalie Faure, Matt Horton, Emilie Flahauw, Adeline Vazquez, Magnus Nordborg, Joy Bergelson, Joel Cuguen, and Fabrice Roux. Linkage and association mapping of arabidopsis thaliana flowering time in nature. *PLoS genetics*, 6(5):e1000940, 2010.

JL Brewster. Environmental physiology of the onion: Towards quantitative models for the effects of photoperiod, temperature and irradiance on bulbing, flowering and growth. In *I International Symposium on Edible Alliaceae 433*, pages 347–374, 1994.

Jelle Bruinsma. *World agriculture: towards 2015/2030: an FAO perspective*. Earthscan, 2017.

Patrick Cahan, Hu Li, Samantha A Morris, Edroaldo Lummertz Da Rocha, George Q Daley, and James J Collins. Cellnet: network biology applied to stem cell engineering. *Cell*, 158(4):903–915, 2014.

Diogo M Camacho, Katherine M Collins, Rani K Powers, James C Costello, and James J Collins. Next-generation machine learning for biological networks. *Cell*, 173(7):1581–1592, 2018.

Giovanna Capovilla, Markus Schmid, and David Posé. Control of flowering by ambient temperature. *Journal of Experimental Botany*, 66(1):59–69, 2015.

Seth Carbon, Amelia Ireland, Christopher J Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, AmiGO Hub, and Web Presence Working Group. Amigo: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.

Jorge J Casal and Sureshkumar Balasubramanian. Thermomorphogenesis. *Annual Review of Plant Biology*, 70:321–346, 2019.

Karen Century, T Lynne Reuber, and Oliver J Ratcliffe. Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. *Plant physiology*, 147(1):20–29, 2008.

Mikhail K Chailakhyan. Internal factors of plant flowering. *Annual Review of Plant Physiology*, 19(1):1–37, 1968.

Seok So Chang, Soon Ki Park, Byung Chul Kim, Bong Joong Kang, Dal Ung Kim, and Hong Gil Nam. Stable genetic transformation of arabidopsis thaliana by agrobacterium inoculation in planta. *The Plant Journal*, 5(4):551–558, 1994.

IPCC Climate Change. The physical science basis: Working group i contribution to the fifth assessment report of the intergovernmental panel on climate change. *K., Tignor, M., Allen, SK, Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, PM, Eds*, page 1535, 2013.

Yin Hoon Chew, Amity M Wilczek, Mathew Williams, Stephen M Welch, Johanna Schmitt, and Karen J Halliday. An augmented arabidopsis phenology model reveals seasonal temperature control of flowering time. *New Phytologist*, 194(3):654–665, 2012.

Yin Hoon Chew, Robert W Smith, Harriet J Jones, Daniel D Seaton, Ramon Grima, and Karen J Halliday. Mathematical models light up plant signaling. *The Plant Cell*, 26(1):5–20, 2014.

P Chouard. Vernalization and its relations to dormancy. *Annual Review of Plant Physiology*, 11(1):191–238, 1960.

Betty YW Chung, Martin Balcerowicz, Marco Di Antonio, Katja E Jaeger, Feng Geng, Krzysztof Franaszek, Poppy Marriott, Ian Brierley, Andrew E Firth, and Philip A Wigge. An rna thermoswitch regulates daytime growth in arabidopsis. *Nature plants*, 6(5):522–532, 2020.

Enrico Coen, Anne-Gaëlle Rolland-Lagan, Mark Matthews, J Andrew Bangham, and Przemyslaw Prusinkiewicz. The genetics of geometry. *Proceedings of the National Academy of Sciences*, 101 (14):4728–4735, 2004.

Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1): D1049–D1056, 2015.

Grant R Cramer, Kaoru Urano, Serge Delrot, Mario Pezzotti, and Kazuo Shinozaki. Effects of abiotic stress on plants: a systems biology perspective. *BMC plant biology*, 11(1):1–14, 2011.

Peter Q Craufurd and Tisha R Wheeler. Climate change and the flowering time of annual crops. *Journal of Experimental botany*, 60(9):2529–2539, 2009.

Amanda J Crawford, Deirdre H McLachlan, Alistair M Hetherington, and Keara A Franklin. High temperature exposure increases plant cooling capacity. *Current Biology*, 22(10):R396–R397, 2012.

Lam Dai Vu, Kris Gevaert, and Ive De Smet. Feeling the heat: searching for plant thermosensors. *Trends in Plant Science*, 24(3):210–219, 2019.

Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717–729, 2010.

Carolin Delker, Marcel Quint, and Philip A Wigge. Recent advances in understanding thermomorphogenesis signaling. *Current Opinion in Plant Biology*, 68:102231, 2022.

Yanglin Ding, Yiting Shi, and Shuhua Yang. Molecular regulation of plant responses to environmental temperatures. *Molecular plant*, 13(4):544–564, 2020.

Jose R Dinneny, Terri A Long, Jean Y Wang, Jee W Jung, Daniel Mace, Solomon Pointer, Christa Barron, Siobhan M Brady, John Schiefelbein, and Philip N Benfey. Cell identity mediates the response of arabidopsis roots to abiotic stress. *Science*, 320(5878):942–945, 2008.

Thomas Dresselhaus and Ralph Hückelhoven. Biotic and abiotic stress responses in crop plants, 2018.

DA Elizondo, RW McClendon, and G Hoogenboom. Neural network models for predicting crop phenology. *Paper-American Society of Agricultural Engineers (USA). no. 923596.*, 1992.

EMBL-EBI. Arrayexpress, functional genomics data, 2023. https://www.ebi.ac.uk/biostudies/arrayexpress.

Jeremiah J Faith, Boris Hayete, Joshua T Thaden, Ilaria Mogno, Jamey Wierzbowski, Guillaume Cottarel, Simon Kasif, James J Collins, and Timothy S Gardner. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

Thilia Ferrier, José Tomás Matus, Jian Jin, and José Luis Riechmann. Arabidopsis paves the way: genomic and network analyses in crops. *Current Opinion in Biotechnology*, 22(2):260–270, 2011.

Catherine Feuillet, Jan E Leach, Jane Rogers, Patrick S Schnable, and Kellye Eversole. Crop genome sequencing: lessons and rationales. *Trends in plant science*, 16(2):77–88, 2011.

Jim C Forbes, James C Forbes, and Drennan Watson. *Plants in agriculture*. Cambridge University Press, 1992.

Fabio Fornara, Amaury de Montaigu, and George Coupland. Snapshot: control of flowering in arabidopsis. *Cell*, 141(3):550–550, 2010.

Duncan T Forster, Sheena C Li, Yoko Yashiroda, Mami Yoshimura, Zhijian Li, Luis Alberto Vega Isuhuaylas, Kaori Itto-Nakama, Daisuke Yamanaka, Yoshikazu Ohya, Hiroyuki Osada, et al. Bionic: biological network integration using convolutions. *Nature Methods*, 19(10):1250–1261, 2022.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

Lei Gan, Zhenzhen Wei, Zuoren Yang, Fuguang Li, and Zhi Wang. Updated mechanisms of gcn5—the monkey king of the plant kingdom in plant development and resistance to abiotic stresses. *Cells*, 10 (5):979, 2021.

Malia A Gehan, Kathleen Greenham, Todd C Mockler, and C Robertson McClung. Transcriptional networks—crops, clocks, and abiotic stress. *Current opinion in plant biology*, 24:39–46, 2015.

Kyung-Eun Gil and Chung-Mo Park. Thermal adaptation and plasticity of the plant circadian clock. *New Phytologist*, 221(3):1215–1229, 2019.

A Gobin. Impact of heat and drought stress on arable crop production in belgium. *Natural Hazards and Earth System Sciences*, 12(6):1911–1922, 2012.

H Charles J Godfray, John R Beddington, Ian R Crute, Lawrence Haddad, David Lawrence, James F Muir, Jules Pretty, Sherman Robinson, Sandy M Thomas, and Camilla Toulmin. Food security: the challenge of feeding 9 billion people. *science*, 327(5967):812–818, 2010.

Marc Goetz, Lauren C Hooper, Susan D Johnson, Julio Carlyle Macedo Rodrigues, Adam Vivian-Smith, and Anna M Koltunow. Expression of aberrant forms of auxin response factor8 stimulates parthenocarpy in arabidopsis and tomato. *Plant physiology*, 145(2):351–366, 2007.

Aaron D Goldman, Ram Samudrala, and John A Baross. The evolution and functional repertoire of translation proteins following the origin of life. *Biology direct*, 5(1):1–12, 2010.

Christine Granier, Catherine Massonnet, Olivier Turc, Bertrand Muller, Karine Chenu, and Francois Tardieu. Individual leaf development in arabidopsis thaliana: a stable thermal-time-based programme. *Annals of botany*, 89(5):595–604, 2002.

Xiaoyu Guo, Dongfeng Liu, and Kang Chong. Cold signaling in plants: Insights into mechanisms and regulation. *Journal of integrative plant biology*, 60(9):745–756, 2018.

P Hadley, EH Roberts, RJ Summerfield, and FR Minchin. A quantitative model of reproductive development in cowpea [vigna unguiculata (l) walp.] in relation to photoperiod and temperature, and implications for screening germplasm. *Annals of Botany*, 51(4):531–543, 1983.

P Hadley, EH Roberts, RJ Summerfield, and FR Minchin. Effects of temperature and photoperiod on flowering in soya bean [glycine max (l.) merrill]: a quantitative model. *Annals of Botany*, 53(5):669–681, 1984.

Samiul Haque, Jabeen S Ahmad, Natalie M Clark, Cranos M Williams, and Rosangela Sozzani. Computational prediction of gene regulatory networks in plant growth and development. *Current opinion in plant biology*, 47:96–105, 2019.

David R Harris and Dorian Q Fuller. Agriculture: definition and overview. *Encyclopedia of global archaeology*, pages 104–113, 2014.

Ulrike Hartmann, Susanne Höhmann, Klaus Nettesheim, Ellen Wisman, Heinz Saedler, and Peter Huijser. Molecular cloning of svp: a negative regulator of the floral transition in arabidopsis. *The Plant Journal*, 21(4):351–360, 2000.

Mirza Hasanuzzaman, Kamrun Nahar, Md Mahabub Alam, Rajib Roychowdhury, and Masayuki Fujita. Physiological, biochemical, and molecular mechanisms of heat stress tolerance in plants. *International journal of molecular sciences*, 14(5):9643–9684, 2013.

M Hasdai, B Weiss, A Levi, A Samach, and R Porat. Differential responses of arabidopsis ecotypes to cold, chilling and freezing temperatures. *Annals of Applied Biology*, 148(2):113–120, 2006.

Jerry L Hatfield and John H Prueger. Temperature extremes: Effect on plant growth and development. *Weather and climate extremes*, 10:4–10, 2015.

Yuehui He and Richard M Amasino. Role of chromatin modification in flowering-time control. *Trends in plant science*, 10(1):30–35, 2005.

Yuehui He, Tao Chen, and Xiaolin Zeng. Genetic and epigenetic understanding of the seasonal timing of flowering. *Plant communications*, 1(1):100008, 2020.

Chris A Helliwell, Craig C Wood, Masumi Robertson, W James Peacock, and Elizabeth S Dennis. The arabidopsis flc protein interacts directly in vivo with soc1 and ft chromatin and is part of a high-molecular-weight protein complex. *The Plant Journal*, 46(2):183–192, 2006.

Jae Bok Heo and Sibum Sung. Vernalization-mediated epigenetic silencing by a long intronic noncoding rna. *Science*, 331(6013):76–79, 2011.

Janet A Higgins, Paul C Bailey, and David A Laurie. Comparative genomics of flowering time pathways using brachypodium distachyon as a model for the temperate grasses. *PLoS one*, 5(4):e10065, 2010.

Archibald Vivian Hill. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *j. physiol.*, 40:iv–vii, 1910.

Camilla B Hill and Chengdao Li. Genetic architecture of flowering phenology in cereals and opportunities for crop improvement. *Frontiers in plant science*, 7:1906, 2016.

Tom W Hiscock. Adapting machine-learning algorithms to design gene circuits. *BMC bioinformatics*, 20: 1–13, 2019.

Stefan Hoops, Raquel Hontecillas, Vida Abedi, Andrew Leber, Casandra Philipson, Adria Carbo, and Josep Bassaganya-Riera. Ordinary differential equations (odes) based modeling. In *Computational Immunology*, pages 63–78. Elsevier, 2016.

Jui-Hung Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in bioinformatics*, 13(3):281–291, 2012.

Richard GH Immink, David Posé, Silvia Ferrario, Felix Ott, Kerstin Kaufmann, Felipe Leal Valentim, Stefan De Folter, Froukje Van der Wal, Aalt DJ van Dijk, Markus Schmid, et al. Characterization of soc1's central role in flowering by the identification of its upstream and downstream regulators. *Plant physiology*, 160 (1):433–449, 2012.

Ravi Iyengar. Why we need quantitative dynamic models, 2009.

Pierre Jacob, Adi Avni, and Abdelhafid Bendahmane. Translational research: exploring and creating genetic diversity. *Trends in Plant Science*, 23(1):42–52, 2018.

Seonghoe Jang, Stefano Torti, and George Coupland. Genetic and spatial interactions between ft, tsf and svp during the early stages of floral induction in arabidopsis. *The Plant Journal*, 60(4):614–625, 2009.

Chih-Hung Jen, Iain W Manfield, Ioannis Michalopoulos, John W Pinney, William GT Willats, Philip M Gilmartin, and David R Westhead. The arabidopsis co-expression tool (act): a www-based tool and database for microarray-based gene expression analysis. *The Plant Journal*, 46(2):336–348, 2006.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Christian Jung and Andreas E Müller. Flowering time control and applications in plant breeding. *Trends in plant science*, 14(10):563–573, 2009.

Jae-Hoon Jung, Mirela Domijan, Cornelia Klose, Surojit Biswas, Daphne Ezer, Mingjun Gao, Asif Khan Khattak, Mathew S Box, Varodom Charoensawan, Sandra Cortijo, et al. Phytochromes function as thermosensors in arabidopsis. *Science*, 354(6314):886–889, 2016.

Kerstin Kaufmann, Frank Wellmer, Jose M Muiño, Thilia Ferrier, Samuel E Wuest, Vijaya Kumar, Antonio Serrano-Mislata, Francisco Madueno, Pawel Krajewski, Elliot M Meyerowitz, et al. Orchestration of floral initiation by apetala1. *science*, 328(5974):85–89, 2010.

Anne E Kelly and Michael L Goulden. Rapid shifts in plant distribution with recent climate change. *Proceedings of the national academy of sciences*, 105(33):11823–11826, 2008.

Jae Joon Kim, Jeong Hwan Lee, Wanhui Kim, Hye Seung Jung, Peter Huijser, and Ji Hoon Ahn. The microrna156-squamosa promoter binding protein-like3 module regulates ambient temperature-responsive flowering via flowering locus t in arabidopsis. *Plant Physiology*, 159(1):461–478, 2012.

Hannah A Kinmonth-Schultz, Melissa J MacEwen, Daniel D Seaton, Andrew J Millar, Takato Imaizumi, and Soo-Hyung Kim. Mechanistic model of temperature influence on flowering through whole-plant accumulation of ft. *BioRxiv*, page 267104, 2018.

Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

M Koornneef, CJ Hanhart, and JH Van der Veen. A genetic and physiological analysis of late flowering mutants in arabidopsis thaliana. *Molecular and General Genetics MGG*, 229(1):57–66, 1991.

Christian Körner. Significance of temperature in plant life. *Plant growth and climate change*, pages 48–69, 2006.

Paolo Korwin Krukowski, Jan Ellenberger, Simone Röhlen-Schmittgen, Andrea Schubert, and Francesca Cardinale. Phenotyping in arabidopsis and crops—are we addressing the same traits? a case study in tomato. *Genes*, 11(9):1011, 2020.

Stephen Kotiang and Ali Eslami. A probabilistic graphical model for system-wide analysis of gene regulatory networks. *Bioinformatics*, 36(10):3192–3199, 2020.

Joel A Kreps, Yajun Wu, Hur-Song Chang, Tong Zhu, Xun Wang, and Jeff F Harper. Transcriptome changes for arabidopsis in response to salt, osmotic, and cold stress. *Plant physiology*, 130(4):2129–2141, 2002.

Heidi Kuehn, Arthur Liberzon, Michael Reich, and Jill P Mesirov. Using genepattern for gene expression analysis. *Current protocols in bioinformatics*, 22(1):7–12, 2008.

Shubhada R Kulkarni, Dries Vaneechoutte, Jan Van de Velde, and Klaas Vandepoele. Tf2network: predicting transcription factor regulators and gene regulatory networks in arabidopsis using publicly available binding site information. *Nucleic acids research*, 46(6):e31–e31, 2018.

S Vinod Kumar and Philip A Wigge. H2a. z-containing nucleosomes mediate the thermosensory response in arabidopsis. *Cell*, 140(1):136–147, 2010.

Dorota Kwiatkowska. Flowering and apical meristem growth dynamics. *Journal of Experimental Botany*, 59(2):187–201, 2008.

Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13, 2008.

Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008.

Douglas A Lauffenburger. Cell signaling pathways as control modules: complexity for simplicity? *Proceedings of the National Academy of Sciences*, 97(10):5031–5033, 2000.

Felipe Leal Valentim, Simon van Mourik, David Pose, Min C Kim, Markus Schmid, Roeland CHJ van Ham, Marco Busscher, Gabino F Sanchez-Perez, Jaap Molenaar, Gerco C Angenent, et al. A quantitative and dynamic model of the arabidopsis flowering time gene regulatory network. *PloS one*, 10(2):e0116973, 2015.

Jeong Hwan Lee, Seong Jeon Yoo, Soo Hyun Park, Ildoo Hwang, Jong Seob Lee, and Ji Hoon Ahn. Role of svp in the control of flowering time by ambient temperature in arabidopsis. *Genes & development*, 21(4):397–402, 2007.

Jungeun Lee, Mijin Oh, Hanna Park, and Ilha Lee. Soc1 translocated to the nucleus by interaction with agl24 directly regulates leafy. *The Plant Journal*, 55(5):832–843, 2008.

Christian Møldrup Legaard, Thomas Schranz, Gerald Schweiger, Ján Drgoňa, Basak Falay, Cláudio Gomes, Alexandros Iosifidis, Mahdi Abkar, and Peter Gorm Larsen. Constructing neural network-based models for simulating dynamical systems. *ACM Computing Surveys*, 2021.

Willeke Leijten, Ronald Koes, Ilja Roobeek, and Giovanna Frugis. Translating flowering time from arabidopsis thaliana to brassicaceae and asteraceae crop species. *Plants*, 7(4):111, 2018.

Laura Lema-Perez, Rafael Muñoz-Tamayo, Jose Garcia-Tirado, and Hernan Alvarez. On parameter interpretability of phenomenological-based semiphysical models in biology. *Informatics in Medicine Unlocked*, 15:100158, 2019.

Louise A Lewis and Richard M McCourt. Green algae and the origin of land plants. *American journal of botany*, 91(10):1535–1556, 2004.

Bingjie Li, Kang Gao, Huimin Ren, and Wenqiang Tang. Molecular mechanisms governing plant responses to high temperatures. *Journal of Integrative Plant Biology*, 60(9):757–779, 2018.

Dan Li, Chang Liu, Lisha Shen, Yang Wu, Hongyan Chen, Masumi Robertson, Chris A Helliwell, Toshiro Ito, Elliot Meyerowitz, and Hao Yu. A repressor complex governs the integration of flowering signals in arabidopsis. *Developmental cell*, 15(1):110–120, 2008.

Yupeng Li, Stephanie A Pearl, and Scott A Jackson. Gene networks in plant biology: approaches in reconstruction and analysis. *Trends in plant science*, 20(10):664–675, 2015.

Jörg Linde, Sylvie Schulze, Sebastian G Henkel, and Reinhard Guthke. Data-and knowledge-based modeling of gene regulatory networks: an update. *EXCLI journal*, 14:346, 2015.

Lu Liu, Chang Liu, Xingliang Hou, Wanyan Xi, Lisha Shen, Zhen Tao, Yue Wang, and Hao Yu. Ftip1 is an essential regulator required for florigen transport. *PLoS biology*, 10(4):e1001313, 2012.

James CW Locke, Megan M Southern, László Kozma-Bognár, Victoria Hibberd, Paul E Brown, Matthew S Turner, and Andrew J Millar. Extension of a genetic network model by iterative experimentation and mathematical analysis. *Molecular systems biology*, 1(1):2005–0013, 2005.

Jeff Long and M Kathryn Barton. Initiation of axillary and floral meristems in arabidopsis. *Developmental biology*, 218(2):341–353, 2000.

Allison J Lopatkin and James J Collins. Predictive biology: modelling, understanding and harnessing microbial complexity. *Nature Reviews Microbiology*, 18(9):507–520, 2020.

Christian D Lorenzo, Maximiliano Sanchez-Lamas, Mariana S Antonietti, and Pablo D Cerdán. Emerging hubs in plant light and temperature signaling. *Photochemistry and photobiology*, 92(1):3–13, 2016.

Anna Lovrics, Yu Gao, Bianka Juhász, István Bock, Helen M Byrne, András Dinnyés, and Krisztián A Kovács. Boolean modelling reveals new regulatory connections between transcription factors orchestrating the development of the ventral spinal cord. *PloS one*, 9(11):e111430, 2014.

Mikael Lucas, Laurent Laplaze, and Malcolm J Bennett. Plant systems biology: network matters. *Plant, cell & environment*, 34(4):535–553, 2011.

Nicholas M Luscombe, Dov Greenbaum, and Mark Gerstein. What is bioinformatics? an introduction and overview. *Yearbook of medical informatics*, 10(01):83–100, 2001.

Ulrich Lutz, David Pose, Matthias Pfeifer, Heidrun Gundlach, Joerg Hagmann, Congmao Wang, Detlef Weigel, Klaus FX Mayer, Markus Schmid, and Claus Schwechheimer. Modulation of ambient temperature-dependent flowering in arabidopsis thaliana by natural variation of flowering locus m. *PLoS Genetics*, 11(10):e1005588, 2015.

Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. Springer, 2006.

John C Mathers. Plant foods for human health: research challenges. *Proceedings of the Nutrition Society*, 65(2):198–203, 2006.

Rajesh Mehrotra, Gauri Gupta, Riccha Sethi, Purva Bhalothia, Narayan Kumar, and Sandhya Mehrotra. Designer promoter: an artwork of cis engineering. *Plant molecular biology*, 75(6):527–536, 2011.

David W Meinke, J Michael Cherry, Caroline Dean, Steven D Rounsley, and Maarten Koornneef. Arabidopsis thaliana: a model plant for genome analysis. *Science*, 282(5389):662–682, 1998.

Maria Katherine Mejia-Guerra, Marcelo Pomeranz, Kengo Morohashi, and Erich Grotewold. From plant gene regulatory grids to network dynamics. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(5):454–465, 2012.

Huaiyu Mi, Anushya Muruganujan, John T Casagrande, and Paul D Thomas. Large-scale gene function analysis with the panther classification system. *Nature protocols*, 8(8):1551–1566, 2013.

Huaiyu Mi, Dustin Ebert, Anushya Muruganujan, Caitlin Mills, Laurent-Philippe Albou, Tremayne Mushayamaha, and Paul D Thomas. Panther version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive api. *Nucleic acids research*, 49(D1):D394–D403, 2021.

Todd P Michael, Todd C Mockler, Ghislain Breton, Connor McEntee, Amanda Byer, Jonathan D Trout, Samuel P Hazen, Rongkun Shen, Henry D Priest, Christopher M Sullivan, et al. Network discovery pipeline elucidates conserved time-of-day–specific cis-regulatory modules. *PLoS genetics*, 4(2):e14, 2008.

Scott D Michaels. Flowering time regulation produces much fruit. *Current opinion in plant biology*, 12(1): 75–80, 2009.

Scott D Michaels and Richard M Amasino. Flowering locus c encodes a novel mads domain protein that acts as a repressor of flowering. *The Plant Cell*, 11(5):949–956, 1999.

Scott D Michaels, Gary Ditta, Cindy Gustafson-Brown, Soraya Pelaz, Martin Yanofsky, and Richard M Amasino. Agl24 acts as a promoter of flowering in arabidopsis and is positively regulated by vernalization. *The Plant Journal*, 33(5):867–874, 2003.

Scott D Michaels, Isabel C Bezerra, and Richard M Amasino. Frigida-related genes are required for the winter-annual habit in arabidopsis. *Proceedings of the National Academy of Sciences*, 101(9):3281–3285, 2004.

Paramjit Singh Minhas, Jagadish Rane, and Ratna Kumar Pasala. Abiotic stresses in agriculture: An overview. *Abiotic stress management for resilient agriculture*, pages 3–8, 2017.

Peter V Minorsky. Achieving the in silico plant. systems biology and the future of plant biological research, 2003.

RAC Mitchell, VJ Mitchell, SP Driscoll, J Franklin, and DW Lawlor. Effects of increased co2 concentration and temperature on growth and yield of winter wheat at two levels of nitrogen application. *Plant, Cell & Environment*, 16(5):521–529, 1993.

Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, 14(10):719–732, 2013.

Ron Mittler, Eduardo Blumwald, et al. Genetic engineering for modern agriculture: challenges and perspectives. *Annual review of plant biology*, 61(1):443–462, 2010.

Ludmila Mlynárová, Jan-Peter Nap, and Ton Bisseling. The swi/snf chromatin-remodeling gene atchr12 mediates temporary growth arrest in arabidopsis thaliana upon perceiving environmental stress. *The Plant Journal*, 51(5):874–885, 2007.

JL Monteith. Climatic variation and the growth of crops. *Quarterly Journal of the Royal Meteorological Society*, 107(454):749–774, 1981.

Jihyun Moon, Horim Lee, Minsoo Kim, and Ilha Lee. Analysis of flowering pathway integrators in arabidopsis. *Plant and Cell Physiology*, 46(2):292–299, 2005.

Miguel A Moreno-Risueno, Wolfgang Busch, and Philip N Benfey. Omics meet networks—using systems approaches to infer regulatory networks in plants. *Current opinion in plant biology*, 13(2):126–131, 2010.

James IL Morison and Michael D Morecroft. *Plant growth and climate change*. John Wiley & Sons, 2008.

Iftach Nachman, Aviv Regev, and Nir Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(suppl_1):i248–i256, 2004.

Kazuo Nakashima, Yusuke Ito, and Kazuko Yamaguchi-Shinozaki. Transcriptional regulatory networks in response to abiotic stresses in arabidopsis and grasses. *Plant physiology*, 149(1):88–95, 2009.

NCBI. Gene expression omnibus, 2023. https://www.ncbi.nlm.nih.gov/geo/.

Chris J Needham, Iain W Manfield, Andrew J Bulpitt, Philip M Gilmartin, and David R Westhead. From gene expression to gene regulatory networks in arabidopsis thaliana. *BMC systems biology*, 3(1):1–18, 2009.

Hilde Nelissen, Maurice Moloney, and Dirk Inzé. Translational research: from pot to plot. *Plant biotechnology journal*, 12(3):277–285, 2014.

Donald E Nelson, Peter P Repetti, Tom R Adams, Robert A Creelman, Jingrui Wu, David C Warner, Don C Anstrom, Robert J Bensen, Paolo P Castiglioni, Meghan G Donnarummo, et al. Plant nuclear factor y (nf-y) b subunits confer drought tolerance and lead to improved corn yields on water-limited acres. *Proceedings of the National Academy of Sciences*, 104(42):16450–16455, 2007.

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

Anand Nishad and Ashis Kumar Nandi. Recent advances in plant thermomemory. *Plant Cell Reports*, 40 (1):19–27, 2021.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384, 2013.

Ian R Noble, Saleemul Huq, Yuri A Anokhin, J Carmin, Dieudonne Goudou, Felino P Lansigan, Balgis Osman-Elasha, and Alicia Villamizar. *limate change 2014: impacts, adaptation, and vulnerability. Part*

A: Global and sectoral aspects. Contribution of working group II to the fifth assessment report of the intergovernmental panel on climate change, chapter Adaptation needs and options. Cambridge University Press, 2014.

Elad Noor, Sarah Cherkaoui, and Uwe Sauer. Biological insights through omics data integration. *Current Opinion in Systems Biology*, 15:39–47, 2019.

Andrea Ocone, Laleh Haghverdi, Nikola S Mueller, and Fabian J Theis. Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–i96, 2015.

Damian R Page and Ueli Grossniklaus. The art and design of genetic screens: Arabidopsis thaliana. *Nature Reviews Genetics*, 3(2):124–136, 2002.

Camille Parmesan and Mick E Hanley. Plantsland climate change: complexities and surprises. *Annals of botany*, 116(6):849–864, 2015.

Ripon Patgiri, Ganesh Chandra Deka, and Anupam Biswas. Principles of big graph: In-depth insight. 2023.

G Pei, L Chen, and W Zhang. Wgcna application to proteomic and metabolomic data analysis. In *Methods in enzymology*, volume 585, pages 135–158. Elsevier, 2017.

Dolores R Piperno. The origins of plant cultivation and domestication in the new world tropics: patterns, process, and new developments. *Current anthropology*, 52(S4):S453–S470, 2011.

Athanasios Polynikis, SJ Hogan, and Mario Di Bernardo. Comparing different ode modelling approaches for gene regulatory networks. *Journal of theoretical biology*, 261(4):511–530, 2009.

John R Porter and Mikhail A Semenov. Crop responses to climatic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1463):2021–2035, 2005.

John R Porter, Liyong Xie, Andrew J Challinor, Kevern Cochrane, S Mark Howden, Muhammad Mohsin Iqbal, David B Lobell, and Maria Isabel Travasso. *Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter Food security and food production systems. Cambridge University Press, 2014.

David Pose, Leonie Verhage, Felix Ott, Levi Yant, Johannes Mathieu, Gerco C Angenent, Richard GH Immink, and Markus Schmid. Temperature-dependent regulation of flowering by antagonistic flm variants. *Nature*, 503(7476):414–417, 2013.

Kristin V Presnell and Hal S Alper. Systems metabolic engineering meets machine learning: A new era for data-driven metabolic engineering. *Biotechnology journal*, 14(9):1800416, 2019.

Nicholas J Provart, Jose Alonso, Sarah M Assmann, Dominique Bergmann, Siobhan M Brady, Jelena Brkljacic, John Browse, Clint Chapple, Vincent Colot, Sean Cutler, et al. 50 years of arabidopsis research: highlights and future directions. *New Phytologist*, 209(3):921–944, 2016.

Marcel Quint, Carolin Delker, Keara A Franklin, Philip A Wigge, Karen J Halliday, and Martijn Van Zanten. Molecular and genetic control of plant thermomorphogenesis. *Nature plants*, 2(1):1–9, 2016.

Tomas Radivoyevitch. Mass action models versus the hill model: An analysis of tetrameric human thymidine kinase 1 positive cooperativity. *Biology Direct*, 4:1–17, 2009.

Omar Rafique and Ajaz Hussain Mir. Weighted dimensionality reduction and robust gaussian mixture model based cancer patient subtyping from gene expression data. *Journal of Biomedical Informatics*, 112:103620, 2020.

Oliver J Ratcliffe, Roderick W Kumimoto, Becky J Wong, and José Luis Riechmann. Analysis of the arabidopsis mads affecting flowering gene family: Maf2 prevents vernalization by short periods of cold. *The Plant Cell*, 15(5):1159–1169, 2003.

Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P Mesirov. Genepattern 2.0. *Nature genetics*, 38(5):500–501, 2006.

Willem A Rensink and C Robin Buell. Arabidopsis to rice. applying knowledge from a weed to enhance our understanding of a crop species. *Plant Physiology*, 135(2):622–629, 2004.

Narges Rezaie, Farilie Reese, and Ali Mortazavi. Pywgcna: A python package for weighted gene co-expression network analysis. *Bioinformatics*, page btad415, 2023.

Seung Y Rhee, Kenneth D Birnbaum, and David W Ehrhardt. Towards building a plant cell atlas. *Trends in plant science*, 24(4):303–310, 2019.

Seung Yon Rhee, William Beavis, Tanya Z Berardini, Guanghong Chen, David Dixon, Aisling Doyle, Margarita Garcia-Hernandez, Eva Huala, Gabriel Lander, Mary Montoya, et al. The arabidopsis information

resource (tair): a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community. *Nucleic acids research*, 31(1):224–228, 2003.

Simone Romano, Nguyen Xuan Vinh, James Bailey, and Karin Verspoor. Adjusting for chance clustering comparison measures. *The Journal of Machine Learning Research*, 17(1):4635–4666, 2016.

Kook Hui Ryu, Ling Huang, Hyun Min Kang, and John Schiefelbein. Single-cell rna sequencing resolves molecular relationships among individual plant cells. *Plant physiology*, 179(4):1444–1456, 2019.

Tomoaki Sakamoto and Seisuke Kimura. Plant temperature sensors. *Sensors*, 18(12):4365, 2018.

Yoh Sakuma, Kyonoshin Maruyama, Feng Qin, Yuriko Osakabe, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. Dual function of an arabidopsis transcription factor dreb2a in water-stress-responsive and heat-stress-responsive gene expression. *Proceedings of the National Academy of Sciences*, 103(49): 18822–18827, 2006.

José Domingo Salazar, Treenut Saithong, Paul E Brown, Julia Foreman, James CW Locke, Karen J Halliday, Isabelle A Carré, David A Rand, and Andrew J Millar. Prediction of photoperiodic regulators from quantitative gene circuit models. *Cell*, 139(6):1170–1179, 2009.

Yara-Elena Sanchez-Corrales, Elena R Alvarez-Buylla, and Luis Mendoza. The arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process. *Journal of theoretical biology*, 264(3):971–983, 2010.

David C Sands, Cindy E Morris, Edward A Dratz, and Alice L Pilgeram. Elevating optimal human nutrition to a central goal of plant breeding and production of plant-based foods. *Plant Science*, 177(5):377–389, 2009.

Akiko Satake, Tetsuhiro Kawagoe, Yukari Saburi, Yukako Chiba, Gen Sakurai, and Hiroshi Kudoh. Forecasting flowering phenology under climate warming by modelling the regulatory dynamics of flowering-time genes. *Nature communications*, 4(1):2303, 2013.

Heiko Schoof, Rebecca Ernst, Vladimir Nazarov, Lukas Pfeifer, Hans-Werner Mewes, and Klaus FX Mayer. Mips arabidopsis thaliana database (matdb): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Research*, 32(suppl_1):D373–D376, 2004.

Franziska Schramm, Jane Larkindale, Elke Kiehlmann, Arnab Ganguli, Gisela Englich, Elizabeth Vierling, and Pascal Von Koskull-Döring. A cascade of transcription factor dreb2a and heat stress transcription

factor hsfa3 regulates the heat stress response of arabidopsis. *The Plant Journal*, 53(2):264–274, 2008.

Katia Scortecci, Scott D Michaels, and Richard M Amasino. Genetic interactions between flm and other flowering-time genes in arabidopsis thaliana. *Plant molecular biology*, 52(5):915–922, 2003.

Motoaki Seki, Mari Narusaka, Junko Ishida, Tokihiko Nanjo, Miki Fujita, Youko Oono, Asako Kamiya, Maiko Nakajima, Akiko Enju, Tetsuya Sakurai, et al. Monitoring the expression profiles of 7000 arabidopsis genes under drought, cold and high-salinity stresses using a full-length cdna microarray. *The Plant Journal*, 31(3):279–292, 2002.

Ümit Seren, Dominik Grimm, Joffrey Fitz, Detlef Weigel, Magnus Nordborg, Karsten Borgwardt, and Arthur Korte. Arapheno: a public database for arabidopsis thaliana phenotypes. *Nucleic Acids Research*, page gkw986, 2016.

Charlotte Seydel, Julia Biener, Vladimir Brodsky, Svenja Eberlein, and Thomas Nägele. Predicting plant growth response under fluctuating temperature by carbon balance modelling. *Communications biology*, 5(1):1–13, 2022.

Candice C Sheldon, Dean T Rouse, E Jean Finnegan, W James Peacock, and Elizabeth S Dennis. The molecular basis of vernalization: the central role of flowering locus c (flc). *Proceedings of the National Academy of Sciences*, 97(7):3753–3758, 2000.

Rinku Singh and GS Singh. Traditional agriculture: a climate-smart approach for sustainable food production. *Energy, Ecology and Environment*, 2(5):296–316, 2017.

R Sivasubramanian, Nitika Mukhi, and Jagreet Kaur. Arabidopsis thaliana: a model for plant research. In *Plant biology and biotechnology*, pages 1–26. Springer, 2015.

Nataliya Sokolovska, Olga Permiakova, Sofia K Forslund, and Jean-Daniel Zucker. A semi-supervised approach to discover bivariate causality in large biological data. In *Machine Learning and Data Mining in Pattern Recognition: 14th International Conference, MLDM 2018, New York, NY, USA, July 15-19, 2018, Proceedings, Part I 14*, pages 406–420. Springer, 2018.

Susan Solomon, Dahe Qin, Martin Manning, Kristen Averyt, and Melinda Marquis. *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*, volume 4. Cambridge university press, 2007.

Donald F Specht. Probabilistic neural networks. *Neural networks*, 3(1):109–118, 1990.

N Sreenivasulu, SK Sopory, and PB Kavi Kishor. Deciphering the regulatory mechanisms of abiotic stress tolerance in plants by genomic approaches. *Gene*, 388(1-2):1–13, 2007.

Napaporn Sriden and Varodom Charoensawan. Large-scale comparative transcriptomic analysis of temperature-responsive genes in arabidopsis thaliana. *Plant Molecular Biology*, pages 1–19, 2022.

Anusha Srikanth and Markus Schmid. Regulation of flowering time: all roads lead to rome. *Cellular and molecular life sciences*, 68(12):2013–2037, 2011.

Jörg Stelling and Ernst Dieter Gilles. Mathematical modeling of complex regulatory networks. *IEEE Transactions on NanoBioscience*, 3(3):172–179, 2004.

Eric J Stockinger, Sarah J Gilmour, and Michael F Thomashow. Arabidopsis thaliana cbf1 encodes an ap2 domain-containing transcriptional activator that binds to the c-repeat/dre, a cis-acting dna regulatory element that stimulates transcription in response to low temperature and water deficit. *Proceedings of the National Academy of Sciences*, 94(3):1035–1040, 1997.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

Ying Sun and José R Dinneny. Q&a: How do gene regulatory networks control environmental responses in plants? *BMC biology*, 16(1):1–4, 2018.

Sridevi Sureshkumar, Craig Dent, Andrei Seleznev, Celine Tasset, and Sureshkumar Balasubramanian. Nonsense-mediated mrna decay modulates flm-dependent thermosensory flowering response in arabidopsis. *Nature Plants*, 2(5):1–7, 2016.

Nobuhiro Suzuki, Hiroe Sejima, Rachel Tam, Karen Schlauch, and Ron Mittler. Identification of the mbf1 heat-response regulon of arabidopsis thaliana. *The Plant Journal*, 66(5):844–851, 2011.

Daniel R Taub, Brian Miller, and Holly Allen. Effects of elevated co2 on the protein concentration of food crops: a meta-analysis. *Global Change Biology*, 14(3):565–575, 2008.

Nuno Tenazinha and Susana Vinga. A survey on methods for modeling and analyzing integrated biological networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(4):943–958, 2010.

Hrudayanath Thatoi and Jayanta Kumar Patra. Biotechnology and pharmacological evaluation of medicinal plants: An overview. *Journal of herbs, spices & medicinal plants*, 17(3):214–248, 2011.

Michael F Thomashow. Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. *Annual review of plant biology*, 50(1):571–599, 1999.

Andrea Thorn. Artificial intelligence in the experimental determination and prediction of macromolecular structures. *Current Opinion in Structural Biology*, 74:102368, 2022.

Wilfried Thuiller, Sandra Lavorel, Miguel B Araújo, Martin T Sykes, and I Colin Prentice. Climate change threats to plant diversity in europe. *Proceedings of the National Academy of Sciences*, 102(23):8245–8250, 2005.

Matteo Togninalli, Ümit Seren, Dazhe Meng, Joffrey Fitz, Magnus Nordborg, Detlef Weigel, Karsten Borgwardt, Arthur Korte, and Dominik G Grimm. The aragwas catalog: a curated and standardized arabidopsis thaliana gwas catalog. *Nucleic acids research*, 46(D1):D1150–D1156, 2018.

Lam-Son Phan Tran, Kazuo Nakashima, Kazuo Shinozaki, and Kazuko Yamaguchi-Shinozaki. Plant gene networks in osmotic stress response: from genes to regulatory networks. *Methods in enzymology*, 428: 109–128, 2007.

Rajiv K Tripathi and Olivia Wilkins. Single cell gene regulatory networks in plants: Opportunities for enhancing climate change stress resilience. *Plant, Cell & Environment*, 44(7):2006–2017, 2021.

Hiroyuki Tsuji, Shojiro Tamaki, Reina Komiya, and Ko Shimamoto. Florigen and the photoperiodic control of flowering in rice. *Rice*, 1(1):25–35, 2008.

Harrij Van Velthuizen. *Mapping biophysical factors that influence agricultural production and rural vulnerability*. Number 11. Food & Agriculture Org., 2007.

Klaas Vandepoele, Mauricio Quimbaya, Tine Casneuf, Lieven De Veylder, and Yves Van de Peer. Unraveling transcriptional control in arabidopsis using cis-regulatory elements and coexpression networks. *Plant physiology*, 150(2):535–546, 2009.

Leonie Verhage, Gerco C Angenent, and Richard GH Immink. Research on floral timing by ambient temperature comes into blossom. *Trends in plant science*, 19(9):583–591, 2014.

Isaac Virshup, Sergei Rybakov, Fabian J Theis, Philipp Angerer, and F Alexander Wolf. anndata: Annotated data. *BioRxiv*, pages 2021–12, 2021.

Michael Jeffrey Volk, Ismini Lourentzou, Shekhar Mishra, Lam Tung Vo, Chengxiang Zhai, and Huimin Zhao. Biosystems design by machine learning. *ACS synthetic biology*, 9(7):1514–1533, 2020.

Joachim Von Braun. *The world food situtation: new driving forces and required actions*. Intl Food Policy Res Inst, 2007.

Xu-Wen Wang, Dandi Qiao, Michael H Cho, Dawn L DeMeo, Edwin K Silverman, and Yang-Yu Liu. A statistical physics approach for disease module detection. *Genome Research*, 32(10):1918–1929, 2022.

Xuewei Wang, Ertugrul Dalkic, Ming Wu, and Christina Chan. Gene module level analysis: identification to networks and dynamics. *Current opinion in biotechnology*, 19(5):482–491, 2008.

Hairong Wei, Staffan Persson, Tapan Mehta, Vinodh Srinivasasainagendra, Lang Chen, Grier P Page, Chris Somerville, and Ann Loraine. Transcriptional coordination of the metabolic network in arabidopsis. *Plant physiology*, 142(2):762–774, 2006.

Detlef Weigel and Jane Glazebrook. Genetic analysis of arabidopsis mutants. *Cold Spring Harbor Protocols*, 2008(3):pdb–top35, 2008.

Detlef Weigel, John Alvarez, David R Smyth, Martin F Yanofsky, and Elliot M Meyerowitz. Leafy controls floral meristem identity in arabidopsis. *Cell*, 69(5):843–859, 1992.

Stephen M Welch, Judith L Roe, and Zhanshan Dong. A genetic neural network model of flowering time control in arabidopsis thaliana. *Agronomy Journal*, 95(1):71–81, 2003.

Frank Wellmer and José L Riechmann. Gene networks controlling the initiation of flower development. *Trends in Genetics*, 26(12):519–527, 2010.

David J Weston, Lee E Gunter, Alistair Rogers, and Stan D Wullschleger. Connecting genes, coexpression modules, and molecular signatures to environmental stress phenotypes in plants. *BMC Systems Biology*, 2:1–17, 2008.

Tim Wheeler and Joachim Von Braun. Climate change impacts on global food security. *Science*, 341 (6145):508–513, 2013.

TR Wheeler, GR Batts, RH Ellis, P Hadley, and JIL Morison. Growth and yield of winter wheat (triticum aestivum) crops in response to co2 and temperature. *The Journal of Agricultural Science*, 127(1): 37–48, 1996.

Stefanie Wienkoop, Sacha Baginsky, and Wolfram Weckwerth. Arabidopsis thaliana as a model organism for plant proteome research. *Journal of proteomics*, 73(11):2239–2248, 2010.

Philip A Wigge, Min Chul Kim, Katja E Jaeger, Wolfgang Busch, Markus Schmid, Jan U Lohmann, and Detlef Weigel. Integration of spatial and temporal information during floral induction in arabidopsis. *Science*, 309(5737):1056–1059, 2005.

Amity M Wilczek, Judith L Roe, Mary C Knapp, Martha D Cooper, Cristina Lopez-Gallego, Laura J Martin, Christopher D Muir, Sheina Sim, Alexis Walker, Jillian Anderson, et al. Effects of genetic perturbation on seasonal life history plasticity. *Science*, 323(5916):930–934, 2009.

GDV Williams. Deriving a biophotothermal time scale for barley. *International Journal of Biometeorology*, 18:57–69, 1974.

Charles G Willis, Brad Ruhfel, Richard B Primack, Abraham J Miller-Rushing, and Charles C Davis. Phylogenetic patterns of species loss in thoreau's woods are driven by climate change. *Proceedings of the National Academy of Sciences*, 105(44):17029–17033, 2008.

Liming Xiong, Karen S Schumaker, and Jian-Kang Zhu. Cell signaling during cold, drought, and salt stress. *The plant cell*, 14(suppl_1):S165–S183, 2002.

Kazuko Yamaguchi-Shinozaki and Kazuo Shinozaki. Organization of cis-acting regulatory elements in osmotic-and cold-stress-responsive promoters. *Trends in plant science*, 10(2):88–94, 2005.

Kazuko Yamaguchi-Shinozaki and Kazuo Shinozaki. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu. Rev. Plant Biol.*, 57:781–803, 2006.

Yuzhen Ye and Adam Godzik. Comparative analysis of protein domain organization. *Genome research*, 14 (3):343–353, 2004.

Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8:1–14, 2007.

Seung Kwan Yoo, Kyung Sook Chung, Joonki Kim, Jeong Hwan Lee, Sung Myun Hong, Seong Jeon Yoo, So Yeon Yoo, Jong Seob Lee, and Ji Hoon Ahn. Constans activates suppressor of overexpression of constans 1 through flowering locus t to promote flowering in arabidopsis. *Plant physiology*, 139(2): 770–778, 2005.

Eve Ann Zachgo, Ming Li Wang, Julia Dewdney, David Bouchez, Christine Camilleri, Stephen Belmonte, Lu Huang, Maureen Dolan, and Howard M Goodman. A physical map of chromosome 2 of arabidopsis thaliana. *Genome Research*, 6(1):19–25, 1996.

Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

Hong Zhang, Fei Zhang, Yiming Yu, LI Feng, Jinbu Jia, BO Liu, Bosheng Li, Hongwei Guo, and Jixian Zhai. A comprehensive online database for exploring 20,000 public arabidopsis rna-seq libraries. *Molecular Plant*, 13(9):1231–1233, 2020.

Hua Zhang, Callista Ransom, Philip Ludwig, and Steven Van Nocker. Genetic analysis of early flowering mutants in arabidopsis defines a class of pleiotropic developmental regulator required for expression of the flowering-time switch flowering locus c. *Genetics*, 164(1):347–358, 2003.

James Z Zhang, Robert A Creelman, and Jian-Kang Zhu. From laboratory to field. using information from arabidopsis to engineer salt, cold, and drought tolerance in crops. *Plant physiology*, 135(2):615–621, 2004.

Tian-Qi Zhang, Zhou-Geng Xu, Guan-Dong Shang, and Jia-Wei Wang. A single-cell rna sequencing profiles the developmental landscape of arabidopsis root. *Molecular plant*, 12(5):648–660, 2019.

Zhiwei Zhao, Craig Dent, Huafeng Liang, Junqing Lv, Guandong Shang, Yawen Liu, Fan Feng, Fei Wang, Junhong Pang, Xu Li, et al. Cry2 interacts with cis1 to regulate thermosensory flowering via flm alternative splicing. *Nature Communications*, 13(1):1–15, 2022.

Appendix A
Details of Results

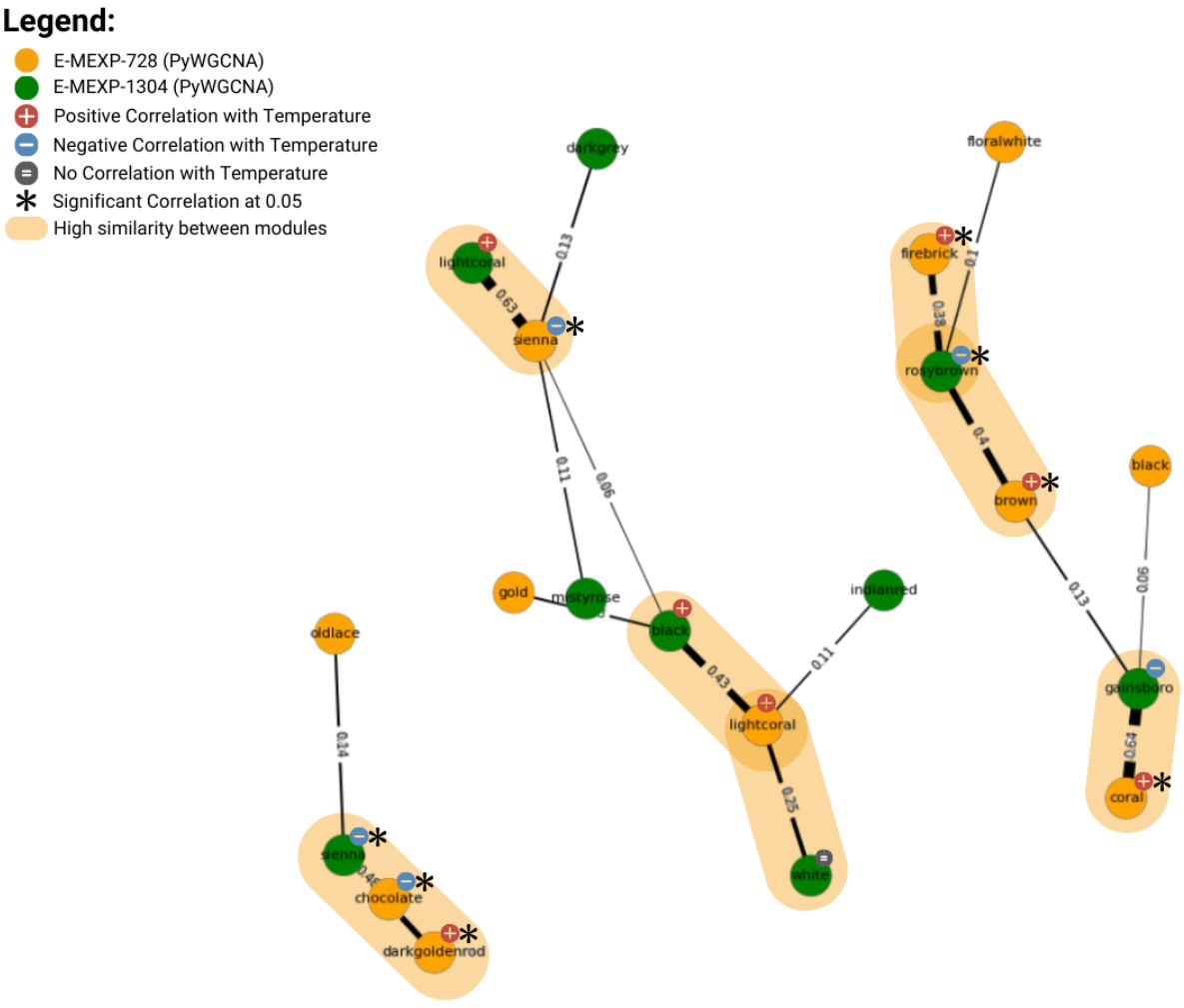


Figure A1: Network visualization of module similarity in the E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] datasets using Python-based Weighted Gene Correlation Network Analysis (PyWGCNA). Orange nodes represent modules extracted from E-MEXP-728, while green nodes represent modules extracted from E-MEXP-1304. Edge values represent the Jaccard similarity scores, indicating the degree of similarity between pairs of modules. The thickness of the edges is proportionate to the level of similarity, with thicker edges signifying stronger similarity. Modules marked with a yellow halo indicate a high degree of agreement between the pair (Jaccard value > 0.2). A red plus sign and a blue minus sign denote a positive and negative correlation with temperature, respectively, while a grey equal sign signifies no correlation with temperature. Modules exhibiting statistically significant relationships between temperature and module expression are marked with an asterisk.

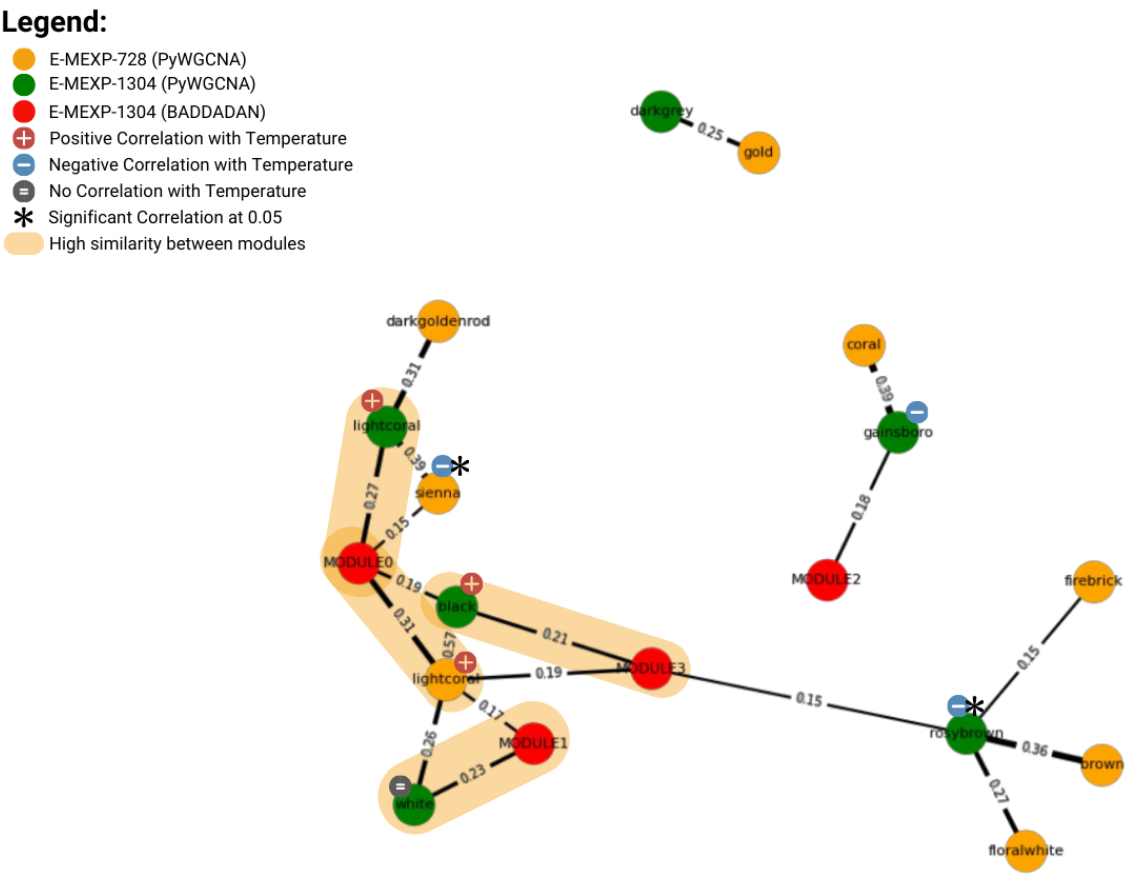


Figure A2: Network visualization of module similarity in the E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] datasets using Python-based Weighted Gene Correlation Network Analysis (PyWGCNA) and Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) methods. Orange nodes represent modules extracted from E-MEXP-728 with PyWGCNA, green nodes represent modules extracted from E-MEXP-1304 obtained with PyWGCNA, and red nodes represent modules extracted from E-MEXP-1304 using BADDADAN. Edge values denote the Jaccard similarity scores, reflecting similarity between the two pair of modules. The thickness of the edges is proportionate to the level of similarity, with thicker edges signifying stronger similarity. Modules marked with a yellow halo indicate a high degree of agreement between the pair (Jaccard value > 0.2). A red plus sign and a blue minus sign denote a positive and negative correlation with temperature, respectively, while a grey equal sign signifies no correlation with temperature. Modules exhibiting statistically significant relationships between temperature and module expression are marked with an asterisk.

Appendix B
Support work

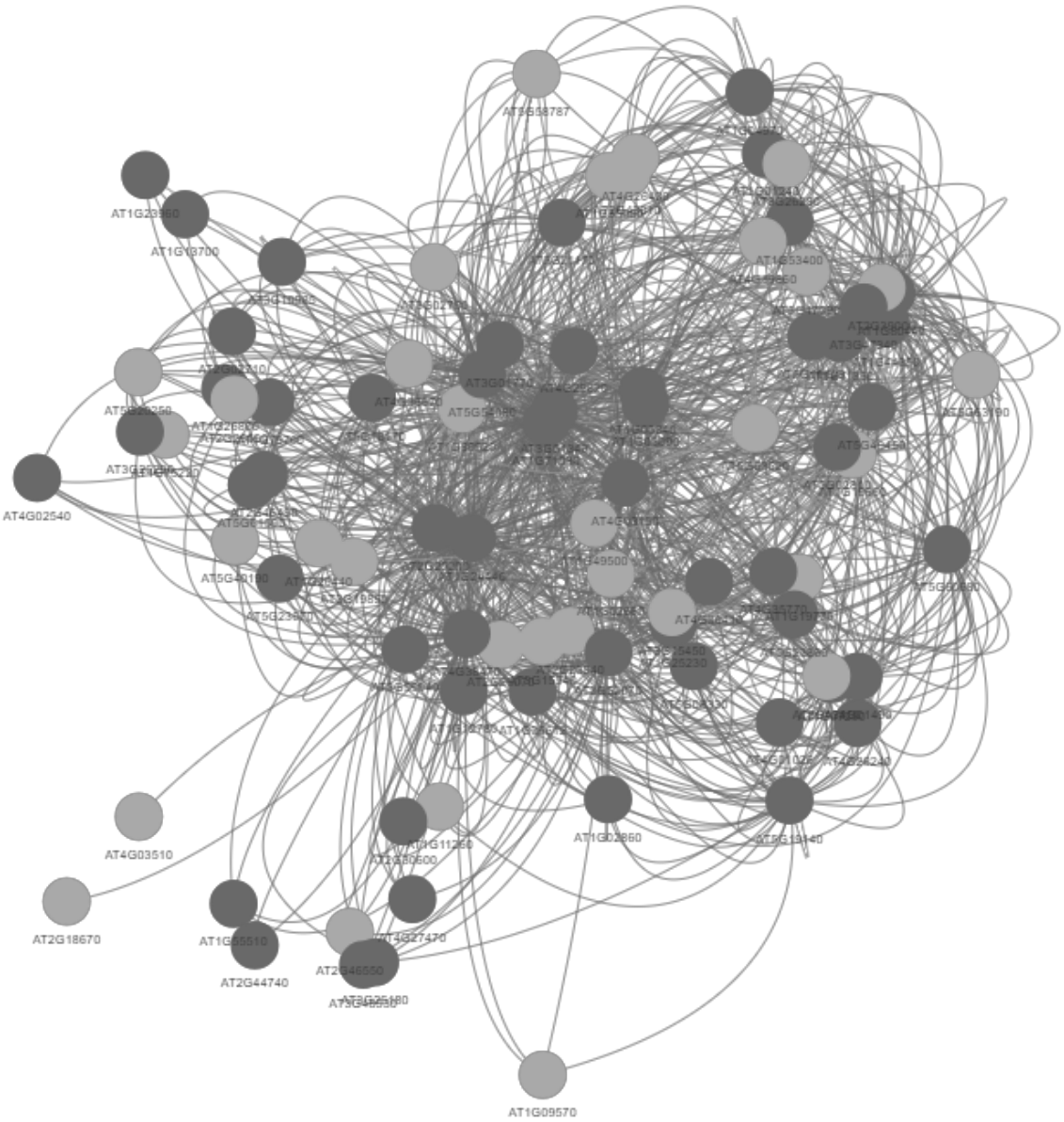


Figure B1: Network visualization for the extracted modules of the E-MEXP-728 Balasubramanian et al. [2006] dataset through the PyWGCNA approach. The visualization was filtered to only present 100 genes and 1000 connections given that the size of the complete network.

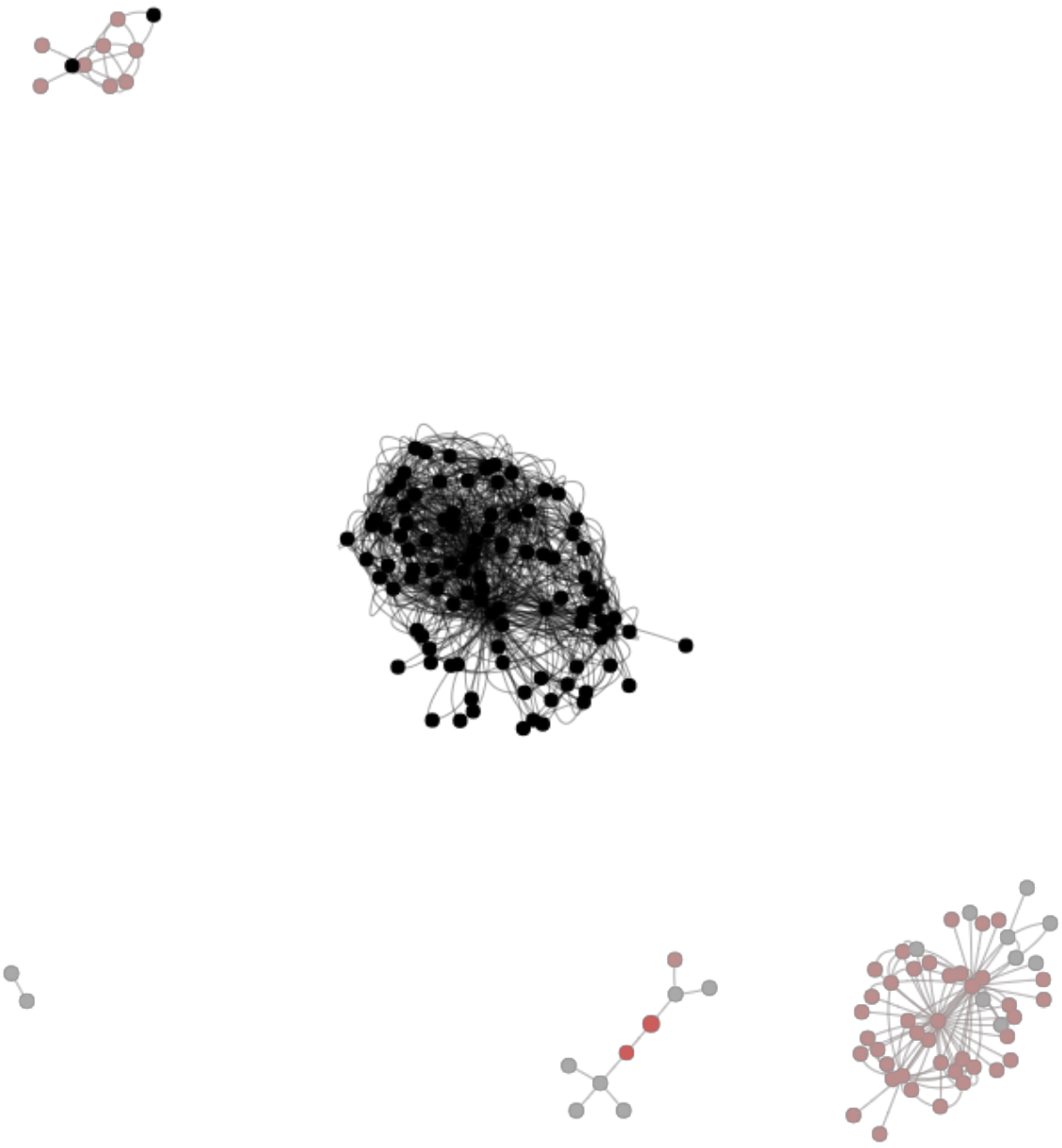


Figure B2: Network visualization for the extracted modules of the E-MEXP-1304 Michael et al. [2008] dataset through the PyWGCNA approach. The visualization was filtered to only present 100 genes and 1000 connections given that the size of the complete network.

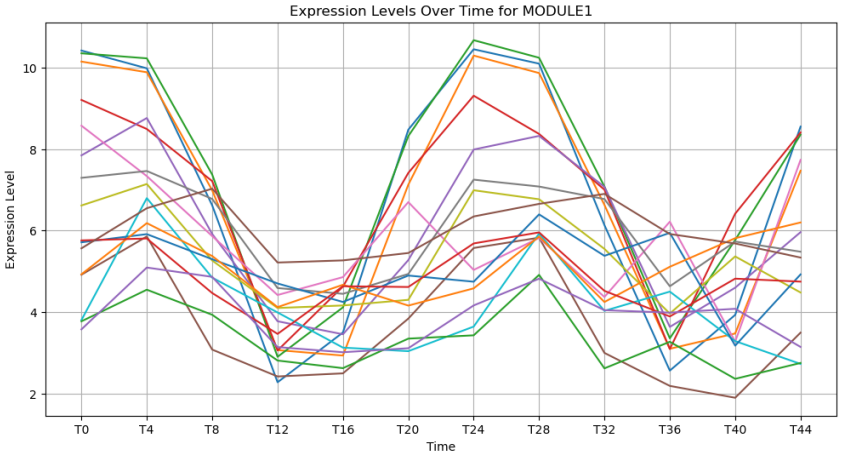


Figure B3: Plot representing the changes in expression levels over time for all 16 genes attributed to 'MODULE1' with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach. Each line and colour represents a unique gene.

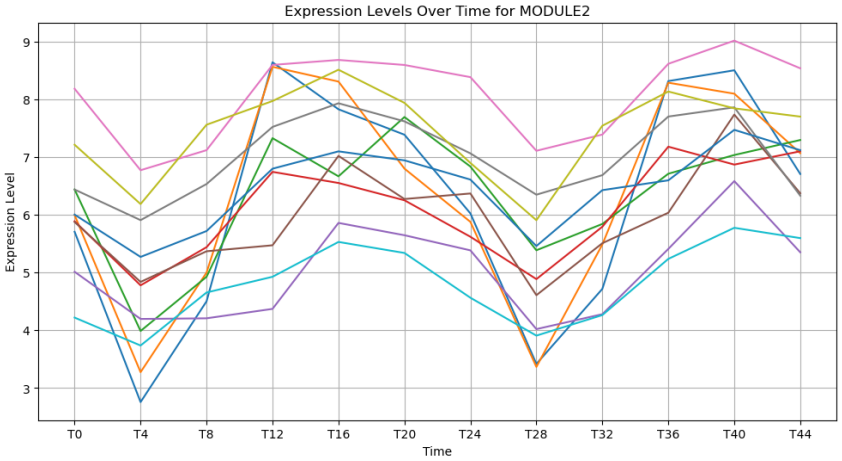


Figure B4: Plot representing the changes in expression levels over time for all 11 genes attributed to 'MODULE2' with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach. Each line and colour represents a unique gene.

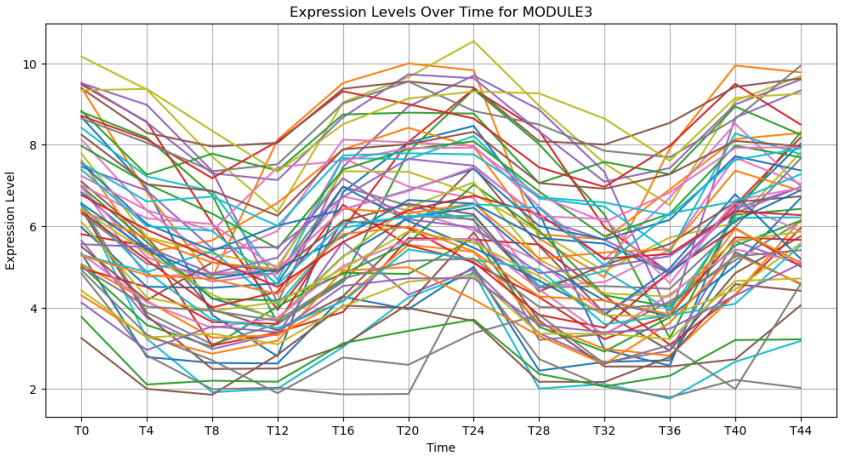


Figure B5: Plot representing the changes in expression levels over time for all 60 genes attributed to 'MODULE3' with the Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approach. Each line and colour represents a unique gene.

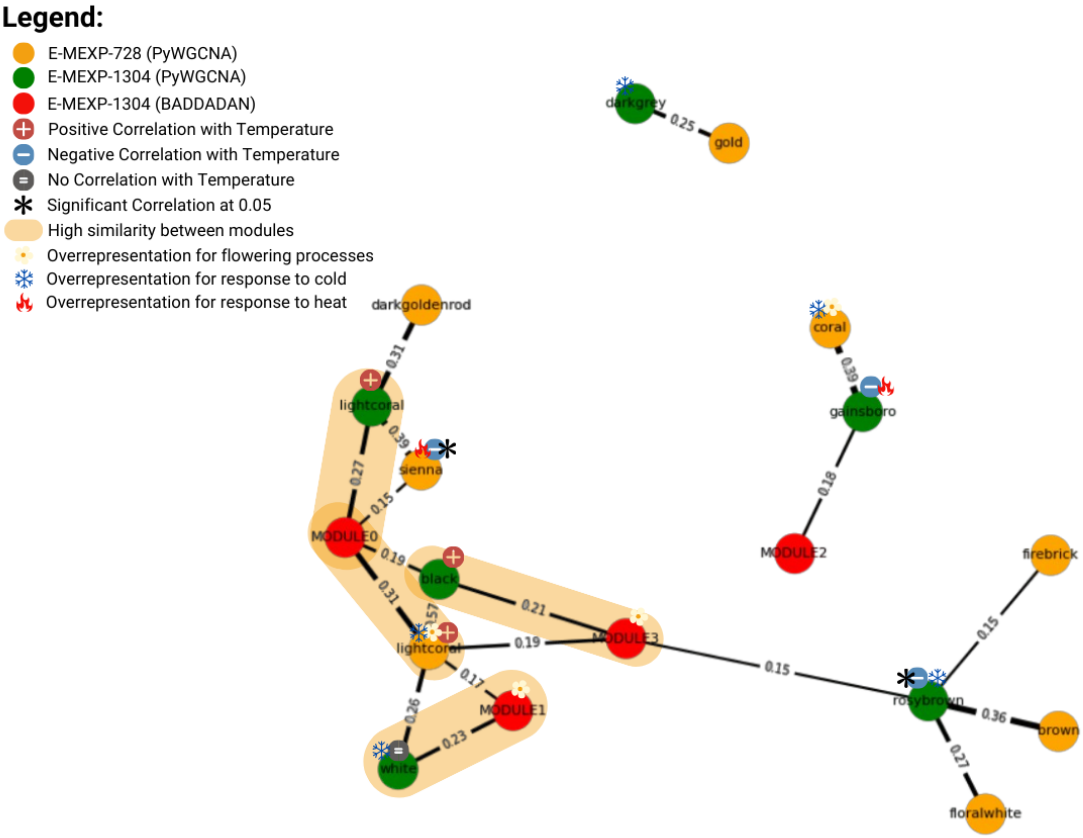


Figure B6: Network visualization illustrating module similarity and outcome representations for the modules extracted from the E-MEXP-728 Balasubramanian et al. [2006] and E-MEXP-1304 Michael et al. [2008] datasets using Python-based Weighted Gene Correlation Network Analysis (PyWGCNA) and Bioinformatics Approach to Describe Dynamical Activations of a Dimension-reduced Arabidopsis Network (BADDADAN) approaches. This visualization features symbolic representations denoting the results obtained from both the module-trait and functional enrichment analyses, specifically focusing on temperature and flowering response. Orange nodes represent modules extracted from E-MEXP-728 with PyWGCNA, green nodes represent modules extracted from E-MEXP-1304 obtained with PyWGCNA, and red nodes represent modules extracted from E-MEXP-1304 using BADDADAN. Edge values denote the Jaccard similarity scores, reflecting the similarity between the two pairs of modules. The thickness of the edges is proportionate to the level of similarity, with thicker edges signifying stronger similarity. Modules marked with a yellow halo indicate a high degree of agreement between the pair (Jaccard value > 0.2). A red plus sign and a blue minus sign denote a positive and negative correlation with temperature, respectively, while a grey equal sign signifies no correlation with temperature. Modules exhibiting statistically significant relationships between temperature and module expression are marked with an asterisk. A flower symbol signifies overrepresentation of flowering-related processes, an ice symbol represents overrepresentation in response to cold, and a flame symbol indicates overrepresentation in response to heat.

This project was developed in collaboration with the Bioinformatics group at Wageningen University & Research, in Wageningen, The Netherlands. This six-month mobility exchange was made possible with the financial support from the Erasmus+ traineeships funding.

With the support of the
Erasmus+ Programme
of the European Union

