UAB
Universitat Autònoma de Barcelona

Applications of Artificial Intelligence for predictive genomics and phenotyping

Federico Jurado Ruiz

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons: https://creativecommons.org/licenses/?lang=ca

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons: https://creativecommons.org/licenses/?lang=es

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license: https://creativecommons.org/licenses/?lang=en







Doctoral thesis

Universitat Autònoma de Barcelona
Facultad de Biociències
Departament de Biologia Animal, Biología Vegetal i Ecologia

Applications of Artificial Intelligence for predictive genomics and phenotyping

Dissertation presented by Federico Jurado Ruiz for the degree of Doctor in Plant Biology and Biotechnology.

Thesis director

Dr. Maria José Aranzana Civit

PhD Candidate

Federico Jurado Ruiz

Agradecimientos

Una tesis es algo que se consigue con gran esfuerzo propio, pero también con apoyo de muchas personas en el camino. Sin estas personas el camino sería extremadamente agotador y deprimente. Por ello quiero agradecer a las personas que han hecho de este camino un viaje único. Primero quiero agradecer a mis padres, Juan y Valentina, y a mi hermano Enrique que me han apoyado en toda mi vida y en concreto en este camino académico, siempre confiando y creyendo en mí. Doy las gracias a Veredas, mi prometida, sin cuyo apoyo, amor y paciencia no habría podido terminar este proyecto. Por supuesto, doy las gracias a mi mentora, directora y amiga Txosse por su atenta y paciente guía, siempre dispuesta a explorar temas diferentes y apasionantes que nos han llevado a proyectos innovadores. También agradezco a todos mis compañeros y amigos en el CRAG con quienes he compartido alegrías y también momentos difíciles, Nicole, Carles, Iván, Andrea, Silvia, Maroof, Azahara, Juan, Andrés, Simone y muchos más. También quiero darle las gracias al departamento, Ibo, Amparo, Bea y más, por el apoyo y por estar ahí siempre cuando tenía preguntas. Por último, quiero dar las gracias a mis colaboradores Mark, Joseph, Phuong y más por una experiencia magnífica y enriquecedora durante mi visita a Wageningen.

La investigación científica no es un camino sencillo y nos exige ser atrevidos con nuestras propuestas. Cuando nos permitimos soñar y planteamos preguntas extrañas y divertidas abrimos la puerta a ideas que siembran la semilla del cambio.

Acknowledgements

A thesis is something that is achieved with great personal effort, but also with the support of many people along the way. Without these people, the journey would be extremely exhausting and depressing. Therefore, I want to thank the people who have made this journey a unique experience. First, I want to thank my parents, Juan and Valentina, and my brother Enrique, who have supported me throughout my life and specifically on this academic path, always trusting and believing in me. I thank Veredas, my fiancée, whose support, love, and patience made it possible for me to complete this project. Of course, I thank my mentor, supervisor, and friend Txosse for her attentive and patient guidance, always willing to explore different and exciting topics that have led us to innovative projects. I also thank all my colleagues and friends at CRAG with whom I have shared joys and difficult moments, Nicole, Carles, Ivan, Andrea, Silvia, Maroof, Azahara, Juan, Andrés, Simone, and many more. I also want to thank the department, Ibo, Amparo, Bea and more, for their support and for always being there when I had questions. Finally, I want to thank my collaborators Mark, Joseph, Phuong, and others for a magnificent and enriching experience during my visit to Wageningen.

Scientific research is not an easy path and requires us to be bold with our proposals. When we allow ourselves to dream and pose strange and fun questions, we open the door to ideas that plant the seed of change.

Table of Contents

Agradecimientos

Acknowledgements

List of abbreviations

Summary 1

Resumen 2

Resum 3

General introduction 5

Agriculture and breeding through time 7

Sequencing 10

First generation sequencing 10

Second generation sequencing 12

Third generation sequencing 13

Sequencing for breeding. 16

Associating genetic information and traits 18

Neural networks basics and uses 20

The basic unit 20

Dense layers and multilayer architecture 22

Convolutional layers 23

Detection and segmentation 24

Autoencoders 26

How do they learn? Training a network 27

Current applications of artificial intelligence to genomics and phenotyping 29

Applications on phenomics 30

Applications to genomics 32

Smart Breeding 34

Objectives 37

Chapter 1: Shape Analyzer: An open-source application for fruit morphology phenotyping 41

1.1. Abstract 43

1.2. Introduction 44

1.3. Materials and Methods. 47

1.3.1. Image Datasets. 47

1.3.2. Detection and segmentation. 48

1.3.3. Metrics given by shape analyzer 48

1.3.4. Development of the desktop application. 51

1.3.5. Detection and segmentation evaluation. 53

Validation of Shape Analyzer for genomic studies 55

1.3.6. Plant material, DNA extraction, SNP genotyping and Fruit phenotyping. 55

1.3.7. Heritability, Linkage map construction and QTL mapping. 56

1.4. Results 59

1.4.1. Image detection and segmentation accuracy 59

1.4.2. Shape analyzer application on the QTL mapping of apple shape and size traits. 60

1.5. Discussion 67

1.5.1. Boosting open software and Easy-to-Use tools for the morphometrics community 67

1.5.2. Mask R-CNN performance and adequacy. 68

1.5.3. Evaluating shape analyzer measures adequacy. 69

1.6. Conclusion 70

1.7. Code availability 71

Chapter 2: LeTra: A leaf tracking workflow based on convolutional neural networks and intersection over union. 73

2.1. Abstract 75

2.2. Introduction 76

2.3. Results 79

2.3.1. General Approach 79

2.3.2. Leaf detection and segmentation 80

2.3.3. Leaf Tracking 82

2.3.4. Leaf count as indicator of development progression 83

2.3.5. Photosynthetic capacity at plant level 84

2.3.6. Photosynthetic capacity at leaf level 85

2.3.7. Photosynthetic response is influenced by leaf age and order 86

2.4. Discussion 87

2.5. Conclusion 92
2.6. Methods 93
2.6.1. Plant material and growing conditions 93
2.6.2. Phenotyping and data collection 94
2.6.3. Training and Validation datasets for detection and segmentation, and tracking evaluation dataset. 95
2.6.4. Detection and segmentation 96
2.6.5. Instance tracking 97
2.6.6. Leaf counting and photosynthesis capacity analysis of two different genotypes 98
2.6.7. Metrics definition 99
2.7. Availability of data and materials 101
Chapter 3: GenoDrawing: An autoencoder framework for image prediction from SNP markers 113
3.1. Abstract 115
3.2. Introduction 116
3.3. Materials and Methods 119
3.3.1. General Approach 119
3.3.2. Plant materials, image acquisition and processing 121
3.3.3. Genotypic data 121
3.3.4. Neural networks models: autoencoder and embedding predictor 122
3.3.5. Embedding predictor target dataset generation 123

3.3.6. Computational specifications 124

3.4. Results 124

3.4.1. Resulting Framework 124

3.4.2. Autoencoder and embedding predictor training. 125

3.4.3. Comparing the impact of trait-related and randomly selected SNPs on embedding prediction models for GenoDrawing 127

3.4.4. Effect of Adding Non-Shape-Related Markers on The Embedding Predictor Model Accuracy. 131

3.5. Discussion 133

3.5.1. Evaluation of Autoencoder Performance, Loss Function Selection, and Embedding Predictor Structure. 133

3.5.2. Using the mean embedding as a representation of the phenotype 135

3.5.3. Relevance of SNP selection in genome prediction models. 136

3.5.4. Evaluating embedding predictor performance using FSI and SR metrics in GenoDrawing. 137

3.5.5. Limitations 139

3.5.6. Data Availability: 139

Chapter 4: Generating images from SNP data through artificial intelligence 141

4.1. Abstract 143

4.2. Introduction 144

4.3. Methods 147
4.3.1. Dataset 147
4.3.2. Neural networks architecture and training 147
Vector quantization for images. 147
Vector quantization for SNPs. 150
Connection through a transformer. 152
4.3.3. Evaluation of the predicted images. 152
4.3.4. Evaluation of SNP reconstruction accuracy 153
4.3.5. Code and libraries used. 153
4.4. Results 154
4.4.1. Image reconstruction 154
4.4.2. SNP reconstruction 155
4.4.3. Connection through transformer 157
4.5. Discussion 160
4.5.1. Quality of the quantized images. 160
4.5.2. SNP vector quantizer effectiveness. 160
4.5.3. Predictions evaluation 161
4.6. Conclusions 162
General Discussion. 165
Synthesis of the findings and highlights 167
Current state, future prospectives and personal views 169
Conclusions 173

Supplementary Materials.....177
Figures for Chapter 1:.....177
Figures for Chapter 2:.....178
Figures for Chapter 4:.....186
Tables for Chapter 4:.....187
References.....198
Annex 219

List of abbreviations

ADAM - Adaptive momentum estimation.

AE – AutoEncoder network.

AI - Artificial intelligence.

AN – Artificial neuron.

ANN – Artificial neural network.

Bp – Base pair.

BC – Back cross.

CF - chlorophyll fluorescence.

cGANs - Cycle GANs.

CNN – Convolutional neural network.

ddNTPs - di-deoxynucleotide triphosphates.

DNA - Deoxyribonucleic acid.

DL – Deep learning.

ELIDA - Enzymatic luminometric inorganic pyrophosphate detection assay.

ENBED - Ensemble Nucleotide Byte-level Encoder-Decoder.

FCN - Fully Convolutional Network.

FP – False positives.

FN – False negatives.

F_v/F_m - Maximum quantum yield.

F_0 - Minimum value of CF.

F_m - Maximum level of CF.

FSI – Fruit shape index.

GANs - Generative adversarial networks.

GBS - Genotype-by-sequencing.

GPT – Generative pre-trained transformers.

GS – Genomic selection.

GWAS - Genome-Wide Association Studies.

GxExM - Genotype x environment x management.

HOTA - Higher Order Tracking Accuracy.

HTP - High-throughput phenotyping.

InDels- Insertions and/or Deletions.

IoT - Internet of things.

IoU - Intersection over union.

LLM – Large language model.

MAE - Mean absolute error.

MAS – Marker assisted selection.

MAB – Marker assisted breeding.

ML - Machine learning.

MLM – Mixed linear models.

MLP - multilayer perceptron.

MSE - Mean Squared Error.

NGS – Next generation sequencing.

NLP - Natural language processing.

NN – Neural network, in this thesis synonym to ANN.

NPEC - Netherlands Plant Eco-phenotyping Center.

ONT – Oxford Nanopore Technologies.

PacBio – Pacific Biosciences.

PCR - polymerase chain reaction.

PPi - inorganic pyrophosphate.

PSI - Photon Systems Instruments.

PX – Pixel.

QTL - Quantitative Trait Loci.

ReLU - Rectified Linear Unit.

RPN - Region proposal network.

rSNPs – Randomly selected SNPs.

SB - Smart Breeding.

SGD - Stochastic gradient descent.

SNP - Single-nucleotide polymorphisms.

SR – Shoulder ratio.

SSD - Single Shot Multibox Detector.

TLU - Threshold logic unit.

TP – True positives.

tRNA – Transfer RNA.

tSNPs – Trait related SNPs.

UAV - Unmanned aerial vehicles.

VGG – Visual Geometry Group.

ViT – Vision Transformer.

VQ – Vector Quantization.

WGS – Whole genome sequencing.

WGR – Whole genome re-sequencing.

YOLO - You Only Look Once.

ZMWs - Zero-mode waveguides.

ϕPSII - Efficiency of photosystem II.

μ - mean.

σ^2 – standard deviation.

Summary

This thesis proposes different ways to apply neural network-based models to a range of problems in plant genetics research. The focus is placed on three different tasks, and solutions are proposed for each. The first task is measurement, for which an open-source tool utilizing an instance segmentation model has been developed. This model has been trained for three fruits: apple, pear, and almond, with specific measures implemented for each to facilitate shape studies. To ease access and use by the scientific community, a minimalist interface has also been developed. The second task is time series measurement. For this, an instance segmentation model has been used and coupled with a new association method. This method has enabled a pioneering study of differences in photosynthetic performance between two genotypes of *Arabidopsis thaliana* at the leaf level. This study shows how leaf-level photosynthetic performance patterns do not follow the average pattern in one genotype, while they do in the other. The final task is phenotype prediction. In this case, the applicability of generative artificial intelligence to phenotype prediction from genetic markers has been investigated. Predicting phenotypes related to shape and appearance is particularly complex since the selection of relevant features can be subjective. Therefore, this thesis demonstrates a method for generating phenotype images from genetic markers (GenoDrawing) and possible ways to advance this new line of research.

Resumen

En esta tesis se proponen diferentes formas de aplicar los modelos basados en redes neuronales a diferentes problemas dentro de la investigación en genética vegetal. Para ello se sitúa el foco en tres tareas diferentes y se proponen soluciones para cada una. La primera tarea es la medición y para abordarla se ha desarrollado una herramienta de código abierto que utiliza un modelo de segmentación de instancias. Este modelo se ha entrenado para tres frutos, manzana, pera y almendra y se han implementado medidas específicas para cada uno a fin de facilitar los estudios de forma. Para facilitar el acceso y uso por parte de la comunidad científica se ha desarrollado también una interfaz minimalista. La segunda tarea es la medición en series temporales. Para ello se ha utilizado un modelo de segmentación de instancias y se ha acoplado junto con un nuevo método de asociación. Este método ha permitido un estudio pionero de las diferencias en rendimiento fotosintético entre dos genotipos de A*rabidopsis thaliana* a nivel de hojas. En este estudio se muestra como los patrones de rendimiento fotosintético a nivel de hoja no siguen el patrón medio en un genotipo mientras que en otro si lo hacen. La última tarea es la predicción de fenotipos. En este caso se ha investigado la aplicabilidad de la inteligencia artificial generativa a la predicción del fenotipo desde marcadores genéticos. La predicción de fenotipos relacionados con la forma y apariencia resulta especialmente compleja dado que la selección de características relevantes puede resultar subjetiva. Por tanto, en esta tesis se demuestra un método de generación de imágenes del fenotipo desde marcadores genéticos (GenoDrawing), así como posibles formas de avanzar esta nueva línea de investigación.

Resum

En aquesta tesi es proposen diferents formes d'aplicar els models basats en xarxes neuronals a problemes dins de la investigació en genètica vegetal. Per això, se situa el focus en tres tasques diferents i es proposen solucions per a cadascuna. La primera tasca és la mesura i per abordar-la s'ha desenvolupat una eina de codi obert que utilitza un model de segmentació d'instàncies. Aquest model s'ha entrenat per a tres fruits: poma, pera i ametlla, i s'han implementat mesures específiques per a cadascun a fi de facilitar els estudis de forma. Per facilitar l'accés i l'ús per part de la comunitat científica s'ha desenvolupat també una interfície minimalista. La segona tasca és la mesura en sèries temporals. Per això, s'ha utilitzat un model de segmentació d'instàncies i s'ha acoblat juntament amb un nou mètode d'associació. Aquest mètode ha permès un estudi pioner de les diferències en rendiment fotosintètic entre dos genotips d'Arabidopsis thaliana a nivell de fulla. En aquest estudi es mostra com els patrons de rendiment fotosintètic a nivell de fulla no segueixen el patró mitjà en un genotip mentre que en l'altre sí que ho fan. L'última tasca és la predicció de fenotips. En aquest cas, s'ha investigat l'aplicabilitat de la intel·ligència artificial generativa a la predicció del fenotip des de marcadors genètics. La predicció de fenotips relacionats amb la forma i aparença resulta especialment complexa, ja que la selecció de característiques rellevants pot resultar subjectiva. Per tant, en aquesta tesi es demostra un mètode de generació d'imatges del fenotip des de marcadors genètics (GenoDrawing), així com possibles formes d'avançar en aquesta nova línia de recerca.

General introduction

Agriculture and breeding through time

Agriculture and plant breeding have moved far from their traditional roots. In the past, it was mostly based on what farmers and growers had learned through experience and observation. We have moved from phenotype-based selection methods to advanced techniques which includes genetic information to help us grow better crops. This section examines how farming and breeding started with these traditional methods and has grown into the science it is now.

The origins of agriculture are subject to extensive debate and research. It is unclear which were the motivators for the transition from a nomadic hunter-gatherer society to a settled society focused on agriculture (Piperno, 2018). It is more commonly agreed that agriculture started approximately 12,000 years ago as a complex gradual process in which humans and plants played a mutual influence game (Allaby et al., 2022), forming what some consider the first economic revolution. The **Neolithic** agricultural revolution impacted every aspect of human society with dietary changes that affected our physical development and life expectancy and enabled the development of an economic system. Inventors could now obtain returns from their properties, which, in exchange boosted the development of tools that assisted in different agricultural processes (Bowles & Choi, 2019). Animal husbandry, which might have started before agriculture (Scanes, 2018), played a crucial role as the principal energy source for these processes, being used for plowing, mills, and transport. In parallel, the process of crop domestication began (Peng et al., 2011) with advances largely driven by the intuition and

experience of early humans. Gardeners started selecting seeds and crops not only for their high yield but, importantly during that epoch, for their high seed-to-yield ratio (Thomas Stalke et al., 2021). This approach ensured that humans had not only seeds for the next season but also nutrition and was the beginning of a guided genetic selection process. Passed through generations, it was the cause of the emergence of many modern cultivars and laid the foundational practices of traditional breeding. It was a slow and gradual process, but essential in laying the groundwork for understanding how to manipulate and improve plant traits (Herrera & Garcia-Bertrand, 2018). From the **Neolithic** era to the **Middle Ages** (spanned from the 5th century to the 15th century), the primary focus of agricultural advancements laid on the more efficient utilization of land, the development of sophisticated irrigation systems, and the introduction of new cultivars, which were essential for diversifying and securing food sources (van der Veen, 2010). This period witnessed the transformation of primitive farming techniques into more structured and sustainable practices. Terracing, crop rotation, and the use of natural fertilizers were among the innovations that improved yields and soil fertility. The introduction of new crops and livestock resulting from trade and conquests also played a pivotal role in shaping regional agriculture. However, it was not until the **Age of Enlightenment** (late 17th and 18th centuries) that humanity embarked on a revolutionary journey in both agriculture and science. This era brought forth a scientific approach to farming, moving beyond empirical knowledge to a more systematic understanding of botany, soil science, and genetics. The Enlightenment saw the beginnings of agronomy as a formal science and the application of conscient scientific principles to

farming, leading to significant improvements in crop production and animal husbandry. Innovations such as crop experimentation, selective breeding, and the early stages of mechanization marked this period as a turning point in agricultural history (Jones, 2016), setting the stage for the modern agricultural practices we know today. In this context, Thomas Fairchild, an English gardener, created the first intentional flower hybrid around 1720 (Leapman, 2012) while he was trying to elucidate the existence of sex in plants. His experimental "mule" flower, though not widely accepted at the time, was a pioneering step in the field of hybridization. During the **Industrial Revolution**, which spanned from the late 18th to the mid-19th century, significant changes in agriculture paralleled those in industry. This period saw the advent of mechanized farming, dramatically changing agricultural practices. The introduction of machinery like the seed drill, mechanical reapers, and threshing machines increased efficiency and productivity, reducing the reliance on manual labor. This era also fostered improved methods in agricultural management and land use, further boosting food production to meet the demands of a rapidly growing population. It was in the midst of this transformative period, in 1865, that Gregor Mendel, an Augustinian monk, made a groundbreaking discovery. He observed that traits pass down through generations in pea plants, leading him to formulate the laws of inheritance (Mendel, 1865). Mendel's work was not fully recognized until 40 years later and laid the foundation for the field of genetics, revolutionizing our understanding of heredity and breeding practices. Building on Mendel's legacy, the early 20th century witnessed further developments in genetics, which profoundly influenced agricultural science. The advent of genetics introduced

pivotal concepts, further expanding the understanding of heredity and breeding. One of these concepts was Quantitative Traits (QT), mentioned as early as 1923 (Sax, 1923), which tried to measure how much genes can influence traits. The increasing understanding of DNA functionality and DNA structure in 1953 (Franklin & Gosling, 1953; Watson & Crick, 1953; Wilkins et al., 1953) emerged new possibilities. This new knowledge later would enable breeders to select and develop plants based on their genetic potential rather than just observable traits. Subsequently, around 1955, we can find references to the idea of genes as markers (Everson & Schaller, 1955) but it was not until the development of DNA sequencing when we can start talking about Marker-Assisted Selection (MAS) and the current state of breeding.

Sequencing

First generation sequencing

Sequencing is arguably one of the biggest breakthroughs in genetics history. Twelve years after the discovery of the DNA structure, Robert Holley and his colleagues, whose work focused on the study of tRNA and alanine tRNA, were able to decipher the first RNA sequence. Their approach consisted of the digestion of the alanine tRNA, using pancreatic ribonuclease, into seventeen fragments with composition determined in previous works through chromatography (Holley, Everett, et al., 1965). By determining which were the terminal and starting fragments and then studying the fragments given, they were able to reconstruct the first sequence of 77 base pairs (bp) (Holley,

Apgar, et al., 1965). At the same time, Frederick Sanger published multiple works from the 1965 to 1975 which would culminate with a method (Sanger & Coulson, 1975) that allowed to obtain the first complete genome, belonging to the bacteriophage ϕX174 (Sanger, Air, et al., 1977). Not much later, the still used Sanger sequencing method, was published (Sanger, Nicklen, et al., 1977). This method is based on the usage of modified di-deoxynucleotide triphosphates (ddNTPs) which are fluorescent. Unlike their normal counterparts, ddNTPs lack the 3' hydroxyl group necessary for forming the phosphodiester bond with the next nucleotide, thus halting DNA synthesis once incorporated. In its initial form, this method utilized radioactive labels to identify the end points of the synthesized DNA fragments, with each of the four ddNTP types added to separate reaction mixtures. The terminated fragments were then separated by gel electrophoresis and visualized on an X-ray film through autoradiography, allowing the sequence to be read directly from the pattern of bands. With the advent of fluorescence-based detection in the 80s, the Sanger method evolved significantly. Fluorescent labels, each emitting at a different spectrum for the four ddNTPs, replaced radioactive tagging. This innovation facilitated the automation of sequencing and enabled the simultaneous processing of all four ddNTP reactions in a single run, using capillary electrophoresis and laser detection systems. Sanger sequencing might be the oldest method for sequencing, but it is still heavily used under multiple circumstances (Bloemen et al., 2022; Wen et al., 2020) as it is the most precise sequencing method for short sequences. This method, however, is constrained by the length of the sequence, around 800-1000bp (Crossley et al., 2020), and the time require to produce them. This limitations did not stop the scientific

community from achieving the first full human genome in 2006 with a total cost of approximately 2.7 billion dollars and 13 years of work (Wetterstrand, 2021).

Second generation sequencing

As the field of genomics pressed forward, the need for faster, more cost-effective sequencing methods led to the emergence of second-generation sequencing also known as next generation sequencing (NGS) technologies. This new era was characterized by a departure from the labor-intensive and time-consuming methods of Sanger sequencing to massively parallel processing systems. These innovative platforms allowed for the sequencing of millions of small DNA fragments simultaneously, drastically reducing the cost and time required for sequencing the entire genome. The first technology of this kind was the pyrosequencing in 1996 (Ronaghi et al., 1996) by 454, later acquired by Roche. This technology is based on the use of emulsion PCR and the detection of inorganic pyrophosphate (PPi) released during DNA synthesis. The key to this method is that the amplification of DNA strands allows for the emission of light signals. These signals are produced when a DNA base is correctly added to a growing strand, leading to the release of PPi, which is then converted into a detectable light signal through a series of enzymatic reactions. This process enables the identification of each base added during the sequencing using the enzymatic luminometric inorganic pyrophosphate detection assay (ELIDA) (Nyrén et al., 1993).

Despite the groundbreaking nature of 454 pyrosequencing, its application has waned over time, overshadowed by the advancements

brought by Solexa, which was acquired by Illumina in 2007. Solexa's method, now a cornerstone of Illumina's sequencing capabilities, revolutionized NGS with its innovative approach to sequencing. This technique involves attaching adapters to both ends of DNA strands, facilitating the formation of amplification bridges on a solid surface. Crucially, the incorporation of each nucleotide triggers the emission of a fluorescence signal, detectable across multiple laser channels. Following nucleotide addition, the 3' end of the DNA chain is chemically cleaved, converting the amplification bridge back to a single stranded state, still anchored to the surface, and primed for subsequent amplification cycles (Bentley et al., 2008). Illumina's technology would culminate into their sequencing platforms that stands out for its scalability and cost-effectiveness. This has made Illumina's sequencing technology the most broadly adopted across various fields of research.

The impact of second-generation sequencing extends beyond technical innovations. It has democratized genomic research, making it possible for smaller labs to engage in high-throughput sequencing projects that were once the exclusive domain of large institutions. Moreover, the flood of genomic data generated by NGS has spurred developments in bioinformatics, challenging researchers to devise new algorithms and computational tools to analyze, store, and interpret the vast datasets produced.

Third generation sequencing

As transformative as NGS has been, the field continues to evolve, with ongoing efforts to overcome the limitations of short-read sequencing.

The search for methods that produce longer reads, higher throughput, and even lower costs drives the development of third-generation sequencing technologies. In this field there are two companies leading the field, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). PacBio made a significant breakthrough in 2009 with the introduction of their third-generation sequencing technology (Eid et al., 2009). Their approach centers around the use of zero-mode waveguides (ZMWs), small nano wells that confine individual DNA polymerase molecules and their template DNA strands. During the sequencing process, each nucleotide incorporated by the polymerase is labeled with a unique fluorescent marker. As the DNA strand is synthesized, these fluorescent markers are excited and emit light signals corresponding to each added nucleotide. The fluorescent markers is also clipped by the polymerase during the addition, ending the pulse signal. This real-time observation of DNA synthesis allows for the sequencing of significantly longer reads compared to previous technologies, with the added benefit of providing a direct view of the DNA replication process.

A few years later, in 2014, ONT would present the MinION a technology that would surprise for its portable size and capabilities. This technology is based on the use of nanopores introduced in the early 90s, where individual DNA strands are threaded through nanoscopic pores embedded in a synthetic membrane. As the DNA transits through the nanopore, it causes characteristic disruptions in an ionic current flowing through the pore. These disruptions are then analyzed to determine the sequence of nucleotides in the DNA strand (Jain et al., 2016). ONT's method stands out for its ability to produce

ultra-long reads, making it particularly useful for sequencing complex genomic regions and for applications where real-time sequencing data is advantageous. However, ONT sequencing tends to have a higher error rate compared to both NGS and PacBio technologies, particularly in homopolymer regions, and the read accuracy can vary based on the length and complexity of the DNA sequence (Delahaye & Nicolas, 2021).

Recent advances in sequencing technology continue to address these limitations. Efforts are being made to improve accuracy and reduce the costs of both PacBio and ONT technologies. Enhanced error-correction algorithms and developments in sequencing chemistry and hardware are being implemented to increase the reliability and efficiency of long-read sequencing (Satam et al., 2023).

Sequencing for breeding.

As mentioned in the first section, the field of breeding has undergone a remarkable evolutionary process since its inception, which can be broadly categorized into four distinct eras (S. Jiang et al., 2020). **Breeding 1.0** encompasses the early domestication period and initial forays into trait selection based on observable characteristics. **Breeding 2.0** began with the discovery of genetics, using Mendelian inheritance for more systematic trait selection, but still with limited knowledge in the molecular side. The advances in sequencing, specifically NGS, kick-started the third generation of breeding. Although the Sanger method provides a reliable sequencing approach, its lower throughput limits its utility in the context of large-scale breeding programs. The breakthrough came with the development of high-throughput sequencing technologies that have enabled whole-genome sequencing (WGS). The new sequencing techniques allowed for drafting full genomes in many unknown organisms, paving the way for whole genome re-sequencing (WGR) (Torkamaneh et al., 2018). WGR is particularly relevant in breeding as it facilitates the identification of small variations in the DNA sequence that may or may not entail functional differences, but crucially, they serve as genetic markers for characterization. This small variations are commonly either insertions or deletions of less than 50 bases, known as InDels, or simply specific changes in the bases, known as single-nucleotide polymorphisms (SNPs)(Varshney et al., 2020). The ability to detect these modifications has far-reaching implications in breeding. It enables the creation of detailed genetic fingerprints for each individual, providing a high-resolution view of the genetic diversity

within a population. This granularity is essential for identifying genetic variations correlated with desirable traits, supporting more informed selection decisions in breeding programs. With the introduction of genotype-by-sequencing (GBS) techniques, the scope of breeding research has expanded even further. GBS streamlines the process of genotyping, making it more efficient and cost-effective (D'Agostino & Tripodi, 2017). This approach involves sequencing random DNA fragments from the genome and then identifying SNPs within these fragments. GBS is particularly valuable in breeding as it allows for the rapid screening of large populations, facilitating the discovery of genotype-phenotype correlations. The last generation of breeding will be covered in section 1.5.

Associating genetic information and traits

Marker-assisted breeding has been revolutionized by the increased accessibility and affordability of SNP chips, altering traditional breeding methodologies. These chips enable the detailed characterization of populations, facilitating the establishment of statistical associations with desired traits. Among the various methodologies employed to build associations, QTL (Quantitative Trait Loci) mapping and GWAS (Genome-Wide Association Studies) stand out as the two predominant approaches and which to use strongly depends on the experimental design. QTL mapping is primarily utilized to pinpoint regions within the genome (QTL) that influence a certain trait. The predominant experimental designs are F2 biparental populations which can derive from (1) F1 selfing hybrids or (2) BC events of an F1 hybrid x of the parental lines. On these kinds of experimental designs, methods such as interval mapping (Jansen, 1993) can be employed to identify genetic markers closely linked to desirable traits (Iwata et al., 2016). Although QTL mapping is still used, the current trend in research is GWAS. One of the GWAS advantages is the flexibility in the experimental design. A GWAS can be performed in large populations mixing different origins which allows for a finer mapping of the markers providing the information. There are many different methods used to perform a GWAS but the predominant trends is the mixed linear models (MLM) and their different variants and improvements. In short, MLM is a kind of regression which accounts for both fixed effects (genetic variants) and random effects (population structure and relatedness) to improve accuracy and reliability. The chapter published by Gupta (Gupta et al.,

2019) contains an excellent collection of the different methods from to original paper using mixed linear models (Yu et al., 2006) to one of the most used today, FarmCPU (X. Liu et al., 2016).

Most of the methods based on MLM are powerful but also present a few shortcomings. Some of the issues found using these methods are their reliance on assumptions (such as normal distributed traits), the challenges inherent to epistasis, the detection of non-linear traits and the often appearance of significant SNPs not related to the causal SNP (Sun et al., 2021). One of the current topics of research is the application of machine learning (ML) techniques to improve the predictive capabilities of statistical modelling in a GWAS, or to completely replace it. Machine learning comprehends a wide range of techniques that are based on learning algorithms which adjust themselves using experience. Some of the most known methods in ML are ensemble models and neural networks. Neural networks will be explained in detail in the next section as it has been the main methodology used in this thesis. Ensemble methods are based on pooling mechanisms, which depending on the approach, can be compound hundreds of decision trees from which a consensus is extracted (Dong et al., 2020). ML as a post-GWAS approach is showing to be extremely useful to filter false positives and find the true relevant SNPs refining the results of a GWAS (Nicholls et al., 2020). Ensemble methods have one of the biggest advantages on this task as decision trees can be explored and the SNPs, along with their relevance, can be obtained from the predictive model. On the other hand, neural networks (NN) behave as a black box, which often makes the results very difficult to understand (I. E. Nielsen et al., 2022).

Neural networks basics and uses

In this thesis there will be several mentions to different neural networks architectures and models. This section provides all the basic ideas on the different layers and architectures used in this thesis.

The basic unit

The unprecedented popularity of ChatGPT has significantly elevated the status of neural networks and artificial intelligence but the basics of modern NN date from 1957 with a laboratory report which named the perceptron (Rosenblatt, 1957, 1958). Rosenblatt idea was to emulate how a biological neuron works, using multiple inputs like dendrites and giving an output signal as the axon, creating what we name a threshold logic unit (TLU), figure 1.1. The TLU calculates a weighted sum of its inputs, represented as $z = \sum_{i=1}^{n}(x_i w_i) + b_i$ where **z** is the result from the weighted sum across **n** inputs, **x** the inputs, **w** the weight assign to that input and **b** a bias added to the sum. Originally the function applied to the perceptron was a simple threshold in which if z < 0 it would output 0 and if z ≥ 0 the output would be 1, making a simple binary output. However, modern neural networks have evolved to utilize a variety of more complex activation functions to process z, moving beyond the binary threshold to enable the modeling of non-linear relationships (B. Ding et al., 2018), which has made also the term change from TLU to Artificial Neuron (AN). Regarding the activations, the Rectified Linear Unit (ReLU) function and the Sigmoid function are noteworthy for their prevalent use. The ReLU function, defined as $f(z) = \max(0, z)$ effectively enables the

network to capture non-linear patterns without the computational burden associated with more complex functions. The Sigmoid function, defined as $f(z) = \frac{1}{1+e^{-z}}$ maps the input z to a range between 0 and 1, offering a smooth gradient that is crucial for the convergence during the learning process.

Commonly, the size of a NN is measured using the term "parameters". This term refers to the number of weights (w) adjusted during the learning process. The higher this value gets the bigger the challenge for computation and training. In the case of a TLU we would have as many parameters as inputs, so i.e. if we have a TLU using two inputs such as width and height for an object, we will have a network of 2 parameters.

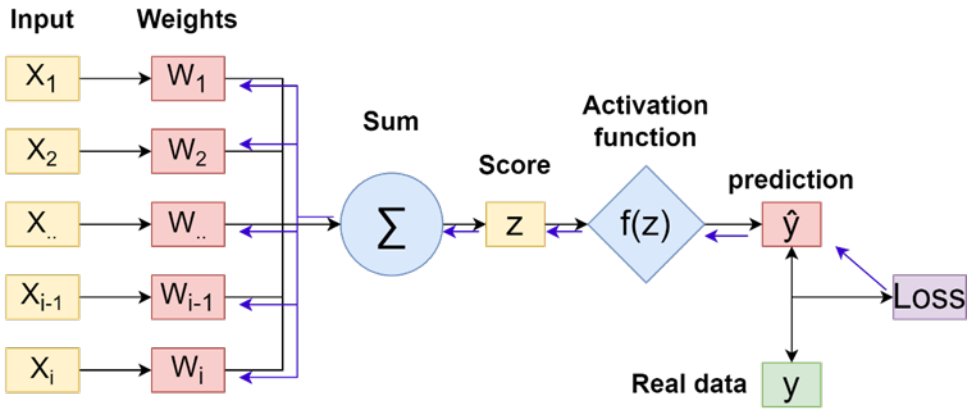


Figure 1.1: Illustration of a basic artificial neuron. Input values X_1,X_2..X_i are multiplied by corresponding weights W_1,W_2,…,W_i and the products are summed to an output z. This output is then passed through an activation function as $f(z)$ to produce the prediction denoted as ŷ. The predicted output is compared against the actual

output y using a loss function to compute the prediction error. Then the weights are corrected in a process known as backpropagation.

Dense layers and multilayer architecture

Building upon the foundational principle of the AN, dense layers represent a fundamental construction element within NNs. A dense layer, also known as a fully connected layer, is formed by stacking multiple ANs in parallel, with each ANs processing all incoming inputs. This architecture excels in extracting multiple features from the input data, albeit at a significant computational cost. The computational demand of a dense layer is quantified by the product of the number of inputs (I) and the number of ANs within the layer resulting in I*AN weights to learn. Such computational intensity can become prohibitive, especially in applications involving high-dimensional data, such as analyzing data from large SNPs arrays. For instance, a dense layer tasked with processing a 1 million SNP array, compound of 1024 ANs, would have to learn more than a billion parameters, requiring substantial computational resources for the training process. Consequently, the application of dense layers is often reserved for scenarios where input dimensions are manageable.

To mitigate the computational challenges, NNs typically employ multiple dense layers in succession. This sequential arrangement allows for a hierarchical processing of information, where each layer extracts increasingly abstract features from its inputs, which provides the network with the capability to discern complex patterns and relationships within the data. Usually, this structure is referred to as

multilayer perceptron (MLP) even though is not actually composed of TLUs.

Convolutional layers

Managing high-dimensional data presents substantial challenges, particularly when it comes to image analysis. Consider a modest color image with dimensions of 300x300 pixels; this results in 90,000 inputs per color channel, adding to a grand total of 270,000 inputs. The complexity escalates further with images exceeding the 300x300 pixel threshold, which is often the case in various research and application fields. Kunihiko Fukushima introduced a groundbreaking approach to image analysis in 1980 named the neocognitron (Fukushima, 1980). Inspired by the visual nervous system's structure and functioning, it laid the foundational principles for what would evolve into current convolutional neural networks (CNNs). The neocognitron emulated the biological visual cortex, employing a series of convolutional layers to progressively extract and recognize patterns. It introduced two key operations through its specialized cells: the S-cells (simple cells) responsible for feature extraction via convolution, and the C-cells (complex cells) tasked with feature pooling. This dual process of feature extraction and pooling forms the core of current CNN architectures. In modern CNNs, convolutional layers dynamically learn features such as edges and textures by applying filters (figure 1.2), while subsequent pooling layers reduce the spatial dimensions of the feature maps. This down sampling process efficiently decreases the amount of computation and the number of parameters needed for later layers, all while retaining

critical information (Z. Li et al., 2022). Thus, CNNs manage to significantly reduce input volume layer by layer, ensuring that relevant features are emphasized for accurate classification or recognition tasks. This method of information compression has been used to create many complex detection models and architectures (He et al., 2017; P. Jiang et al., 2022). When a model which is based on information reduction (encode) is paired with a model that does the inverse operation (information retrieval) we build an encoder-decoder

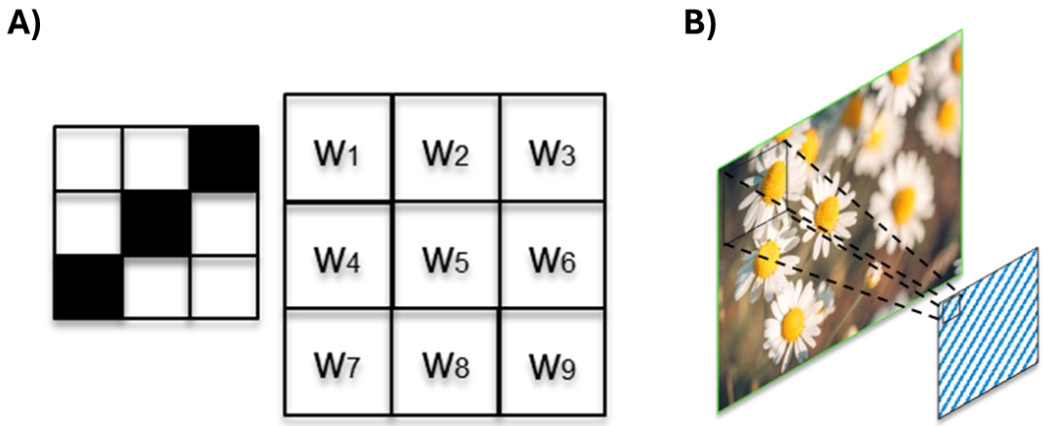


Figure 1.2: Convolutional neural networks illustrated. A) Display how the weights are reshaped into a matrix to detect patterns in images. B) These matrices are known as filters which are then passed through the full image step by step.

Detection and segmentation

Building on the principles of CNNs, many different models were developed to advance the fields of detection and image segmentation. These tasks leverage the hierarchical feature extraction capabilities of

CNNs to identify and localize objects within images and to delineate precise boundaries between different objects or regions.

Object detection, or detection, involves identifying the presence of objects in an image and locating them with bounding boxes. CNNs are employed to scan the image at multiple scales and aspect ratios, using convolutional layers to extract features that are indicative of various objects. Following the feature extraction, additional dense layers are used to predict both the class labels (e.g., apple, pear, strawberry) and the bounding box coordinates for each detected object (x1, y1, x2, y2). Computing the full model for large images can be slow and it would produce at maximum one detection, so this process typically involves a region proposal network (RPN) for object classification and localization. This network is in charge of proposing regions of the image that are most probable to contain an interesting object. The region is then processed by the object detection network which proposes where are the bounding boxes and which kind of object it is (Taghizadeh & Chalechale, 2022). Popular architectures like Faster R-CNN (Girshick, 2015), You Only Look Once (YOLO) (P. Jiang et al., 2022), and Single Shot Multibox Detector (SSD)(W. Liu et al., 2016) exemplify this approach, optimizing the balance between accuracy and computational efficiency.

Image segmentation goes a step further by performing classification to pixel level with the aim of changing the representation of an image into something more meaningful. There are two main types of segmentation: semantic segmentation and instance segmentation. Semantic segmentation assigns a class label to each pixel in the image,

making no distinction between different instances of the same class. In contrast, instance segmentation not only labels each pixel with a class but also differentiates between distinct instances of the same class. One notable architecture for semantic segmentation is the Fully Convolutional Network (FCN), which replaces fully connected layers with convolutional layers to output spatial maps instead of classification scores. These maps are then up sampled back to the size of the original image, providing a pixel-wise classification. There are many segmentation architectures, which include convolutions but the most used one, especially in biomedicine, is the U-Net (N. Siddique et al., 2021). The idea behind the U-Net is using information learned through the encoder, to reinforce the decoder which returns the map containing the labels at pixel level. For instance segmentation, models like Mask R-CNN (He et al., 2017) extend Faster R-CNN by adding a branch for predicting segmentation masks on each region of interest, enabling precise object localization and segmentation within the same framework. Mask R-CNN is the model used during this thesis in chapters 3 and 4 and has proven to be extremely useful in the instance segmentation task.

Autoencoders

Autoencoders are a class of NN architecture designed for unsupervised learning tasks, primarily focusing on the efficient encoding and decoding of data. Their primary function is to learn a compressed, dense representation of the input data, known as an encoding, which captures the most salient features of that data. This process is facilitated through a network architecture that consists of

two main components: the encoder and the decoder. The encoder is responsible for the information compression into a series of values which are named latents or embeddings. These embeddings are then received by the decoder which aims to reconstruct the original input. By training on this task, the latents become more and more relevant to the reconstruction and therefore more informative. There are many works based on the study of embeddings and their representation. For instance, they are commonly used for anomalies detection (Czibula et al., 2021; N. Shvetsova et al., 2021) and to cluster never seen examples (Hadipour et al., 2022). More importantly for this thesis, the embedding representation can be used for prediction as auxiliary networks can target embedded values from new information. Autoencoders have evolved since the basic architecture into probabilistic autoencoders, also known as variational autoencoders which learn to produce new examples from the embeddings (R. Wei & A. Mahmood, 2021). Also other patterns of configuration for the embeddings have been developed, such as vector quantization (van den Oord et al., 2017) which has derived in combinations of autoencoders and attention mechanisms (Esser et al., 2021).

How do they learn? Training a network

All the layers and architectures explained up to this section mean very little if we are not able to train the networks. The process of training means adjusting the weights of a network so it can model the inputs and produce the desired result. To train an ANN we need first a loss function. This loss function is in charge of measuring how far was the prediction from the ground truth data. One common loss function is

the Mean Squared Error (MSE) which estimates the error as the mean of the average squared difference between the real data and the prediction, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. There are many functions to estimate the loss depending on the context. For regression tasks losses such as huber, mean absolute error (MAE) and MSE are the most common while in classification the cross entropy and the variants are the most frequent (Q. Wang et al., 2022). Recently more abstract and complex losses have been developed, especially for image generation and analysis. One of the most interesting was proposed by researchers at OpenAI and Adobe, the perceptual loss (R. Zhang et al., 2018). Perceptual loss is based on detection models capabilities to extract relevant features that represent the image. The last few layers of a model can be used as a form of compressed representation and therefore can be used to compare two images in an abstract representation. This approach has proven to be, as the manuscript states "unreasonably effective" to improve image generation.

Once the error is measured a method to update the weights is required. This process is known as backpropagation and optimization and was the cause of stagnation of neural networks for almost four decades since their invention due to their computational requirements. It is an optimization problem in which first, we need to estimate the gradients using partial derivatives and the chain rule to calculate how much the weights must be corrected. Once the gradients are estimated then an optimization algorithm is used to update the weights. There are few algorithms that solve the optimization being two the most dominant in the field, the stochastic gradient descent (SGD) (Robbins & Monro, 1951; Tian et al., 2023), and the adaptive

momentum estimation (ADAM) (Kingma & Ba, 2014; Reyad et al., 2023).

This is a general approximation, but many other factors must be considered while training a neural network. It is common that a deep network might become “too good” at prediction due to overfitting which, in few words, is the network memorizing the training dataset and not generalizing. An overfitted model is easily noticed as the differences in precision between the dataset used during the training and validation dataset (one never seen before by the model) is noticeable. There are also other emergent issues related to the network number of layers such as gradients getting too large, exploding gradients, or too small, vanishing gradients. To combat all these issues there are many strategies which in general are called regularization techniques. The most common are dropouts layers, which randomly inactivate some neurons (Santos & Papa, 2022), and normalization layers, which keep the gradients within a normal range (Garbin et al., 2020).

Current applications of artificial intelligence to genomics and phenotyping

Artificial intelligence (AI) models have direct applications that are currently one of the hot topics in the field. In this chapter we will expand on their different applications and how they integrate into Smart Breeding (SB), also known as breeding 4.0(S. Jiang et al., 2020).

Applications on phenomics

The advancement of high-throughput phenotyping platforms marks a significant breakthrough in both farming and breeding sectors. Traditionally, manual phenotyping has dominated breeding programs, yet a shift towards automation is emerging, with numerous companies integrating these technologies to expedite their research and development efforts. These phenotyping platforms vary in their tools and setups, depending on the environment, whether in controlled settings like greenhouses or in open fields. In greenhouses, innovations such as hyperspectral cameras mounted on rails provide top-down surveillance (Flood et al., 2016), distinguishing individual plants, while robotics paired with cameras facilitate performance evaluation and plant management (H. I. Tangarife & A. E. Díaz, 2017). These monitoring techniques in controlled environment are naturally evolving into vertical farming which tries to build factories that provide with vegetables all year-round in an automated environment (Ji et al., 2023). In open field, the usage of unmanned aerial vehicles (UAV) combined with internet of things (IoT) based system is becoming the preferred choice (Boursianis et al., 2022) due to their cost-effectiveness ratio. Monitoring systems benefit enormously from detection and segmentation models, as NN outperform all the alternatives in image analysis tasks on the field (Barbole et al., 2022; Gené-Mola et al., 2023).

Detection is not the only area where AI is beneficial for phenomics. The other two major areas of AI, generative AI and reinforcement learning, have had a significant impact. Generative AI, through

techniques such as Generative Adversarial Networks (GANs) (J. Gui et al., 2023), plays a pivotal role in augmenting phenotypic datasets. This augmentation is not merely about expanding the quantity of data but enhancing its quality, enabling researchers to simulate a wide range of genetic variations and environmental conditions that would be costly or impractical to produce in real life (Benfenati et al., 2022). For instance, in a recent work, Drees et al. (2021) used cycle GANs (cGANs) to generate synthetic images of plants under various stress conditions, such as drought, at different developmental stages. This capability is especially crucial in the era of climate change, where adaptability to diverse and changing conditions is key to sustainable agriculture.

Reinforcement Learning (RL), on the other hand, finds its application in the dynamic and interactive aspects of plant phenotyping and care. By employing RL algorithms, autonomous systems like UAVs and robotic arms can learn and adapt their behaviors to efficiently monitor and manage crops. These systems can navigate through fields or greenhouses, identifying plants that require attention, be it watering, nutrient application, or disease treatment, and taking the necessary actions with minimal human intervention (Ferrández-Pastor et al., 2023;. Wang et al., 2020). This approach is akin to an intelligent trial-and-error process that learns from each step, optimizing for the most efficient path to achieving goals, such as understanding complex plant traits or interactions between machines and the environment.

The confluence of Generative AI and Reinforcement Learning in phenomics foresees a new era of precision and efficiency in plant

research and agriculture but also exemplifies the transformative power of AI technologies in tackling some of the most pressing challenges in food security and sustainability.

Applications to genomics

The applications of AI in genomics are now more focused than ever on two main areas: classification AI and generative AI. The study of obtained sequences represents a complex area within bioinformatics that has seen rapid evolution over the past two decades. However, since the publication of the foundational paper on Transformers and attention (Vaswani et al., 2017), the field has been experiencing a whirlwind of evolution. Transformers are at the heart of modern generative AI and are used to create a wide array of foundational models such as ChatGPT (Ray, 2023), DALL-E (Ramesh et al., 2022a), and, more relevant to this field, the Nucleotide Transformer (Dalla-Torre et al., 2023), HyenaDNA (Nguyen et al., 2023) and DNABERT (Ji et al., 2021; Zhou et al., 2024). In the last two years, we are witnessing a race to establish a foundational model from which the rest of the DNA-based predictive models will derive. Some of these models are intended for sequence classification and some for sequence generation, but most can be adapted for both tasks. The transformative impact of classification AI in genomics lies in its ability to analyze and interpret complex genetic data, by interrelating the inputs with mechanisms such as attention, or in case of HyenaDNA, with the hyena operator (Poli et al., 2023). These models are commonly pretrained on sequence masking and prediction forcing them to use the context information to predict the missing sequence.

This kind of pretraining allows to create an embedded dimensional space à la autoencoder which contains relevant information. These foundational models can be later adjusted by adding new layers, commonly named as heads, which perform the desired task. For instance, InstaDeep, a company focused on developing applications of transformers for genomics and the authors of the Nucleotide transformer, released a classification variant which predicts gene functionality at nucleotide level in mammals (de Almeida et al., 2024). Foundational models like Ensemble Nucleotide Byte-level Encoder-Decoder (ENBED)(Malusare et al., 2024), enables the generation of synthetic genetic sequences. These synthetic sequences can be used to model genetic diversity, predict the effects of genetic mutations, or even create genetic sequences that do not exist in nature but could potentially lead to advancements in synthetic biology.

All these models exemplify how transformer-based models are being tailored to understand and manipulate the language of life. The race to establish a foundational model in genomics is not merely academic; it signifies a paradigm shift towards a more unified and efficient approach to genetic research and its applications in medicine, agriculture, and biotechnology. Such a model would serve as a versatile tool for generating insights across a wide range of genomic inquiries, streamlining the predictive modeling process, and fostering innovation by providing a shared, robust platform upon which new theories and applications can be built.

Smart Breeding

Smart breeding is rapidly evolving as a response to the urgent need for cultivar adaptation to a changing climate, consumer expectations and crop diseases propagation. This process has been significantly expedited by the advent of high-throughput platforms in genomics and phenomics, which generate vast amounts of data. This data encompasses relevant information for interaction Genotype x environment x management (GxExM), which leads to a diverse array of data types. For instance, even a basic platform would take into account genetic information (e.g., FASTQ, BAM, VCF files), phenotypic data (such as raw images, size measurements, hyperspectral imagery), and environmental and management details (including soil content and temperature variations over time)(Roitsch et al., 2019). This easily builds into complex models that are difficult to develop, as multimodal representation models are sparse. The task at hand for breeding programs is to efficiently allocate resources based on a plethora of information. In this context, reinforcement learning emerges as a powerful tool for optimizing resource allocation and usage. Recent research (Moeinizade et al., 2022) shows that coupling reinforcement learning with a learning model, in this case, a random forest instead of a neural network, can enhance the efficiency of resource allocation in genome selection. However, this approach currently relies solely on genomic information.

The potential for artificial intelligence in smart breeding extends beyond genome selection to encompass a more holistic understanding of GxExM interactions. By integrating data across genomics,

phenomics, and environmental management, AI-driven models could predict the optimal breeding strategies under varying conditions, thereby significantly reducing the time and resources needed to develop new cultivars. Moreover, the integration of multimodal data into these models could unlock insights into complex traits and their responses to environmental factors, enhancing the precision and effectiveness of breeding programs. To achieve this vision for the future of breeding, multimodality with complex information is a requirement and exploring possible ways to achieve it will be crucial.

Objectives

In this thesis we aim to develop and use the current artificial intelligence applications to improve plant phenotyping, genomics, and breeding. The concrete objectives are as follows:

1. Improve fruit phenotyping using detection and segmentation methods based on neural networks as a substitute to current measurement methods based on color segmentation
2. Enable phenotyping leaf photosynthetic capacity along plant growth by coupling detection and segmentation based on neural networks with a high throughput phenotyping platform.
3. Prove the applicability of generative AI to image prediction from genetic information.

These objectives have been addressed through four chapters. In Chapter 1 we develop the Shape Analyzer, a new application to evaluate fruit morphology from fruit sections and develop an open-source repository and an interface publicly available for the scientific community.

In Chapter 2 we develop a methodology (LeTra), connecting detection and segmentation with a leaf tracking method to identify leaves along their growth and evaluate its use in the study of Arabidopsis photosynthesis performance between genotypes

In chapter 3 we create a neural network architecture that allows to retrieve images from genetic markers (SNPs) and evaluate its performance studying the similarities between predicted and real phenotypes, as well as the importance and impact of SNP preselection on predictions.

To conclude, in Chapter 4 we propose alternatives for the SNP preprocessing and handling on neural networks as well as new methods for image generation from SNPs.

Chapter 1: Shape Analyzer: An open-source application for fruit morphology phenotyping

1.1. Abstract

Phenotyping is undergoing a transformative evolution, driven largely by the integration of artificial intelligence. Despite significant advances in field phenotyping, there is a notable gap in open-source morphometric tools tailored for fruit shape assessment. To address this need, we developed Shape Analyzer, a user-friendly, open-source tool developed in Python, designed specifically for the morphometric analysis of fruits. This tool uses the capabilities of Mask R-CNN for precise fruit detection and segmentation with specialized measurements for apples, pears, and almonds, offering a streamlined solution for breeders and researchers. Our study highlights the efficacy of Shape Analyzer through comprehensive testing, showcasing positive results. These results underscore the ability of the tool to accurately capture fruit contours, essential for detailed morphometric analysis. In a practical demonstration of its utility, Shape Analyzer was employed in a Quantitative Trait Loci (QTL) mapping study on an apple biparental family with positive results. In conclusion, Shape Analyzer emerges as an open alternative in the field of fruit morphometrics. The combination of advanced image analysis techniques with the accessibility and adaptability of Python may solve the limitations of existing tools, paving the way for studies aiming to obtain insights into fruit morphology and its genetics.

1.2. Introduction

Plant phenotyping has been used to monitor a wide range of traits, playing a pivotal role in the better understanding of the functioning of crops, as well as in the identification of the genetic mechanisms involved in trait regulation (Khan et al., 2021; Naeem et al., 2021; Tsai et al., 2020). In recent years, linkage analysis and genome-wide association studies (GWAS) have elucidated numerous candidate genes associated with both commercial and non-commercial crop traits establishing GWAS as the predominant methodology for such investigations. However, a noticeable disparity remains between advancements in genotyping and phenotyping technologies. While the cost and resolution of genotyping have witnessed significant enhancements owing to the emergence of novel platforms for marker detection and identification (Ayalew et al., 2019; Darrier et al., 2019), phenotyping predominantly relies on manual labor. This has led to an imbalance in the availability of reliable data; while one can easily find publicly available single nucleotide polymorphisms (SNPs) and microsatellite (SSR) data for numerous genotypes, phenotype information often remains unknown. Fortunately, this trend is shifting, driven by initiatives such as, for example, the Netherlands Plant Eco-phenotyping Center (NPEC) (Jez et al., 2021), the European Plant Phenotyping Network (EPPN) (https://eppn2020.plant-phenotyping.eu/), the European Research Infrastructures (RI) on plant phenotyping (EMPHASIS) (https://emphasis.plant-phenotyping.eu/), or the project PHENET (https://www.phenet.eu/en). The primary objective of these projects is to augment the availability of genotype-phenotype datasets,

facilitating future research under controlled conditions. This, in turn, is expected to expedite the identification of genes pivotal to crop traits and their interactions with the environment. However, the realization of a comprehensive platform requires the development of diverse methodologies addressing various facets of this expansive domain.

One critical component of high-throughput plant phenotyping is image analysis. Recent advancements in computer vision have significantly contributed to this field, driven largely by the burgeoning developments in artificial intelligence (AI) (Fonteijn et al., 2021; Kaynak, 2021; Mishra et al., 2021; Xu & Li, 2022). These AI technologies find applications in diverse domains, ranging from robot navigation (Iqbal et al., 2020) and molecular prediction (Jumper et al., 2021) to image analysis in biology. Notably, progress in biological image analysis has been substantially influenced by advancements in medical imaging, a field that has seen a surge in AI-based publications (Frey, 2019).

In plant biology, AI methodologies are increasingly being employed to address various challenges in field conditions (Chawade et al., 2019). However, the specific issue of fruit morphometric evaluation remains unresolved. Current methods predominantly rely on manual collection and measurement or on the use of specialized software like ImageJ2 (Rueden et al., 2017) and Tomato Analyzer (Rodríguez et al., 2010). ImageJ2 is a versatile tool supported by a robust community, offering various extensions for tasks such as cell segmentation and tracking (Ershov et al., 2022). Despite its merits, ImageJ2 is limited by its inability to perform tasks that are not pre-implemented, requiring

users to code new functionalities in Java, a language less accessible than interpreted languages like Python.

Tomato Analyzer, on the other hand, is tailored for horticultural applications but is specifically designed for tomatoes. This specialization results in limitations when applied to other fruits, particularly in contour detection and measurement accuracy. Moreover, the software's source code is not publicly available, hindering any attempts at modification or improvement.

Given these limitations, there is a pressing need for more flexible, open-source tools that can be adapted to new challenges in image analysis for fruit phenotyping. To address this gap, we have developed a Python-based workflow to train a convolutional neural network (CNN) and a tool for robust fruit shape analysis. This tool, which we have named Shape Analyzer, aims to provide a time-effective solution for fruit morphometrics evaluation.

1.3. Materials and Methods.

So far, Shape Analyzer has been trained to detect and measure apple, pear and almond fruits.

1.3.1. Image Datasets.

The detection and segmentation models were trained and tested with images of fruit sections (in the case of apple and pear) or of whole fruits (in the case of almond) acquired with Mustek A3 S-Series and HP Scanjet Pro 4500 fn1 image scanner. The number of fruits per image varied depending on their size, for their fit in the scanner. Each image dataset was divided into train-test with a split of 10% individuals for testing.

For apple, the model was trained with 92 scanner images of fruits from the Apple REFPOP collection (Jung et al., 2020) containing between two and ten fruit sections of one cultivar each (Dujak et al., 2023). For almond, we obtained 269 images containing from six to ten fruits each. The pear model was trained with 226 images of fruits collected from a pear germplasm collection from the University of Lleida and IRTA, containing between four and 12 fruit sections each. Images of the apple dataset did not contain a photomacrographic scale to properly estimate the pixel-cm correspondence, while pear and almond images did.

1.3.2. Detection and segmentation.

For detection and segmentation, a Mask R-CNN model (He et al., 2017) was fine-tuned. The implementation of Mask R-CNN was obtained from PyTorch (Paszke et al., 2019) along with weights pretrained in the COCO dataset (Lin et al., 2014). As back-end for the feature extraction network, a resnet50 (He et al., 2015) was selected to keep the training feasible inside a GPU with 8GB of dedicated RAM. The training was done following the PyTorch repository instructions and advice. The original code from the PyTorch repository was added to this project library to facilitate the fine-tuning to all users.

1.3.3. Metrics given by shape analyzer

Most measurements and key point locations were common to all fruit types, while a few were fruit specific. The common attributes were Fruit Shape Index (FSI), Box Ratio, Compactness, Roundness, Solidity, and Symmetry, the latter using the Jaccard index (Figure 1.1).

The Fruit Shape Index (FSI) was calculated as the ratio between the inner height and width of the fruit, with the height being the distance between the top and the bottom parts of the fruit.

The Box Ratio was defined as the ratio between the height and width of the bounding box that delimits the fruit shape. Typically, Box Ratio and FSI are highly similar, but in some cases, the inner height and width might exceed the box height and width. For apples, for the better

comparison of samples, the height was measured using the largest lobule, thus avoiding biases in asymmetric apples.

$$Box\ Ratio\ = \frac{Height_{bbox}}{Width_{bbox}}\ FSI\ = \frac{Height_{inner}}{Width_{inner}}$$

Roundness, Compactness, and Solidity are metrics related to the fruit circularity. Compactness is the ratio of the fruit area to the area of a circle with the fruit perimeter, with a maximum of one for a perfect circle. Roundness is similar to compactness, but local irregularities are excluded by using the convex perimeter (i.e. the outer perimeter excluding the stalk and calix cavities) instead of the cortex perimeter. Solidity measures object density, being a direct ratio between the area and the convex area.

$$Compactness\ = \frac{4 * \pi * Area}{Perimeter}$$

$$Roundness = \frac{4 * \pi * Area}{Convex\ Perimeter}\ \ Solidity = \frac{Area}{Convex\ Area}$$

To measure symmetry, the bounding box containing the object mask was divided in half. The right side was flipped and compared to the left side using the Jaccard coefficient metric (also known as Intersection over Union, IoU).

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

In the case of apples, additional measurements were calculated to assess fruit symmetry. The ratio between the height of the larger and smaller lobules defined the Symmetry Ratio. Additional parameters include the distance between the top (stalk cavity) and bottom (eye basin) lobules and their respective ratios, which help in capturing the conicity of the apple. Finally, the angle at the stalk insertion, as well as the angle at the fruit eye were also defined to provide a comprehensive analysis of the apple shape.

In the case of pear, ShapeAnalyzer calculates the distance from the center of the pear to the bottom using the inner height and width, and the depth of the eye.

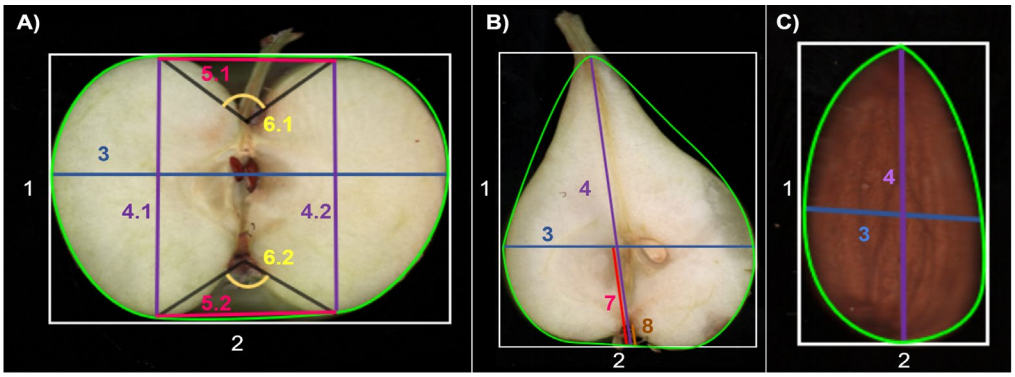


Figure 1.1: Metrics obtained with shape analyzer in apple (A), pear (B) and almond (C) fruits. In all panels, 1 represents the height of the box and 2 its width; they are used to estimate the bounding Box Ratio. 3 is the inner width of the fruit and 4 the inner height; the ratio of both is the Fruit Shape Index. In the case of apple 4.1 and 4.2 measure lobule height. 5.1 and 5.2 measure the distance between the top and bottom lobules (ie. the stalk cavity and the eye basin), being their ratio a

measure of conicity. 6.1 and 6.2 are also unique to apple and represent the angles formed by the eye and the stalk, respectively. In the case of pear, 7 is the distance from the center of the pear to the bottom using the inner height and width, and 8 is the depth of the eye.

1.3.4. Development of the desktop application.

The frontend of Shape Analyzer was implemented using the Vue (You et al., 2024) JavaScript Framework, a versatile and performant framework renowned for its reactivity and ease of use. Complementing this, the Bootstrap (Otto et al., 2024) frontend toolkit, was employed to create an intuitive interface with an array of user-friendly functionalities, facilitating efficient data acquisition, display, and retrieval, which is required in the context of image phenotyping.

The backend logic was developed inside Flask (Pallets, 2024), a micro web framework for Python. Flask's minimalistic design allowed for a focused and tailored implementation, while being compatible with Python's robust machine learning and image analysis libraries. It serves as the brain of the software, including the algorithms and computations to detect and extract the shape and size metrics and attributes.

To ensure that Shape Analyzer is readily accessible to users across various platforms as an installable program, Electron, a framework for building cross-platform JavaScript-based desktop applications, was used. This technology, in conjunction with PyInstaller, which is a library that bundles python code in a single executable package, was

instrumental in constructing a distributable, self- contained executable. This executable, upon deployment, seamlessly initiates both the JavaScript frontend and Python backend components on the user's local machine, requiring no further internet connectivity. The Shape Analyzer application has been developed to be accessible with a minimalistic interface (Figure 1.2.A). The resulting metrics can be downloaded to disk for further analysis as well as seen in the results section of the app (Figure 1.2.B.C.D).

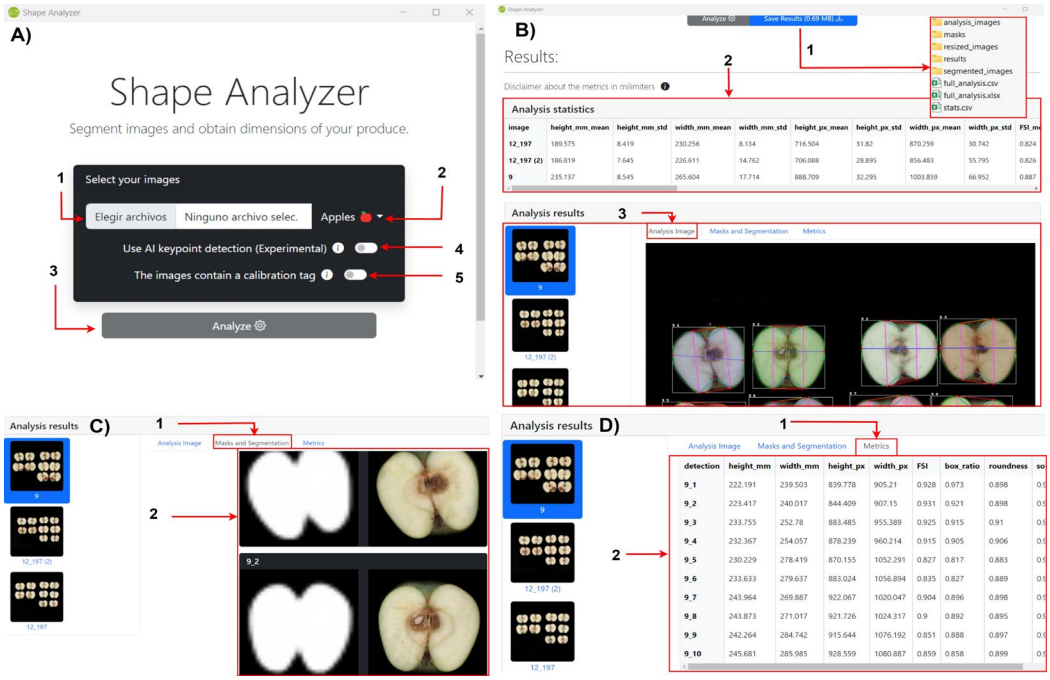


Figure 1.2: Shape Analyzer desktop application. Panel A) shows the application at the main menu. Files are either selected or dragged to A.1 and the fruit type is selected in A.2. Once the files are loaded, the image analysis can be started by clicking on the "analyze" button, A.3. A.4 is an experimental alternative mode for apple key points location if the

fruits in the images are rotated. When there is a calibration tag as described in the methods, A.5 can be selected to allow for a precise conversion of pixels to millimeters. Panel B) shows the results. By clicking on "save results", B.1, a zip file is generated and downloaded with all the information and data of the analysis. The B.2 panel displays the mean and standard deviation values for each metric and each image, the "Analysis image" window in B.3 shows the detection and measurement lines over the image. Panel C) displays the results section in the "Masks and segmentation" window C.1, and C.2 shows the individual images and their corresponding segmentation mask. Last, in Panel D), the "Metrics" tab, D.1, displays the values for the metrics evaluated on each detection, D.2.

1.3.5. Detection and segmentation evaluation.

Mask-RCNN training quality was evaluated using the number of fruits properly detected (True Positive, TP), against the number of false objects detected as true (False Positive, FP) and against the number of fruits missing from the actual ground truth data (False Negative, FN). The evaluation was done in a separated evaluation dataset never used for the training phase. This dataset contained a total of 132 apples, 158 pears and 131 almonds. With these values, precision, recall and F1 score were determined. Precision, defined as the ratio of TP against all the detections of the model, express how truthful are the detections given by the model:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall is defined as the ratio between TP against all TP adding the number of detections missing (FN). This ratio express how sensitive is the model to miss true objects.

$$Recall = \frac{True\ positives}{True\ Positives + False\ Negatives}$$

F1 also known as F-Score is the harmonic mean of both metrics, bringing both concepts into balance:

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

We evaluated the segmentation quality using two concepts, Jaccard coefficient, also known intersection over the union as defined in section 3.3.3, and the Sørensen-Dice Coefficient. Both express how two sets intersect.

$$DSC\ (A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

Validation of Shape Analyzer for genomic studies

1.3.6. Plant material, DNA extraction, SNP genotyping and Fruit phenotyping.

The biparental family PM220 (T34xT99) was used to validate the Shape Analyzer software in breeding materials. This family contained 149 offsprings grown at the IRTA experimental station in Gimenells, Catalonia, Spain. Young leaves were collected for DNA extraction using the CTAB method as outlined by Doyle and Doyle (1990). DNA quality and integrity were analyzed with a Nanopore® spectrophotometer and agarose gel, respectively. All samples, including the two parents, were sent to Illumina®, New Zealand, for genotyping with the 8K RosBREED Apple Illumina® Infinium SNP array (Chagnè et al. 2012). Ten samples had to be discarded for errors in genotyping. Their phenotype information was considered for the heritability estimation but not for the QTL mapping as the genotype information was not reliable, leaving a total of 139 offsprings.

For the phenotyping, five to fifteen fruits per tree were collected at the ripening stage, cut across the longitudinal axis and scanned with Mustek A3 S-Series and HP Scanjet Pro 4500 fn1 image scanner. This process was carried for a period of 2 years taking samples from 2019 and 2020. We used Shape Analyzer to extract metrics for each of the 139 offsprings of the family in two harvest seasons. Each genotype underwent a minimum of 10 evaluations, with some being assessed as

many as 62 times. These evaluations were then averaged to yield a singular value representing the phenotype for the respective attribute using python. These averages were later used to study the population distribution and to perform the QTL mapping.

1.3.7. Heritability, Linkage map construction and QTL mapping.

Broad sense heritability (H^2) was estimated for each of the traits using the R package lme4 (Bates et al., 2015). To build a balanced dataset for the heritability estimation we randomly selected 5 samples per year and per genotype to match the genotype with less samples per year. The heritability computation was repeated for 100 random samplings, to avoid the bias of a single sampling. The random effect model was defined as $y_{ijk} = \mu + g_i + z_j + gz_{ij} + \varepsilon_{ijk}$. Where y_{ijk} is the k th phenotypic value of genotype i on the year j . μ is the grand mean, g_i is the random effect of the i -th genotype, z_j is the fixed effect of the j -th year, gz_{ij} the random interaction effect between the i -th genotype and the j -th year and ε_{ijk} is the error term.

The broad sense heritability is described with the following equation: $H^2 = \sigma_g^2/\sigma_p^2$, where σ_g^2 is the genotypic variance and σ_p^2 is the phenotypic variance. The phenotypic variance was computed by adding the total variance described by the random effects and error term:

$$\sigma_p^2 = \sigma_g^2 + \frac{\sigma_{gz}^2}{n_z} + \frac{\sigma_\varepsilon^2}{n_z n_r}$$

where σ^2_{gz} is the genotype by year interaction variance, σ^2_ε is the error variance, $n_z = 2$ the number of years, $n_r = 5$ the number of samples per year. After fitting the model, the variance of each random effect was used to compute the broad sense heritability following the formula described, for a subset of samplings and traits, the model did not converge, the non-convergent models were not used for heritability computations.

Raw SNP data was processed with GenomeStudio® software in IDAT format. SNPs with call rate <5% and error rate >1% were excluded for the analysis. The linkage map was constructed with JoinMap 5.1 (van Ooijen, 2011), using the cross-pollinated (CP) coding of genotypes. Three mating types were discerned within the biparental family: hk×hk, lm×ll, and nn×np. Of these, hk×hk denoted markers segregating biparentally, while lm×ll and nn×np signified maternally and paternally segregating markers, respectively. Two distinct maps were constructed, one for the maternal parent (T34) and another for the paternal parent (T99). Grouping was done with a threshold of LOD≥3, and distance were calculated with the Kosambi's mapping function (Vinod, 2011) . Reference points for chromosome/linkage group numbers, orientation, and physical positions adhered to the apple 'Golden Delicious' reference genome v1 (Velasco et al., 2010). With the linkage map built, we proceeded to map fruit shape and size QTLs using MapQTL 6 (van Ooijen, 2009). The mean values obtained per genotype were uploaded to the MapQTL software along with the maps, and SNPs. Two separate analyses were performed, one for each parental map. Interval mapping was selected

as the mapping method, and only positions with logarithm of the odds (LOD) superior to 3 were considered as valid peaks.

1.4. Results

Shape Analyzer is an open-source tool to measure fruit morphology in fruit sections, built on top of a modern detection system.

1.4.1. Image detection and segmentation accuracy

The detection and segmentation models were trained and tested with images of fruit sections (in the case of apple and pear) or of whole fruits (in the case of almond). Five key metrics were utilized to evaluate the Mask R-CNN model in the validation panel. For all fruit types, the metrics that indicate the prowess of the model in correctly detecting the fruits (i.e, the Precision, Recall, and F-score (F1)) reached the highest value (equal to 1), while the values of the metrics indicative of the segmentation accuracy (the Intersection over the Union (IoU) and of the Dice coefficient) reached values over 0.97 and 0.99, respectively (Table 1).

Table 1.1: Accuracy and ability of the Mask R-CNN model to correctly detect the fruit sections on each of the validation datasets.

Fruit	Detection Recall	Detection Precision	Detection F1	IoU	Dice Coefficient
Apple	1.0	1.0	1.0	0.972	0.986
Pear	1.0	1.0	1.0	0.973	0.986
Almond	1.0	1.0	1.0	0.98	0.99

To demonstrate the validity and precision of the metrics provided by the Shape Analyzer in fruit sections, we conducted a QTL mapping of apple size and fruit shape attributes in an apple bi-parental F1 family segregating for such traits.

1.4.2. Shape analyzer application on the QTL mapping of apple shape and size traits.

The distributions of the mean values of the measures obtained per each F1 genotype are shown in Figure S1.1. In general, the values followed a normal distribution, with few fruits showing extreme values. Fruit Shape Index (FSI) and Box Ratio values ranged between 0.82 to 1.15, following within the categories from slightly flat to elongated shapes according to the ECPGR categories. The fruit conicity (fruit eye basin/stalk cavity) ranged from 0.77 to 1.15, placing the fruits between the truncate-conical and cylindrical ECPGR classes. The Jaccard-based symmetry indicated high symmetry across the apple population with ranges of 0.91-0.98 and 0.85-0.92 respectively.

The Broad Sense Heritability of the attributes ranged from 53.6% to 86.8% in most of the traits excepting those related to symmetry which scored the lowest values with 13.3% and 16% for symmetry and Jaccard symmetry respectively.

Table 1.2: Heritability results for each of the traits using lme. A hundred rounds of sampling were performed to build a balanced dataset for the estimation (see Materials and Methods). Lowest, highest, and mean heritability achieved during the process are displayed for each of the metrics.

Metric measured	Lowest estimation	Higher estimation	Mean Heritability	Std
Height in pixels	71,2%	73,0%	72,0%	0,6%
Width in pixels	68,8%	71,8%	70,2%	1,0%
FSI	86,1%	87,6%	86,7%	0,5%
Box ratio	86,1%	87,4%	86,8%	0,4%
Roundness	69,6%	75,3%	72,6%	1,8%
Solidity	69,0%	76,0%	72,6%	2,1%
Compactness	66,3%	71,9%	69,0%	2,1%
Cone	71,0%	76,7%	74,0%	2,1%
Symmetry	3,4%	23,2%	13,3%	6,8%
Symmetry Jaccard	7,3%	29,1%	16,0%	8,9%
Top (stalk) angle (Stalk cavity)	50,9%	55,6%	53,6%	2,1%
Bottom (eye) Angle	67%	73%	70%	4%

The metrics of the attributes representing the same concept showed very strong correlation, this is the case of the box ratio and FSI ($\rho = 1$), the lobule and Jaccard symmetries ($\rho = 0.72$), and roundness, solidity and compactness ($\rho \geq 0.77$). None of the attributes evaluated showed correlation with fruit symmetry (average of $\rho = |0.2|$). The values of the top and bottom angles showed moderate to strong negative correlation with the fruit height, and strong positive correlation with fruit compactness.

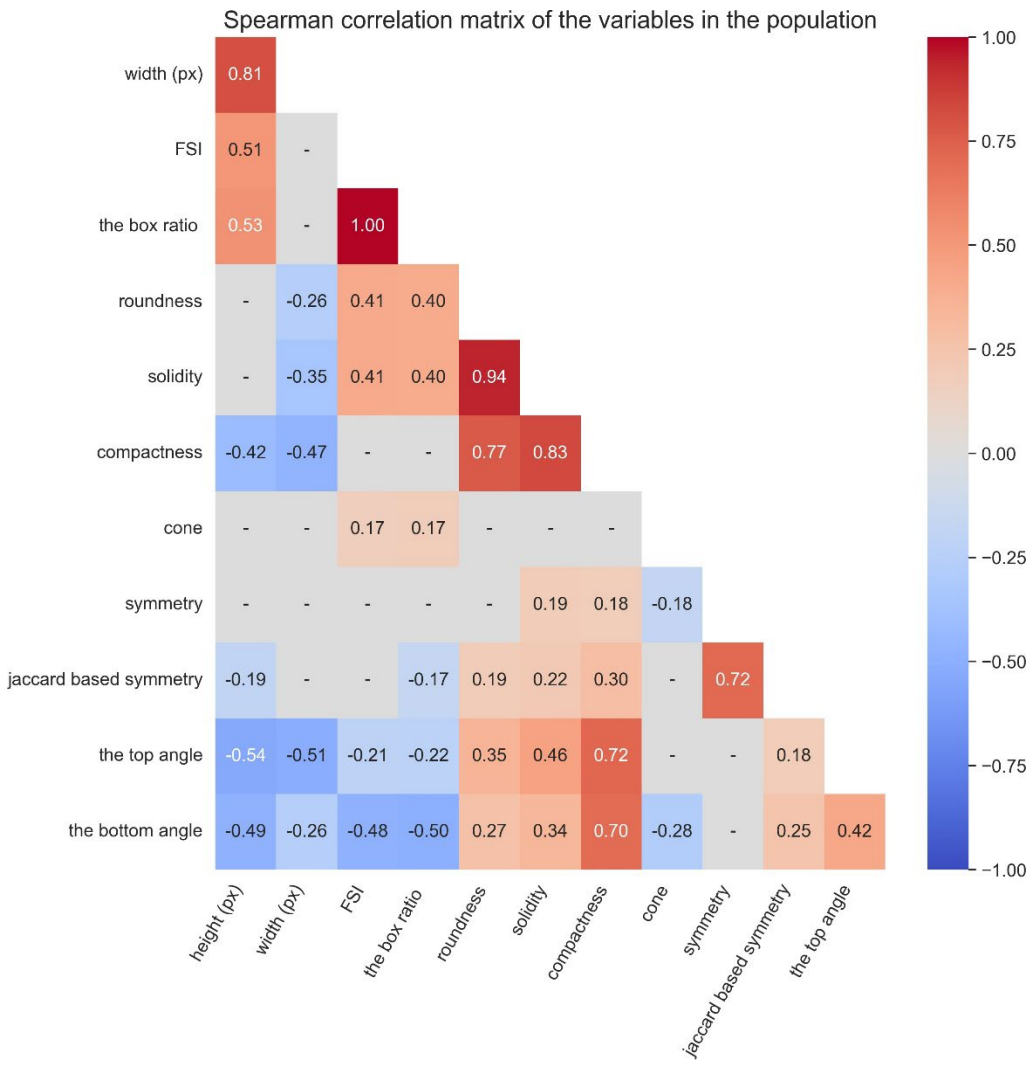


Figure 1.3: Matrix for the Spearman correlation coefficient (ρ) for pairs of variables measured. Only the values with significant correlation ($p < 0.05$) are shown.QTL mapping.

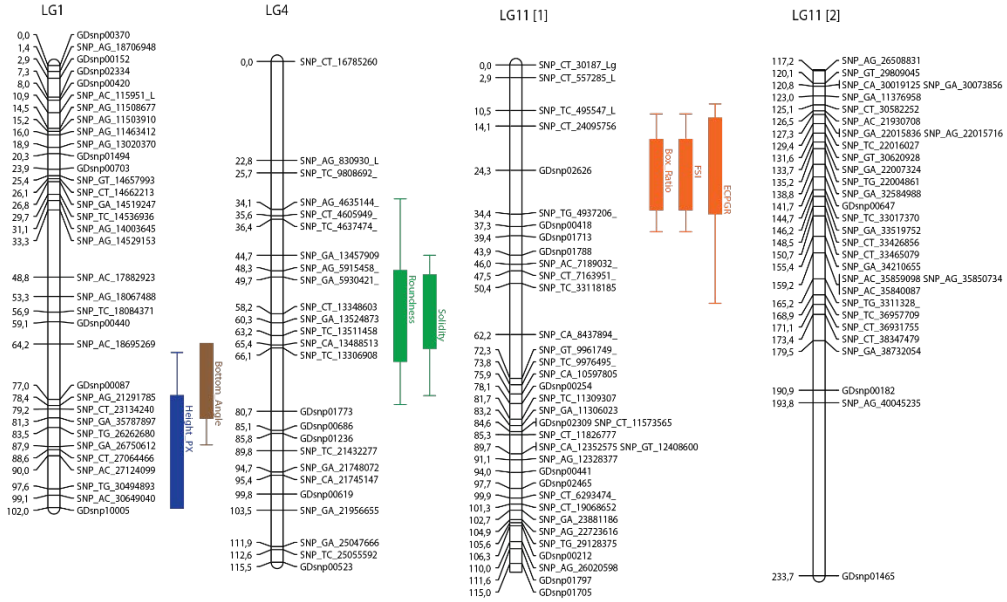
The QTL mapping was stratified based on parental contribution (Figure 1.4). Full maps for each of the parents can be found in the supplementary materials S1 and S2. QTLs (LOD >3) for size, general shape, circularity, conicity, and angle were found in at least one of the

parents. NO QTL was found for symmetry attributes. Parent 1 contributed with QTLs on three linkage groups. QTLs for the 'height' and the 'bottom (eye) angle' were detected near the terminal region of LG1. Two QTLs, one for 'roundness' and one for 'solidity' were located near the central region of LG4, and three QTLs for the 'box ratio', the 'FSI', and the 'ECPGR categorization' QTLs were identified adjacent to the proximal region of LG11. Parent 2 contributed with three QTLs, one for the 'cone' attribute at the top of LG14, and two in LG15: one for the 'Box Ratio' and one for the FSI LG15. The explained variance for all identified QTLs ranged from 9.8% to 12.6%, indicating a moderate effect size and suggesting these loci are important contributors to the phenotypic variation observed (Table 1. 3).

Table 1.3: QTL mapping results for the metrics obtained with Shape Analyzer. Only the metrics with LOD> 3 are shown.

Metric measured	Metric group	Parental donor	Linkage Group	Peak LOD	Variance Explained	Closest Marker to the peak
Height in pixels	Size	1	1	3,86	11,9	GDsnp10005
Bottom angle	Angle	1	1	3,19	9,8	GDsnp00087
Roundness	Circularity	1	4	3,29	10,2	SNP TC 3511458
Solidity	Circularity	1	4	3,29	10,2	SNP TC 13511458
Box ratio	General shape	1	11	4,14	12,6	GDsnp02626
FSI	General shape	1	11	4,13	12,6	GDsnp02626
ECPGR	General shape	1	11	3,54	10,9	GDsnp02626
Cone	Cone	2	14	3,26	10,1	SNP GA 29630036
Box ratio	General shape	2	15	3,57	11	SNP AG 27079169
FSI	General shape	2	15	3,37	10,4	SNP AG 27079169

A) QTLs for Parent 1



B) QTLs for parent 2

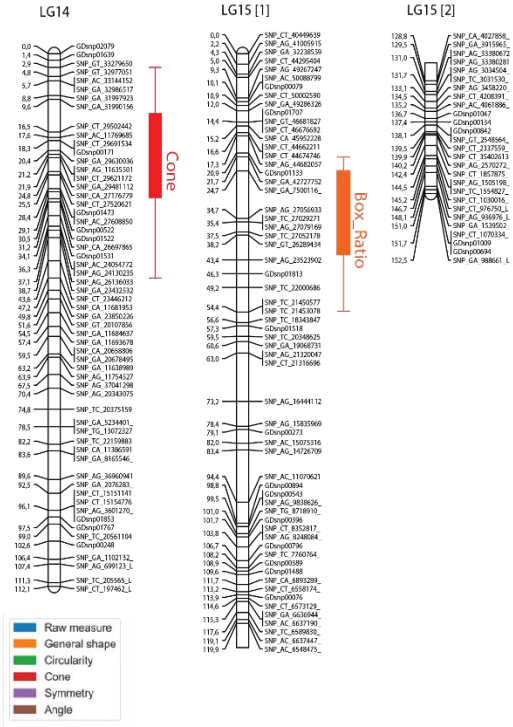


Figure 1.4: QTLs found for the different attributes measured with Shape Analyzer considering *LOD* >3. Two different maps were constructed, one per parent. Parent 1, in panel A) shows QTLs for height, bottom angle in LG 1, roundness and solidity in LG 4 and for box ratio, FSI and the categorization using the ECPGR system on LG 11. Parent 2, in panel B) shows QTLs for the cone metric in LG 14 and for the box ratio in LG 15. Only LGs with QTLs are displayed in the figure.

1.5. Discussion

1.5.1. Boosting open software and Easy-to-Use tools for the morphometrics community

The principal objective behind the development of the Shape Analyzer tool was to create an accessible, modular, and customizable resource for the scientific community, addressing a need in the study of fruit morphometrics. Recognizing the limitations of existing tools, our approach was to harness the versatility and widespread adoption of Python, a programming language renowned for its ease of use and strong presence in the scientific community. Python's inherent user-friendliness and the open-source nature of Shape Analyzer stand in stark contrast to the proprietary constraints of other software for fruit shape analysis, such as the Tomato Analyzer or ImageJ2. While Tomato Analyzer offers specialized functionalities for certain horticultural applications, the closed-source codebase limits its utility for broader research purposes. By making Shape Analyzer's source code publicly available on GitHub, we have not only ensured accessibility but also

opened the door for continuous improvement and customization by researchers worldwide. In comparison with ImageJ2, another widely used tool in image analysis, Shape Analyzer offers enhanced accessibility. While ImageJ2 boasts a robust community and a variety of extensions, it requires users to be proficient in Java to implement functionalities not already pre-set within the software. This requirement can be a significant barrier for researchers who are familiar with programming but not experts in low-level languages. By employing Python for the development of Shape Analyzer, we have lowered the entry barrier, enabling a wider range of scientists to contribute to and benefit from these tools, fostering a community-driven model that encourages innovation and rapid advancement in the field.

1.5.2. Mask R-CNN performance and adequacy.

The remarkable performance of the Mask R-CNN model, as evidenced by the high values of F1, IoU and Dice coefficients, marks an advancement in image analysis for fruit phenotyping. These results underscore the tool's proficiency in delineating precise fruit contours, which is critical for accurate morphometric analysis. Such precision in detection and segmentation paves the way for more nuanced insights into the phenotypic traits of various fruits, allowing for a more precise understanding of fruit morphology. The capacity to incorporate specialized models for segmenting fruit components through targeted training offers considerable adaptability and potential precision. However, this adaptability entails associated costs. Attaining high

precision for a novel subject requires fine-tuning the mask R-CNN model for the specific task, requiring the labeling of a new dataset to facilitate model learning. Even though this can be achieved with a low number of examples, manual work is still required. New advances in segmentation such as Meta segment anything model (SAM) (Kirillov et al., 2023) might alleviate this problem in the future with a pre-segmentation of all the objects in the image, but currently this can only be performed manually.

1.5.3. Evaluating shape analyzer measures adequacy.

Most of the metrics showed normal distribution along with slight skewness in certain cases, suggesting uniformity in the morphological traits across the studied family. Furthermore, the Spearman correlation coefficient analysis underscored strong correlations between metrics representing similar concepts, such as the box ratio and FSI. This correlation not only validates the reliability of the Shape Analyzer in these metrics but also highlights the interrelatedness of different morphological traits. These results display the tool's utility in capturing meaningful and relevant data that can be directly correlated to genetic factors, thereby enhancing our understanding of fruit morphology and its genetic underpinnings.

The purpose of our QTL mapping was primarily to validate the relevance and adequacy of the morphometric measurements obtained through the Shape Analyzer tool. While previous studies have

established associations for fruit morphology (Dujak et al., 2024; Keller et al., 2024), our focus was not on discovering new loci but rather on demonstrating that the metrics provided by Shape Analyzer are significant and can be reliably linked to genetic factors. The family used for this study was genotyped with an array of relatively low density, which produced a map with relatively low resolution and larger chromosomes than expected. Our analysis revealed moderate effect sizes for the identified QTLs, affirming that while these loci are important, they represent only a part of the genetic components influencing fruit morphology. As in Dujak et al. (2023) and in Keller et al. (2024), we identified a QTL for shape at the top of chromosome 11, co-localizing with two genes of the Ovate Family Protein which have been previously described for their role in fruit shape (Rodríguez et al. 2011, Dujak et al. 2023). QTLs for size in Chromosome 1 were also found by Keller et al. (2024). Interestingly, we identify new QTL for fruit in LG4, LG14 and LG15, although explaining a low fraction of the variance. This finding underlines the complexity of genetic control over these traits and the importance of considering a range of metrics for a comprehensive understanding.

1.6. Conclusion

The Shape Analyzer tool, developed in this study, marks a significant advancement in fruit morphometrics, combining the accessibility of Python with the sophistication of modern image analysis techniques. Its introduction challenges the limitations of existing tools, enhancing the scope and ease of research in plant phenotyping. The effectiveness of the Mask R-CNN model, evident in its precise detection and

segmentation capabilities, paves the way for deeper insights into fruit morphology. Moreover, the validation of Shape Analyzer's metrics through QTL mapping and correlation analysis underlines the tool's utility in bridging the gap between phenotypic traits and genetic factors. This study not only contributes a valuable tool to the scientific community but also underscores the importance of accessible, open-source software in advancing research in fruit morphology and beyond.

1.7. Code availability

All the code for the project as well as the compiled application is host in github at the following address:

https://github.com/aranzana-lab/shape-analyzer-app

Chapter 2: LeTra: A leaf tracking workflow based on convolutional neural networks and intersection over union.

Published in: Plant Methods.

Authors: Federico Jurado-Ruiz, Thu-Phuong Nguyen, Joseph Peller, María José Aranzana, Gerrit Polder, Mark G.M. Aarts.

Doi: 10.1186/s13007-024-01138-x

2.1. Abstract

The study of plant photosynthesis is essential for productivity and yield. Thanks to the development of high-throughput phenotyping (HTP) facilities, based on chlorophyll fluorescence imaging, photosynthetic traits can be measured in a reliable, reproducible and efficient manner. In most state-of-the-art HTP platforms, these traits are automatedly analyzed at individual plant level, but information at leaf level is often restricted by the use of manual annotation. Automated leaf tracking over time is therefore highly desired. Methods for tracking individual leaves are still uncommon, convoluted, or require large datasets. Hence, applications and libraries with different techniques are required. New phenotyping platforms are initiated now more frequently than ever; however, the application of advanced computer vision techniques, such as convolutional neural networks (CNN), is still growing at a slow pace. Here, we provide a method for leaf segmentation and tracking through the fine-tuning of Mask R-CNN and intersection over union as a solution for leaf tracking on top-down images of plants. We also provide datasets and code for training and testing on both detection and tracking of individual leaves, aiming to stimulate the community to expand the current methodologies on this topic. We tested the results for detection and segmentation on 523 Arabidopsis thaliana leaves at three different stages of development from which we obtained a mean F-score of 0.956 on detection and 0.844 on segmentation overlap through the intersection over union (IoU). On the tracking side, we tested nine different plants with 191 leaves. A total of 161 leaves were tracked without issues, accounting to a total of 84.29% correct tracking, and a Higher Order Tracking

Accuracy (HOTA) of 0.846. In our case study, leaf age and leaf order influenced photosynthetic capacity and photosynthetic response to light treatments. Leaf-dependent photosynthesis varies according to the genetic background. The method provided is robust for leaf tracking on top-down images. Although one of the strong components of the method is the low requirement in training data to achieve a good base result (based on fine-tuning), most of the tracking issues found could be solved by expanding the training dataset for the Mask R-CNN model.

2.2. Introduction

There are increasingly more high throughput (HTP) phenotyping technologies becoming available, which provides increasingly more methods to support plant breeding efforts for crop improvement. HTP phenotyping platforms are largely based on a wide range of imaging technologies, including visible light (blue 450 nm, green 550 nm and red 600 nm), chlorophyll fluorescence (CF), thermal, hyperspectral imaging and more, reviewed in (Jangra et al., 2021). As a result, a large number of plants at various developmental stages can be measured automatically in a nondestructive manner for target phenotypes. Such large-scale screening enables the selection of the best performing genotypes as well as best phenotypes as predictors of complex traits such as stress tolerance, disease resistance and yield. At the same time the genetic basis of complex traits can be revealed in genetic studies where high-density genetic makers and accurate phenotypes are associated with large-scale population (genetic mapping for quantitative trait locus). Photosynthesis is considered a difficult trait

for breeding or for genetic studies because of its polygenic nature and high sensitivity to environmental changes. This makes photosynthesis parameters a good indicator of biotic and abiotic stresses as well as plant productivity that potentially associates with yield (Chen et al., 2017; Kalaji et al., 2012; Suárez et al., 2022). CF imaging is becoming a powerful and popular technique for measuring efficiency of photosystem II (ϕPSII). The HTP phenotyping platform using CF imaging, the so-called Phenovator (Flood et al., 2016), was developed at Wageningen University, which led to successful studies on natural variation for photosynthesis in the model species *Arabidopsis thaliana* (Prinzenberg et al., 2020; Rooijen et al., 2017). An advanced photosynthesis phenotyping platform, built by Photon Systems Instruments (PSI), is now embedded in the Netherlands Plant Eco-phenotyping Center (NPEC; www.npec.nl), which setup is now available for *A. thaliana* and other species with various plant sizes and architectures. The current image analysis does not include automated individual leaf tracking over time, but as phenotypic variation in leaf morphology and other, e.g. photosynthesis, parameters are observed between developmental stages and between leaves (Bielczynski et al., 2017; Tisne et al., 2008; Wingler et al., 2004), the development of leaf detection and segmentation methods on images is required. Many methods have been developed for leaf segmentation(Itakura & Hosoi, 2018; D. Li et al., 2019; Scharr et al., 2016), but recent advances in artificial intelligence (AI) allows for new approaches. The implementation of the most recent tools for segmentation, such as Mask-RCNN (He et al., 2017) remains to be explored. These models have the advantage of performing both detection and segmentation in one step and being fast in inference times, simplifying the pipeline. All

these methods (AI based or not) still have a problem on time series association. The instance detection order varies, depending on the position of the leaves, requiring an association step to produce a time series for each individual leaf which might be challenging to implement (Yin et al., 2014). As an alternative, we find the use of complex AI methods capable of both tasks, but these methods require abundant training data, which might be expensive to produce timewise. Therefore, in this manuscript we suggest a hybrid approach combining CNN to solve detection and segmentation, with a numerical method to perform association between time points. We also present an analysis of photosynthesis data of two Arabidopsis accessions (Columbia-0 and Ely) as a case study. Results showed that the photosynthetic capacity and photosynthetic response to light treatments are influenced by leaf age and order. There is variation in leaf-dependent photosynthesis, which is of great interest for quantitative genetic studies. The main contributions of this research are summarized as follows: 1) A study on the efficiency of Mask R-CNN to individualize leaves in Arabidopsis. 2) A method based on intersection over union (IoU) to track individual leaf growth on temporal series. 3) A novel comparative study of the photosynthesis efficiency between Columbia-0 and Ely at individual leaf level. 4) A repository with code and a dataset to recreate the work and train other detection models.

2.3. Results

2.3.1. General Approach

The proposed method first employs Mask R-CNN to detect leaves and generate their corresponding binary masks (Figure 2.1.A). Then, the IoU is used to measure the similarity between the masks in consecutive frames and find the correspondence between them (Figure 2.1.B). This correspondence information is further utilized to track the leaves over time. The experimental results demonstrated the effectiveness of the proposed method in accurately detecting and tracking leaves in a top-down setting.

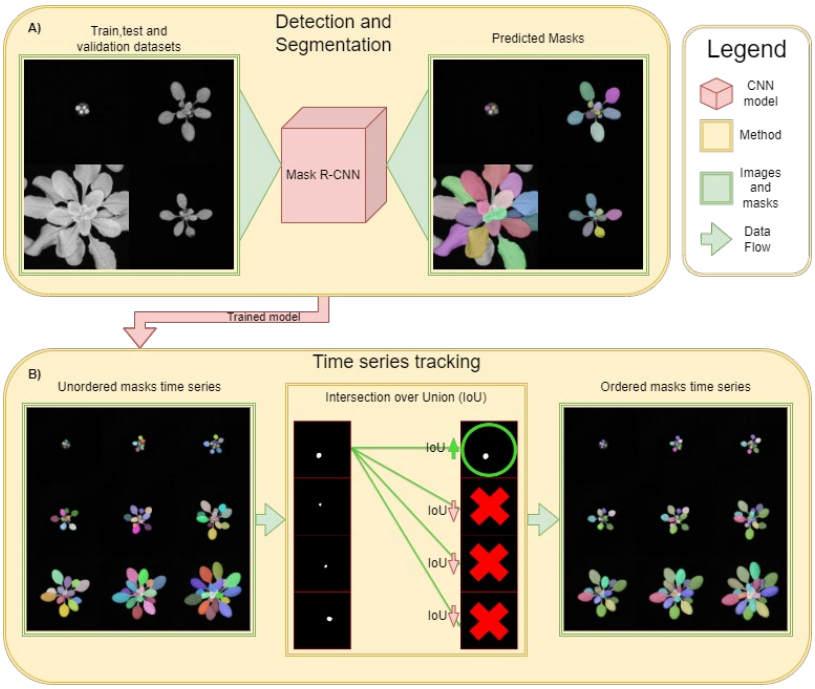


Figure 2.1: Graphical abstract of the proposed approach, consisting of two main stages. In the first stage (Panel 3.1.A), a detection and

segmentation model is trained using images of *A. thaliana* rosettes. In the second stage (Panel 3.1.B), the trained model is employed to detect and segment individual images at each time step. The resulting masks are ranked according to their IoU as denoted by green and red arrows representing high and low IoU values, respectively. This ranking enables the association and reconstruction of object tracking.

2.3.2. Leaf detection and segmentation

A total of 523 different leaves were evaluated, categorized by origin. A total of 304 came from plants in the last stages of development, 151 at the mid stages and 68 at the very early stages. In the first contrast we evaluated how many of them were effectively (True positives, TP) detected against how many of them were not (False negatives, FN) and against how many false detections (false positives, FP) were found. From these leaves we found 294/10/26 (TP/FN/FP) for the late stage, 146/5/7 (TP/FN/FP) for the mid stage and 65/3/1 (TP/FN/FP) for the early stages. Based on these values we estimated the precision (P), recall (R) and F-score (Fs) for each stage, having 0.9187/0.9671/0.9423 (P/R/Fs) for the late stage, 0.954/0.969/0.96 (P/R/Fs) for the mid stage and 0.985/0.9558/0.9701 (P/R/Fs) for the early stage (Table 3.1).

Table 2.1: Results for the detection of the Mask R-CNN model on the validation dataset without considering the segmentation.

Stage	True positive	False negative	False positive	Precision	Recall	F-score
Early	65	3	1	0.98	0.96	0.97
Mid	146	5	7	0.95	0.97	0.96
Late	294	10	26	0.92	0.97	0.94

We also assessed the adequacy of the segmentation for each true positive detection. For this task, we evaluated the IoU between the predicted masks and the ground truth mask. We obtained a mean IoU of 0.79 ± 0.0731 standard deviation (STD) for the early stages, 0.8795 ± 0.0968 for the mid stages, and 0.8631 ± 0.1466 for the late stages(Table 2.2).

Table 2.2: Statistical descriptors for the segmentation. The value represents the IoU between the model predicted masks and the ground truth masks on correct detections for the validation dataset separated by early, mid, and late stages of development.

	Early	Mid	Late
Total	65	146	294
Mean	0.79	0.88	0.86
STD	0.07	0.10	0.15
Min	0.59	0.41	0.20
25%	0.74	0.87	0.87
50%	0.80	0.92	0.91
75%	0.84	0.93	0.93
Max	0.91	0.96	0.97

2.3.3. Leaf Tracking

The HOTA metric was evaluated per individual, which achieved a mean score of 0.84603 ± 0.0623 for nine plants evaluated through 57 timepoints each. The lowest HOTA achieved was 0.7452 and the highest 0.9879. A total of 162 out of 191 leaves were tracked from beginning to end without major issues detected. This accounts for 84.8% of the tracking evaluation dataset (Figure 2.2).

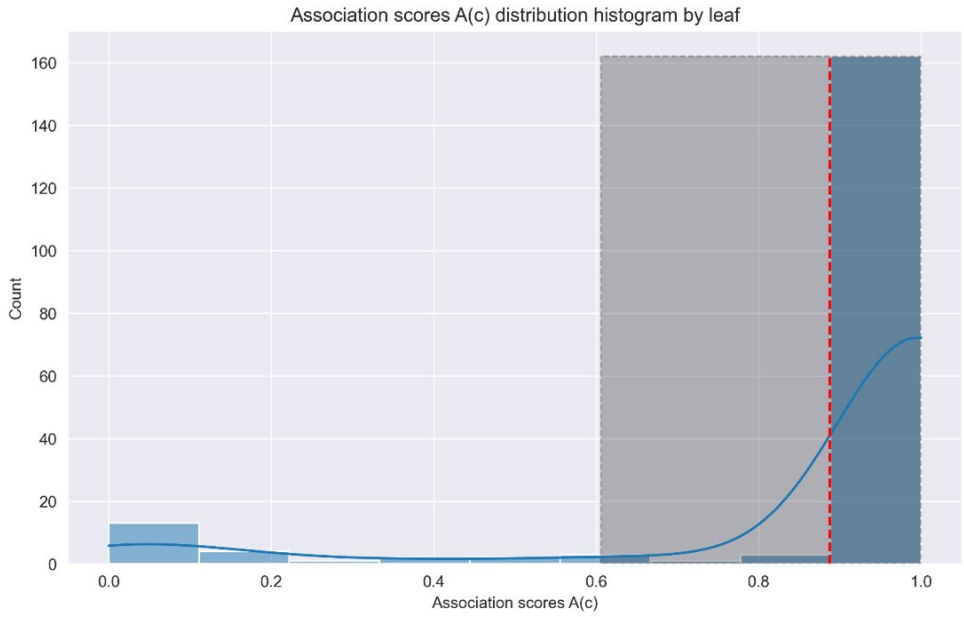


Figure 2.2: Distribution of the Association metric A(c) score counts of the HOTA metric. The red line indicates the mean A(c) score. The blue line indicates the density, and the grey box outlines the STD range of the A(c) score.

2.3.4. Leaf count as indicator of development progression

The total number of rosette leaves over time was retrieved for Col-0 and Ely to evaluate the developmental rate during vegetative growth over the course of the experiment until 26 DAS (Figure 2.3). In the first recorded images on 8 DAS, plants generally had two or three true leaves and two cotyledons. There was no difference between Col-0 and Ely in the leaf count in early days, but a difference became apparent when plants grew older. At the last timepoints, Col-0 had four leaves

more than Ely (p-value close to 0.05, Figure 2.3), which indicates a slightly higher growth rate than Ely. At the time the experiment was finished, Col-0 and Ely were not yet flowering or bolting, which means that the reported leaf numbers are not the final total leaf numbers that these genotypes could make during their vegetative phase.

2.3.5. Photosynthetic capacity at plant level

The photosynthetic capacity at individual plant and leaf level was evaluated as maximum quantum yield (F_v/F_m) and as operating efficiency of photosystem II (ϕPSII). At plant level, Col-0 had a significant better photosynthesis capacity than Ely, with F_v/F_m often 0.04 - 0.05 higher (roughly 5%, Figure 2.4A) and ϕPSII between 0.18 - 0.22 higher (roughly 50%, Figure 2.4B). Both F_v/F_m and ϕPSII of Col-0 and Ely increased over time, while plants grew. The fluctuating light treatments affected both photosynthetic capacity parameters of Ely but not of Col-0. F_v/F_m of Ely increased gradually in days with constant light, but remained more or less constant during fluctuating light treatments (Figure 2.4A).

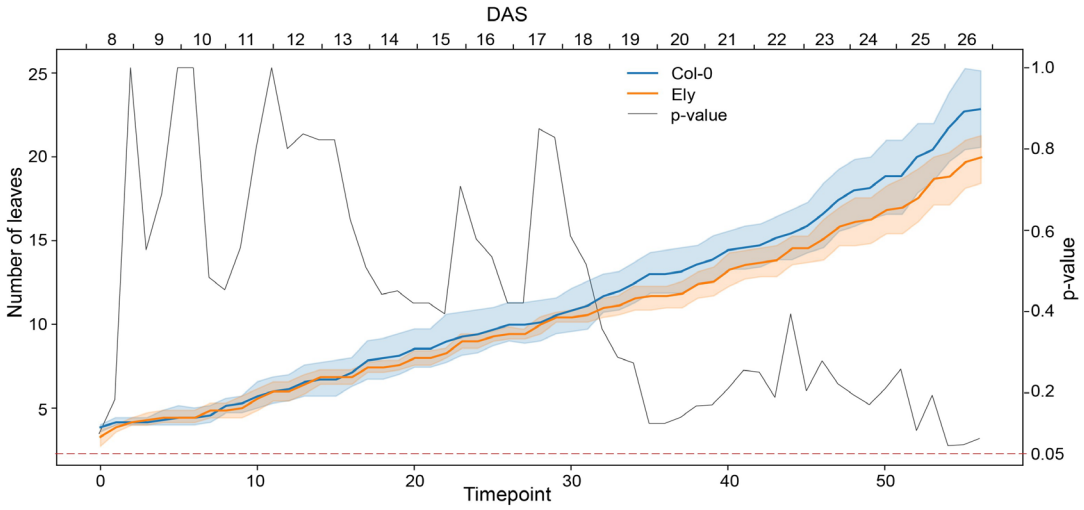


Figure 2.3: Rosette leaf number of *A. thaliana* accessions Col-0 (blue) and Ely (orange) determined over time, either plotted against timepoint or days after sowing, DAS. The minimum and maximum numbers of leaves are shown as shaded regions around the mean of seven replicates (solid line) for each accession. A t-test was performed for every time point to determine the p-values for statistical significance of the difference in leaf number between these two accessions over time (gray solid line).

The diurnal response of ϕPSII in Ely was emphasized during days with fluctuating light treatments, compared to constant light days, which was not or much less obvious for Col-0 (Figure 2.4B).

2.3.6. Photosynthetic capacity at leaf level

A. thaliana plant photosynthesis is expected to change over time depending on plant development. During the vegetative stage, with plants still in the rosette stage, this will be mainly affected by leaf age and leaf development (Bielczynski et al., 2017). Upon identification of

individual leaves, the dynamics of leaf photosynthetic parameters F_v/F_m and ϕPSII over time can be determined, as shown for both Col-0 and Ely in figures 3.5 – 3.9. Leaf photosynthesis generally follows the same trend as observed for whole plant photosynthesis, with a degree of amplitude difference. Leaves 1, 2 and 3 are exceptions (Figures 3.5, 3.6A, 3.7A, 3.8A and 3.9A), likely because they are not properly identified, and represent a mixture of cotyledons and true leaves between replicates. The reason is that in the first input picture, there were four to five leaves detected and randomly assigned. Unexpectedly, F_v/F_m of true leaves had a different pattern than individual plant mean in the Ely accession (Figure 2.8). While F_v/F_m of Col-0 true leaves had a relatively similar pattern compared to that of individual plant mean (Figure 2.6), Ely leaf F_v/F_m kinetic showed that F_v/F_m of a leaf increased sharply (roughly from 0.72 to 0.79), reached its maximum potential when leaves were around a week old and then decreased. As a result, there is substantial variation for F_v/F_m between leaves in Ely at any given moment of time.

2.3.7. Photosynthetic response is influenced by leaf age and order

Leaf tracking also allows evaluation of the influence of leaf age/order on ϕPSII over time and in response to fluctuating light treatments. Figure 2.5 shows the leaf ϕPSII measured in the morning and afternoon at 22 DAS for Col-0 and Ely. Between these measurements, plants were exposed to fluctuating irradiance between 100 and 900 $\mu mol * m^{-2}s^{-1}$ for 5 hours, with duration of each irradiance level taking 60 min. The ϕPSII response is determined as the change in ϕPSII

before and after the fluctuating irradiance treatment. At individual plant level, the ϕPSII response was small, 0.0025 for Col-0 and 0.01 for Ely (Figure 2.4B). However, at the leaf level, younger leaves showed a much larger ϕPSII response than did older leaves (Figure 2.5). The ϕPSII response of each individual plant was therefore mainly determined by the response of the younger leaves.

2.4. Discussion

The aim of this work is to provide both a pipeline and an implementation to perform leaf tracking on top-view images of plants in time series. In this case, the model has been trained in a model organism well known by the plant science community, *Arabidopsis thaliana*, however this pipeline can be applied to any other plant species as long as they can be properly described in a top-view image. Leaf detection and segmentation by itself, is a task that has been performed through different approaches and in different contexts and challenges (Scharr et al., 2016). It is clear from the multitude of approaches that this problem can be tackled in different ways. E.g. edge classification has been used to improve leaf segmentation (Bell & Dee, 2019), but also a combined statistical graph-based method has been used (Kumar & Domnic, 2019) Recent advances in the computer vision field allowed us to implement a novel, more direct approach using convolutional neural networks. The use of mask R-CNN is extensive, and it achieves great results in multiple tasks, with different kinds of objects (Bharati & Pramanik, 2020). Numerous technical alternatives exist for detection and segmentation; however, these alternatives typically specialize in either instance detection or

segmentation, but not both. Mask R-CNN stands out as it is well-documented and can be easily fine-tuned using PyTorch, enhancing its accessibility to the community. Thus, it was a suitable choice for our case. In this case, the model achieves the best results for the detection of early leaves but scores lower for segmentation at that stage. In contrast, the lower results for detection are on the last stages but then it achieves good results on the segmentation. This might be due to the size of the image, with more difficulty to segment a low-resolution object with small size but less difficulty to detect it as the background contains fewer interfering objects, such as other leaves. On the other hand, it might be more challenging to detect individual objects in a crowded image but easier to segment when the detection is right. One issue we found regarding the detection of larger leaves is a relatively high rate of false positives (26 FP out of 330 detections, at a precision of 0.91). This could be addressed by adding an adaptative confidence threshold depending on the stage of the plant, to better balance the trade-off between precision and recall, or by increasing the training dataset size. Recent work has proposed the tracking of dynamic changes at leaf-specific resolution in *A. thaliana* (Swartz et al., 2021) but the proper individualization of different leaves, without recurring to neural networks, has proven to remain a challenge. Through this method, the real challenge lies more on the tracking side than on the individualization of the different leaves. On the other hand, older work provided a method for tracking (Yin et al., 2014), but the implementation can be difficult for a standard user with medium coding experience. A concept such as the overlap between figures (IoU) can be easily implemented in conjunction with the masks

produced by a model given on the PyTorch framework and results in a very intuitive pipeline for most of the users.

The approach of using a mixture between AI and distance measurement, as shown here, has proven to be efficient in other cases as well, such as cell and particle tracking (Ershov et al., 2022). While there are several technical alternatives to the IoU metric for tracking, such as optical flow or center proximity, our implementation has demonstrated sufficient accuracy for leaf tracking, as evidenced by its high accuracy and HOTA scores. Regarding the limitations of the implementation, most of the issues in association arise from false negatives in the detection process. As mentioned in the methodology, a big enough gap between a mask at time frame n and n+1 may produce an association failure. Consequently, failing to keep track for more than 3-4 frames of time allows for enough leaf growth to detriment the method. On the other hand, improving the detection and segmentation accuracy may result in a better association, which can be easily achieved by expanding the number of examples on the training dataset. The biggest challenge to solve is the effect of leaf movement during day-night transition. In certain *A. thaliana* accessions, the leaf movement of the first true leaves can be so strong that there is no overlap between timeframes, which is likely to result in a total loss of association and a new mask assignation. This issue may be tackled by adding one or two images in the transition period from day to night and vice versa, helping to soften the abrupt changes in position. High precision in detection is crucial for accurate leaf tracking, as consecutive detection failures of the same leaf may result in lost tracking due to plant growth. Acknowledging this, we chose the

HOTA evaluation metric, which penalizes both failures in mask association and errors related to mask confusion and missed detections. This metric enables us to effectively evaluate the method's quality and understand its robustness.

On the possible usages and utility, this method allows for individualized leaf growth tracking, and also allows for projected leaf surface area measurement and, depending on the equipment of the phenotyping platform, in fluorescence and infrared measurements per leaf. This kind of analysis may provide insight on differences between individual plants' leaves or patterns within rosettes, providing new traits for genetic association studies. This method is also relatively inexpensive in terms of data to implement, as it is based on fine tuning instead of full training of the detection model. Removing the need for a full training translates into less labelling and less time spent in preparing a dataset, reducing the data preparation burden that normally comes with the use of convolutional neural networks, and also the computational costs. As for usage recommendations, life sciences researchers who wish to utilize our implementation in their own workflows should note that the provided output includes raw data containing masks that may have lost tracking as well as false positives. Therefore, prior to conducting data analysis, it is recommended to filter out such data to minimize the introduction of background noise. The detection and segmentation model has been trained on FC images produced by the phenotyping hardware, which makes this software ideal for phenotyping platforms using the same or comparable hardware. If working with images from a different origin (i.e., RGB images or infrared), the model should be fine-tuned to

obtain the best results. This can be easily achieved by labeling a small dataset and running the scripts provided in the code repository. In case of applying the model to a different plant species, the mask R-CNN model should be retrained on examples of that species. It is important to consider that this methodology works on plants that can be described as top-down images, meaning that plants with high 3D growth will need a different methodology due to the high level of overlap between leaves.

Last, in this manuscript we aim to encourage further development of new methods for better individualization of leaves per plants, therefore the train, test and validation datasets used to train the model are provided, allowing users to expand the library with different approaches for either the segmentation or the tracking. As a case study, we used the CF images to determine photosynthetic capacity at individual plant and leaf level in the *A. thaliana* accessions Col-0 and Ely over time and under various light treatments. The photosynthetic capacity of Ely is lower than that of Col-0 at individual plant level (Figure 2.4), which is consistent with previous studies (El-Lithy et al., 2005; Flood et al., 2020), which explained this to be due to a mutation in the chloroplast *PsbA* gene, encoding the D1 protein of photosystem II, of Ely. The response to light treatment was also more pronounced in Ely than in Col-0. Our results show that younger leaves have a higher photosynthetic capacity than older leaves (Figure 2.5), which is likely to be due to a higher ability to redesign leaf anatomy and the photosynthesis machinery for acclimation to a fluctuating light environment (Bielczynski et al., 2017). This is the first time that leaf photosynthesis dynamics is evaluated in a nondestructive and

automated manner. Variation for photosynthesis parameters between leaves was observed although the differences are subtle. Nevertheless, these potentially enlarge the variation in photosynthesis parameters observed between genotypes, which is often smaller when determined at individual plant level (Prinzenberg et al., 2020; Rooijen et al., 2017). Variation in leaf level photosynthesis parameters is of great interest for quantitative genetic analysis and may resolve the effect of developmental differences on photosynthesis in genetically segregating populations with variation in plant growth among genotypes.

2.5. Conclusion

In this work a method for individual leaf segmentation and tracking on top-view images is provided, as well as an example of implementation of the method. Although there are many technical alternatives, we find this method to be efficient and easy to use for automated workflows on phenotyping platforms such as the one we used. The advantages provided by the use of convolutional neural networks and recent models are clear on precision at detection and individualization, limiting the challenges to tracking. Using a simple concept such as masks overlap, tracking can be solved to a certain degree of confidence, which can be further improved through expanding the training dataset for leaf detection. The basic requirements in data for fine-tuning the detection model are low, giving users the opportunity to take advantage of pretrained weights with little cost.

2.6. Methods

2.6.1. Plant material and growing conditions

A set of 177 *A. thaliana* accessions was used, which includes 175 accessions of the regional Dutch population (Wijfjes, 2021), and the Ely and Columbia-0 (Col-0) accessions. Seeds of accessions were first pregerminated on filter paper with demi water in petri dishes by stratification treatment for five days, after that they were germinated at $24^{\circ}C$ for 16 hours. The germinated seeds were transferred to rockwool block (4 x 4 cm; GrodanTM) to germinate. From this point onwards, the days of the experiment were referred as days after sowing (DAS). This set was grown in a complete randomized block design, with eight blocks. There was a replicate in each block for every Dutch accession, and four replicates were available for Col and Ely, in total 1440 individuals.

Plants were supplied with standard Hyponex nutrient solution (Prinzenberg et al., 2020). The climate room was set to $20^{\circ}C$ and $18^{\circ}C$, day and night temperature, respectively. Photoperiod was set to 12 hours (light on from 8 am to 8 pm), in which light intensity increased and decreased gradually in the first and last hour of the photoperiod to resemble dawn and dusk of the day. Irradiance stayed at 300 $\mu mol{*}m^{-2}s^{-1}$ until application of light treatment for five hours (from 11 am to 4 pm), then returned to 300 $\mu mol{*}m^{-2}s^{-1}$. There were three light treatments: constant irradiance at 300 $\mu mol{*}m^{-2}s^{-1}$, fast fluctuation and slow fluctuation between 900 and 100 $\mu mol * m^{-2}s^{-1}$. The duration of each irradiance level taking 15 and 60 min,

respectively, in the fast and slow light fluctuation treatment. photosynthesis, CF measurements were taken once per day in the night.

2.6.2. Phenotyping and data collection

Photosynthetic parameters were measured based on chlorophyll fluorescence (CF) imaging every day from 8 DAS for 19 days with the Phenovator (PlantScreen Robotics XY system, Photon System Instruments^{TM}). Re-emitted light from photosystem II (PSII) in form of fluorescence is measured and compared whilst exposing plant to a combination of actinic lights (lights that drives photochemistry and photosynthesis), darkness and series of saturating pulses. This method is known as pulse-amplitude modulated fluorescence analysis, and provides non-invasive assessment of PSII efficiency to pass electrons to photosynthetic apparatus (Baker, 2008; Murchie & Lawson, 2013)

In dark-adapted state, a minimum value of CF (F_0) is measured by very low measuring light, which can only result in minimal level of emitted CF but not electron transport induction. Application of saturating pulses in dark-adapted state, driving closure of all reaction centers, induces maximum level of CF (F_m). The difference between Fm and F0 is variable fluorescence, termed F_v, from which F_v/F_m parameter can be calculated, which is used as robust indicator of maximum quantum yield of PSII photochemistry.

Similarly CF is measured at light-adapted state using actinic light, which results in F_p. Saturating pulses can be applied during the actinic

light exposure, thus F_{mp} is obtained. The difference between F_{mp} and F_p in proportion to F_{mp} is used to determine the operating efficiency of PSII photochemistry (ϕPSII). There were three CF imaging a day, consisting of one for dark-adapted maximum quantum yield of PSII (F_v/F_m) and two for light-adapted PSII efficiency at 9.45 am and 5.15 pm. The dark-adapted measurement was performed in the first hour of the day (12 am), thus in total 19 measurements are available. The light-adapted measurements, were made in the morning (9.15 am) and afternoon (5.15 pm) during the photoperiod before and after the light treatment, which led to 38 measurements. Thus for the whole course of the experiment, there were a total of 57 CF measurements. Data collected from the CF imaging was used to produce grayscale images by normalization in a range from 0 to 1 of the data matrices.

2.6.3. Training and Validation datasets for detection and segmentation, and tracking evaluation dataset.

For the training, we used an integrated dataset of *A. thaliana* plants at different stages of development ranging from the display of the first true leaves to the largest size on day 26 after sowing. This dataset was composed of 166 individual plant images with a total of 1773 leaves, from which 149 images were used as subset for training and 17 for testing. A different dataset, composed of 18 individuals in three different developmental stages, was used for validation. The validation dataset was grouped in three categories, early stage (17 images, one removed for not displaying any leaves yet), mid stage (18

images) and late stage (18 images), adding to a total of 53 images (523 leaves) of the same 18 individuals.

To estimate the efficiency of the tracking we used a separate dataset of 9 plants through 57 time points, with a total of 204 leaves. This dataset was never used, neither in the training-testing nor in the validation of the Mask R-CNN model for leaf detection and segmentation.

2.6.4. Detection and segmentation

Detection and segmentation have been performed through a convolutional neural network-based approach by selecting a mask R-CNN model [14] which can tackle the objectives of detection and segmentation in an efficient way. For the feature extraction, we employed a resnet50 backend (He et al., 2016) and fine-tuned the full model parting from pre-trained weights on the MSCOCO dataset (Lin et al., 2014) with a number of classes for classification set to two (leaf and background). The training was performed using stochastic gradient descent (SGD), with a learning rate 0.005 over ten epochs. The original dataset and more details on the training can be found in the code repository as well as notebooks to replicate it. To evaluate the model, we scored the validation dataset in two different contrasts. The initial comparison focused on precision and recall for leaf detection, assessing the model ability to accurately detect leaves while minimizing false positives. We discarded detections with a confidence below 0.9 and counted the number of correct detections against the total number of false positives and false negatives. A correct detection was only assigned if the IoU between the model prediction mask and

the ground truth mask was at least of 0.15. If a predicted mask was found below this threshold and no ground truth mask could be assigned, then it was detected as false positive. Conversely, if a ground truth mask was missing a suitable predicted mask, it was labelled as a false negative.

The second contrast was the measurement of the IoU (Intersection over union) between the ground truth and the predicted masks. Based on the same criterion, we found the pairing mask by discarding detections below the 0.9 confidence and by getting the best overlapping mask for each leaf. In this test, false positives and negatives are disregarded as the aim is to measure the segmentation performance in case of a true positive. Both contrasts were applied and evaluated separately on each of the categories defined in the validation dataset and their respective metrics were collected for each of them.

2.6.5. Instance tracking

The approach used for leaf tracking was based on overlapping masks between timepoints (Figure 2.10). As most of the leaf changes in *A. thaliana* development can be traced on two dimensions, masks between timepoint n and n+1 will be largely overlapping if the timestep or the variance between instances are small enough. In this case the largest timeframe is roughly 8 hours, which represents a small enough gap between different developmental stages of the plant. At each time point, the IoU between a query mask and all the target masks in the next time point was estimated. The target mask with the highest IoU and above a certain threshold (0.15 in this study) was

assigned as the respective pair. We then generated a dictionary containing the query-target relation and appended it to a list, which will keep the order. Thereafter, the target was removed from the target pool. If the query was not found within the targets, we created a pair dictionary with itself and appended it to the list. After looping over all the queries, the remaining targets were assumed to be new leaves, which were appended to the same list as dictionaries containing empty queries and the mask as a target. Once all the pairs had been assigned, we collected all the targets from the list preserving the order and created a new query dataset for the next time step. This process was repeated until we reached the last time step (Figure 2.10).

To evaluate the quality of the tracking we used the HOTA metric (Luiten et al., 2021) and estimated the total leaves that were tracked from the first to the last time point (Last frame it could be detected) without missing the association of the masks and with a maximum of five missing detections for that leaf. This way we generated two categories, right and miss. Thirteen leaves from the tracking evaluation dataset were excluded from both metrics due to limited time points (1 to 3 time points) or small detection errors (very low area), resulting in a total of 191 leaves considered for the evaluation of the metrics.

2.6.6. Leaf counting and photosynthesis capacity analysis of two different genotypes

The two accessions Col-0 and Ely were selected as genotypes for a comparative experiment. For each genotype, seven replicates were

considered. To retrieve the total number of rosette leaves per time point the ϕPSII file produced by the system was used. A summary of the number of instances tracked at each time point was sufficient to perform the estimation. With the leaf number per replicate obtained, a t-test was performed at each time point to evaluate the significance of the differences between genotypes. The estimation of F_v/F_m and ϕPSII per individual plant and leaf was performed on the raw data based on pixel means. Instance segmentation masks were applied to the raw matrix (fimg produced by the phenotyping system) to extract the raw values for the region and mean values were estimated per instance.

2.6.7. Metrics definition

Precision is defined as the ratio between true positives detected against the total detections retrieved by the model.

$$Precision = \frac{TruePositives}{True\ positives\ (TP) + False\ positives\ (FP)} \quad (1)$$

Recall measures the sensitivity of a model taking in consideration True positives against the ground truth detections.

$$Recall = \frac{TruePositives}{True\ positives\ (TP) + False\ Negatives\ (FN)} \quad (2)$$

F-score is the harmonic mean of precision and recall.

$$F - score = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

Intersection over Union of two figures (A and B) is defined as the symmetric difference of figures A and B against the union of those figures.

$$Intersection\ Over\ Union\ (IoU) = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

The Higher Order Tracking Accuracy (HOTA) metric is defined as in the published document (Luiten et al., 2021). This measure joins detection and association scores into a single value.

$$Higher\ Order\ Tracking\ Accuracy\ (HOTA) = \sqrt{\frac{\sum_{c \in \{TP\}} A(c)}{|TP| + |FN| + |FP|}} \quad (5)$$

$$A(c) = \frac{|TPA(c)|}{|TPA(c)| + |FNA(c)| + |FPA(c)|} \quad (6)$$

2.7. Availability of data and materials

All the code in this project was developed using python and different public libraries. The main libraries used in this project were PyTorch (Paszke et al., 2019) for the neural networks, PIL and OpenCV (Bradski, 2000) for the image manipulation, Pandas (McKinney, 2010) and NumPy (Harris et al., 2020) for data analysis, and scikit-learn (Pedregosa et al., 2011) for the IoU . Part of the PyTorch-Vision GitHub repository was used for the model training, specifically the reference scripts found in the folder detection. These scripts are included in the project GitHub under the modelTraining folder without relevant modifications. The full library of this project is available as a public repository at GitHub:

https://github.com/Fedjurrui/Leaf-Tracking.

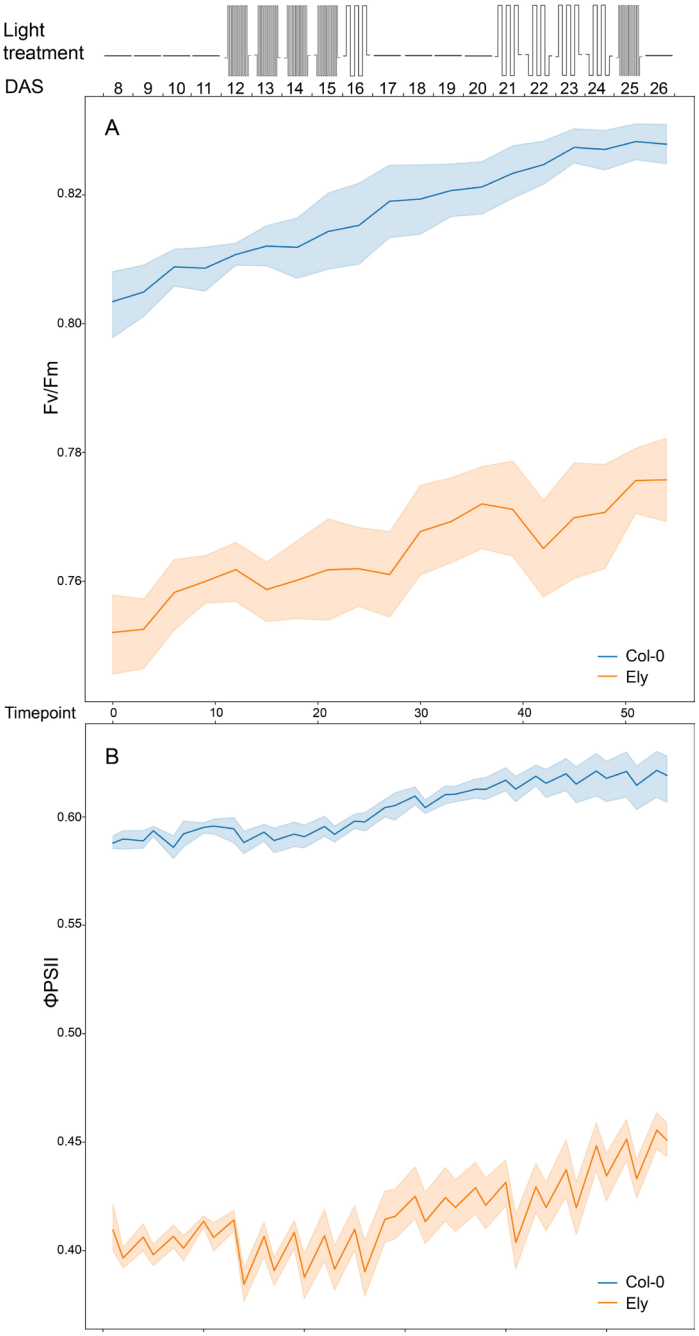


Figure 2.4: Photosynthesis dynamics of *A. thaliana* accessions Col-0 (blue) and Ely (orange) at individual plant level. Schematic representation of the fluctuating light treatment over time is shown in

the top panel, with 300 $\phi\mu mol * m^{-2}s^{-1}$ of constant irradiance, 100 $\phi\mu mol * m^{-2}s^{-1}$ of low irradiance and 900 $\phi\mu mol*m^{-2}s^{-1}$ of high irradiance levels. Maximum quantum yield (F_v/F_m, A) and photosystem II efficiency (ϕPSII, B) are plotted against time (as timepoint or as days after sowing, DAS). The variation for these traits is shown as shaded regions around the mean of 8 replicates (solid line) for each accession.

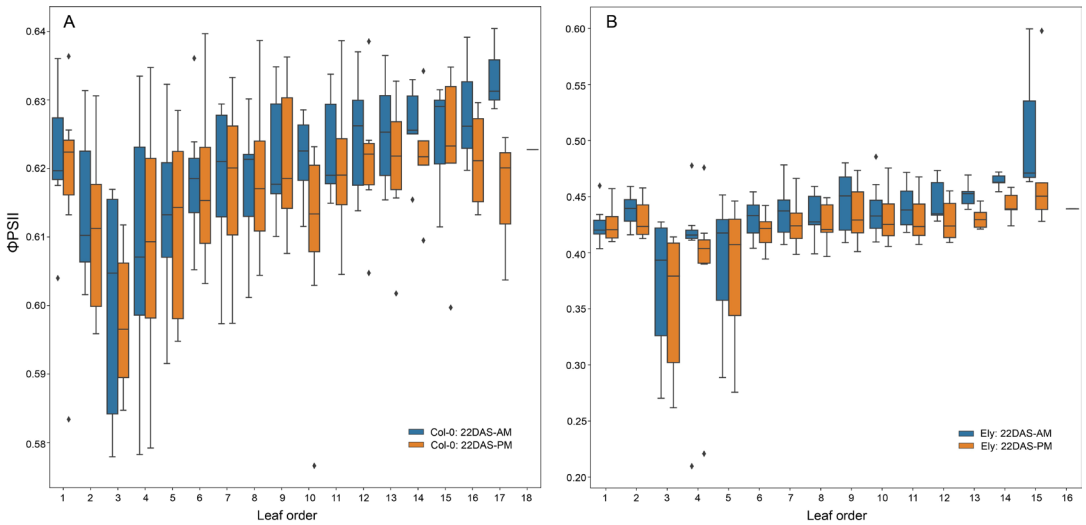


Figure 2.5: Photosystem II efficiencies (ϕPSII) of individual leaves of *A. thaliana* accessions Col0 (A) and Ely (B) before and after a fluctuating light treatment. Boxplots represent the ϕPSII measurements of individual leaves of 8 replicate plants at day 22 after sowing (22 DAS) in both morning before the fluctuating light treatment (AM in cyan) and in the afternoon (PM in orange). Leaf order (x-axis) was obtained according to the order of emergence and detection over the time, meaning the higher leaf order the younger leaf and vice versa.

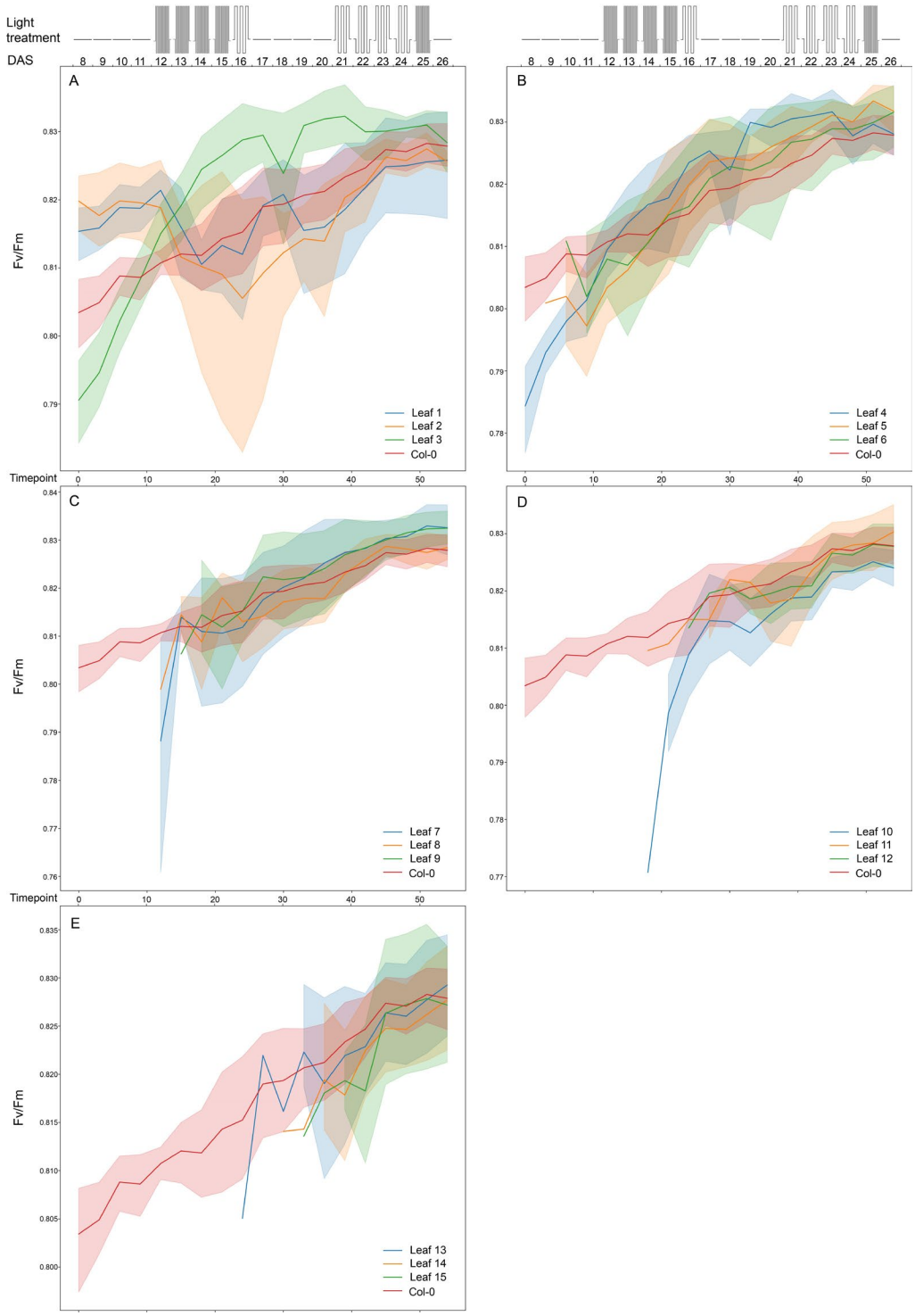

Light treatment
DAS
8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
A
B
C
D
E
Fv/Fm
Timepoint
Leaf 1
Leaf 2
Leaf 3
Col-0
Leaf 4
Leaf 5
Leaf 6
Leaf 7
Leaf 8
Leaf 9
Leaf 10
Leaf 11
Leaf 12
Leaf 13
Leaf 14
Leaf 15

Figure 2.6: Maximum quantum yield (F_v/F_m) of whole Col-0 rosettes and 3 indicated individual leaves per panel (leaf number indicated: (A) for leaf 1 to 3, (B) for leaf 4 to 6, (C) for leaf 7 to 9, (D) for leaf 10 to 12 and (E) for leaf 13 to 15) determined over time (as timepoint or as days after 18 sowing, DAS). Schematic representation of the fluctuating light treatment over time is shown in the top panel, with 300 $\phi\mu mol * m^{-2}s^{-1}$ of constant irradiance, 100 $\phi\mu mol * m^{-2}s^{-1}$ of low irradiance and 900 $\phi\mu mol * m^{-2}s^{-1}$ of high irradiance levels . The mean F_v/F_m values are shown with solid lines, with shaded regions above and below indicating standard errors of the means. Data points represent 8 replicates.

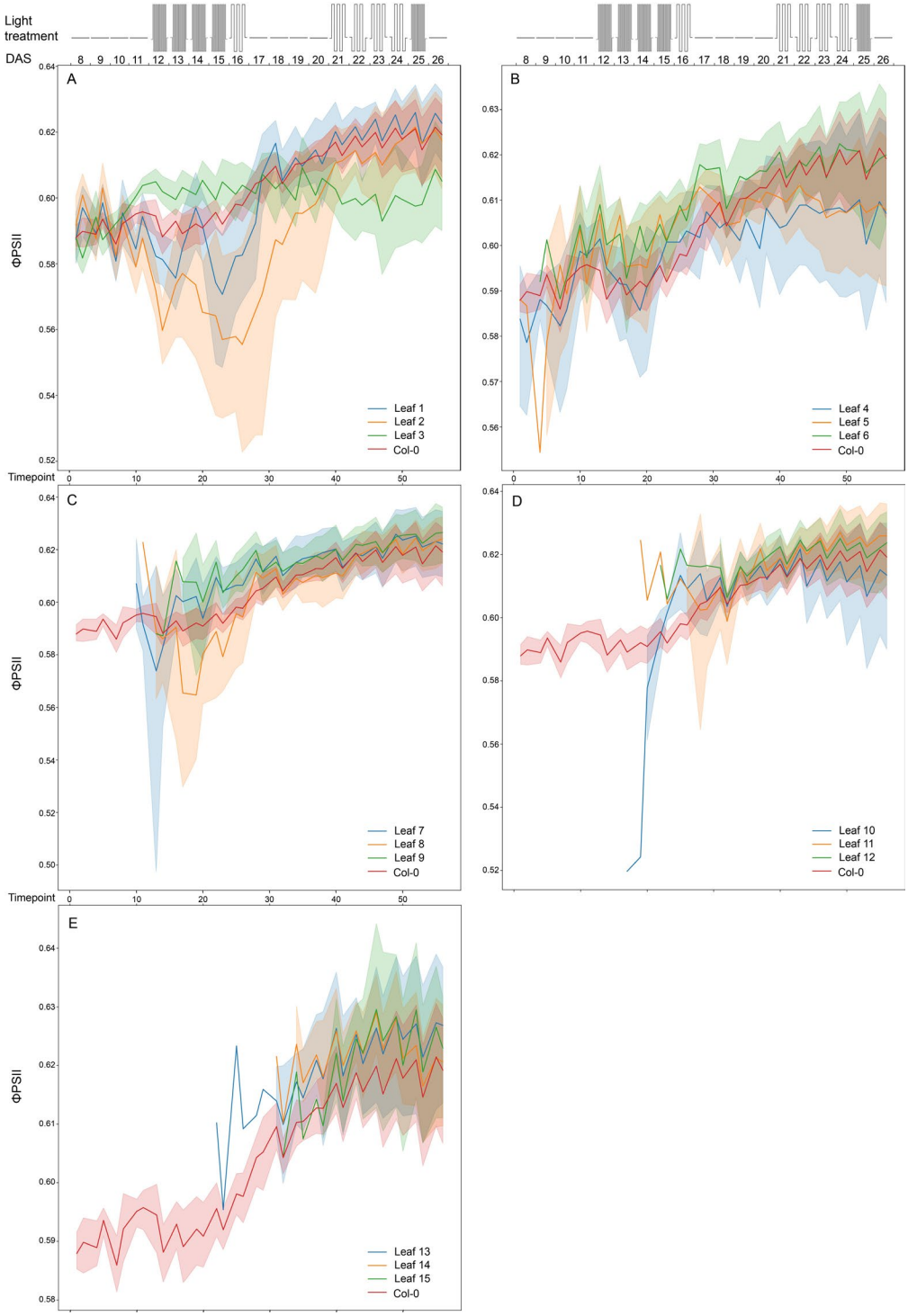

Light treatment
DAS
8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
A
B
C
D
E
ΦPSII
Timepoint
Leaf 1
Leaf 2
Leaf 3
Col-0
Leaf 4
Leaf 5
Leaf 6
Leaf 7
Leaf 8
Leaf 9
Leaf 10
Leaf 11
Leaf 12
Leaf 13
Leaf 14
Leaf 15

Figure 2.7: Photosystem II efficiency (φPSII) of whole Col-0 rosettes and 3 indicated individual leaves per panel (leaf number indicated: (A) for leaf 1 to 3, (B) for leaf 4 to 6, (C) for leaf 7 to 9, (D) for leaf 10 to 12 and (E) for leaf 13 to 15) determined over time (as timepoint or as days after sowing, DAS). Schematic representation of the fluctuating light treatment over time is shown in the top panel, with 300 φμmol ∗ m−2s−1 of constant irradiance, 100 φμmol ∗ m−2s−1 of low irradiance and 900 φμmol ∗ m−2s−1 of high irradiance levels . The mean Fv /Fm values are shown with solid lines, with shaded regions above and below indicating standard errors of the means. Data points represent 8 replicates

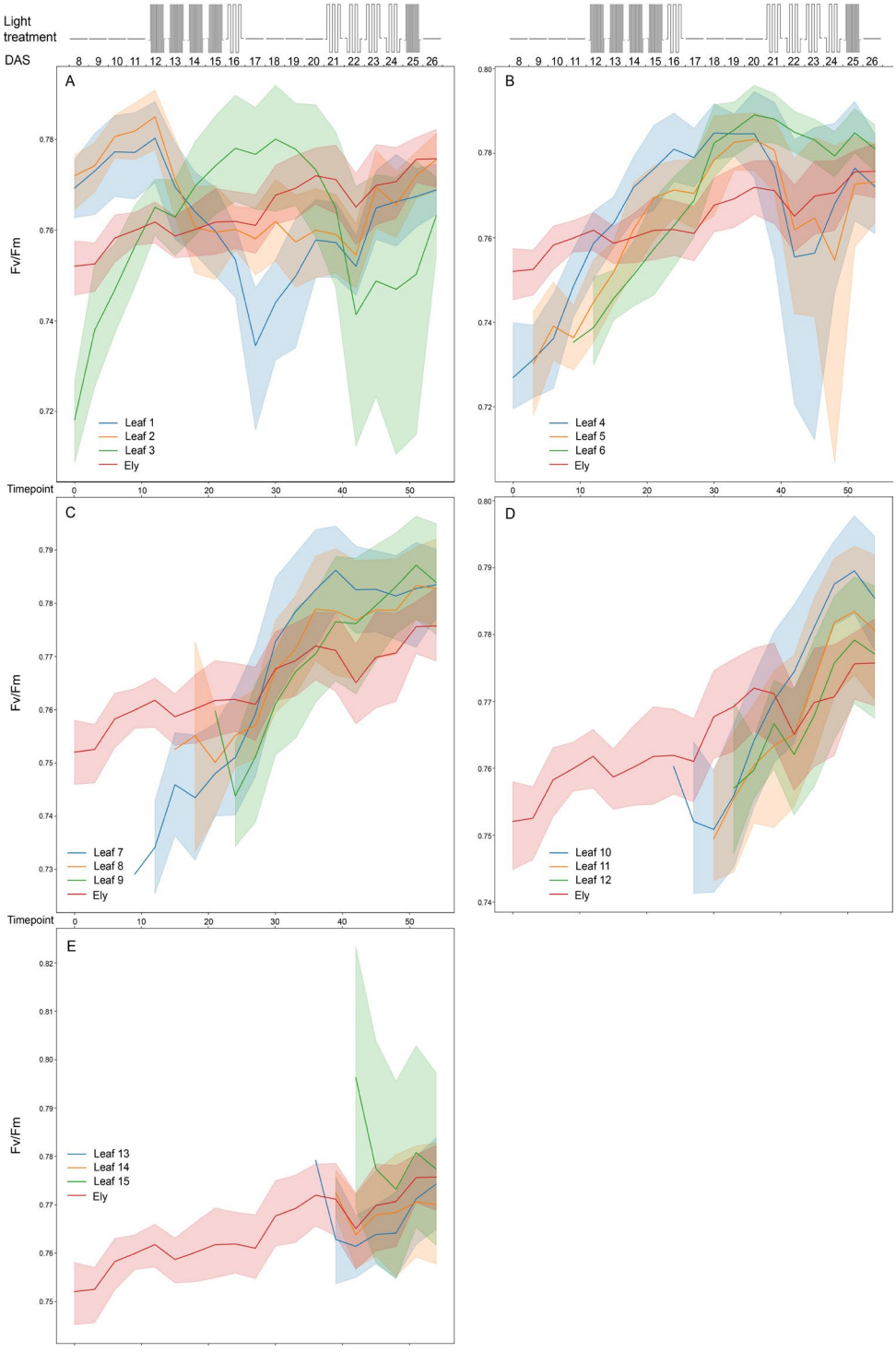

Light treatment
DAS
8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
A
B
C
D
E
Fv/Fm
Timepoint
0 10 20 30 40 50
Leaf 1
Leaf 2
Leaf 3
Ely
Leaf 4
Leaf 5
Leaf 6
Leaf 7
Leaf 8
Leaf 9
Leaf 10
Leaf 11
Leaf 12
Leaf 13
Leaf 14
Leaf 15

Figure 2.8: Maximum quantum yield (Fv /Fm) of whole Ely rosettes and 3 indicated individual leaves per panel (leaf number indicated: (A) for leaf 1 to 3, (B) for leaf 4 to 6, (C) for leaf 7 to 9, (D) for leaf 10 to 12 and (E) for leaf 13 to 15) determined over time (as timepoint or as days after sowing, DAS). Schematic representation of the fluctuating light treatment over time is shown in the top panel, with 300 φμmol ∗ m−2s−1 of constant irradiance, 100 φμmol ∗ m−2s−1 of low irradiance and 900 φμmol ∗ m−2s−1 of high irradiance levels. The mean Fv /Fm values are shown with solid lines, with shaded regions above and below indicating standard errors of the means. Data points represent 8 replicates.

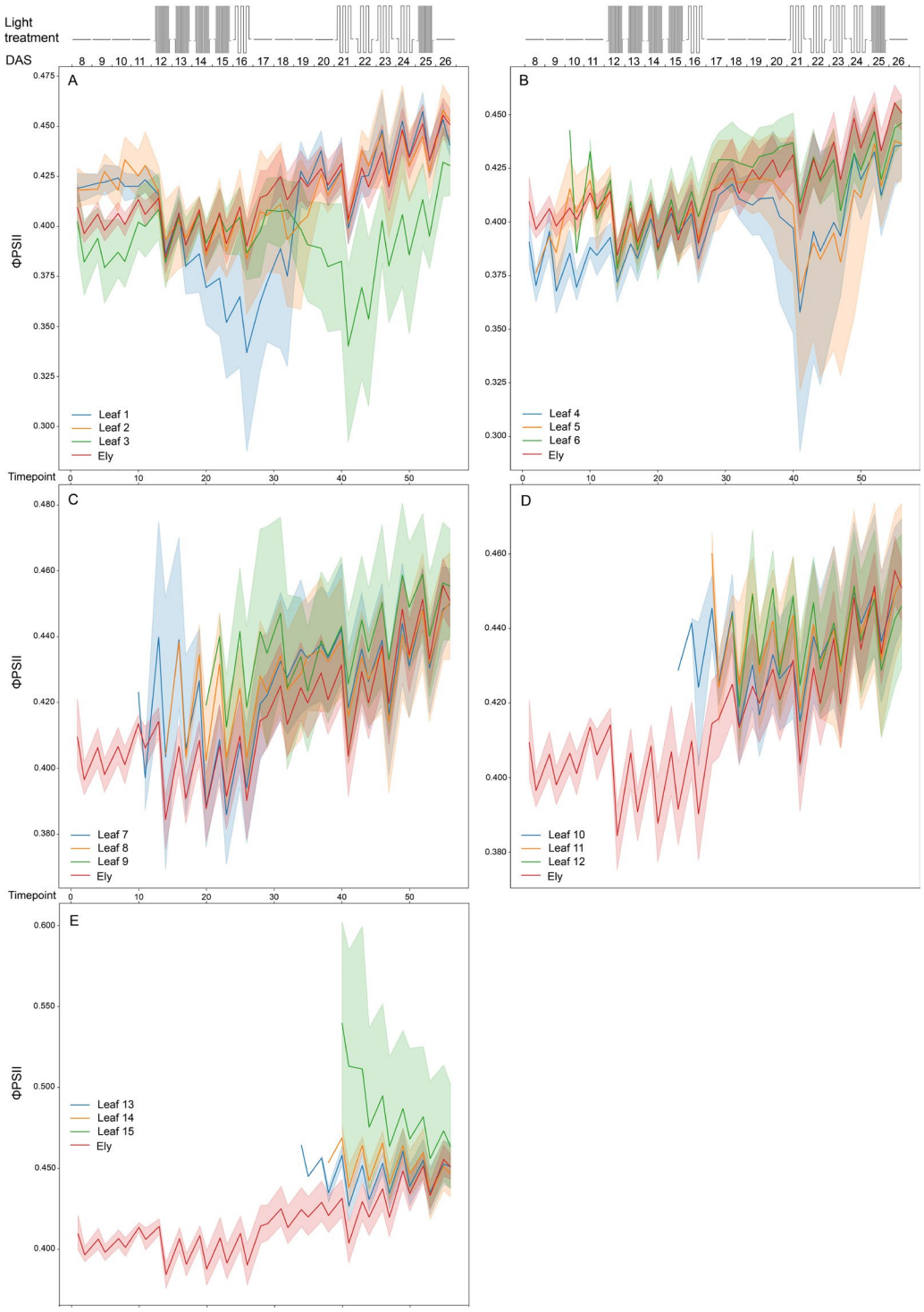

Light treatment
DAS
8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
A
ΦPSII
Leaf 1
Leaf 2
Leaf 3
Ely
Timepoint
B
Leaf 4
Leaf 5
Leaf 6
Ely
C
Leaf 7
Leaf 8
Leaf 9
Ely
D
Leaf 10
Leaf 11
Leaf 12
Ely
E
Leaf 13
Leaf 14
Leaf 15
Ely

Figure 2.9: Photosystem II efficiency (φPSII) of whole Ely rosettes and 3 indicated individual leaves per panel (leaf number indicated: (A) for leaf 1 to 3, (B) for leaf 4 to 6, (C) for leaf 7 to 9, (D) for leaf 10 to 12 and (E) for leaf 13 to 15) determined over time (as timepoint or as days after sowing, DAS). Schematic representation of the fluctuating light treatment over time is shown in the top panel, with 300 φμmol ∗ m−2s−1 of constant irradiance, 100 φμmol ∗ m−2s−1 of low irradiance and 900 φμmol ∗ m−2s−1 of high irradiance levels . The mean Fv /Fm values are shown with solid lines, with shaded regions above and below indicating standard errors of the means. Data points represent 8 replicates.

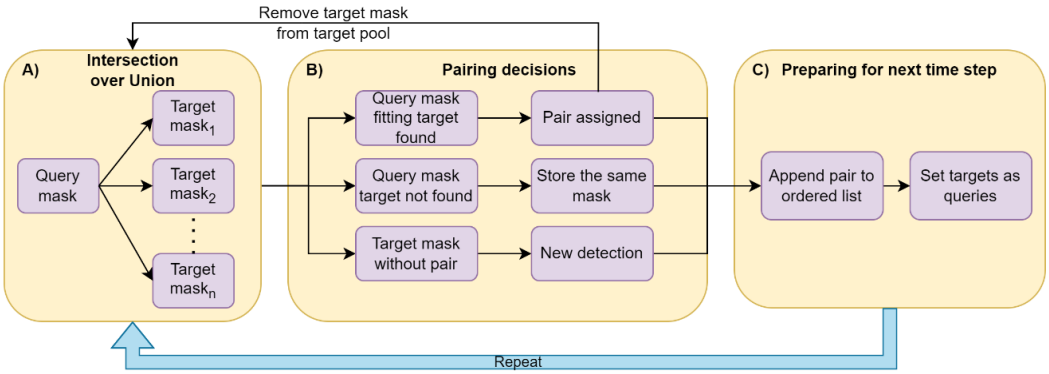


Figure 2.10: Flowchart illustrating the tracking process. Panel A depicts the initial step, where the mask for time step n (query) is retrieved and the IoU is computed for all masks at time step n+1 (targets). In Panel B, the decision loop is made based on the resulting IoU. If the query and target match, the pair is assigned, and the target is removed from the target pool. If the query is not found, indicating that the leaf is either covered in the image or the detection failed, the position is stored by saving the same mask. Finally, the remaining target masks are considered as new detections. Panel C shows the list of masks, in order, which is stored and set as the new queries for the next time step.

Chapter 3: GenoDrawing: An autoencoder framework for image prediction from SNP markers.

Published in: Plant Phenomics.

Authors: Federico Jurado-Ruiz, David Rousseau, Juan A. botia, Maria José Aranzana.

Doi: 10.34133/plantphenomics.0113

3.1. Abstract

Advancements in genome sequencing have facilitated whole genome characterization of numerous plant species, providing an abundance of genotypic data for genomic analysis. Genomic selection and neural networks (NN), particularly deep learning, have been developed to predict complex traits from dense genotypic data. Autoencoders, a NN model to extract features from images in an unsupervised manner, has proven to be useful for plant phenotyping. This study introduces an autoencoder framework, GenoDrawing, for predicting and retrieving apple images from a low-depth single nucleotide polymorphism (SNP) array, potentially useful in predicting traits that are difficult to define. GenoDrawing demonstrates proficiency in its task using a small dataset of shape-related SNPs. Results indicate that the use of SNPs associated with visual traits has significant impact on the generated images, consistent with biological interpretation. While using significant SNPs is crucial, incorporating additional, unrelated SNPs results in performance degradation for simple NN architectures that cannot easily identify the most important inputs. The proposed GenoDrawing method is a practical framework for exploring genomic prediction in fruit tree phenotyping, particularly beneficial for small to medium breeding companies to predict economically significant heritable traits. Although GenoDrawing has limitations, it sets the groundwork for future research in image prediction from genomic markers. Future studies should focus on using stronger models for image reproduction, SNP information extraction, and dataset balance in terms of phenotypes for more precise outcomes.

3.2. Introduction

Advances in high throughput genome sequencing and genotyping methods have brought to reality the whole genome characterization of many plant species, including models and crops, as well as the analysis of diversity at population level. Under this scenario of huge amounts of genotypic data, the link between genotype and phenotype in crops has the tremendous potential to identify genes or genome regions involved in the natural variation of relevant agricultural traits, as well as to predict the performance of offsprings in specific environments. For traits controlled by major genes or moderate to strong Quantitative Trait Loci (QTLs), linkage mapping in biparental families or genome-wide association analysis (GWAS) in germplasm collections allow the development of markers that can, eventually, be used for marker-assisted breeding (MAB). By contrast, complex quantitative traits regulated by multiple QTLs with minor effects are difficult to predict with few markers. Genomic selection (GS) models have been developed to overcome this limitation (Bhat et al., 2016; Iwata et al., 2015; H. Zhang et al., 2019; Zhao et al., 2014). In short, they predict complex traits from dense genotypic data and a set of qualitative, continuous, or discrete measurable trait descriptors in a training population to ultimately estimate the performance of the individuals of a breeding population from their genomic profile. More recently, neural networks (NN) are being suggested as a powerful tool for genomic prediction that may surpass some challenges associated to the classical GS models, such as the requirement of assuming data distributions of the linear models, or the requirement of priors' specifications in Bayesian models (Pérez-Enciso & Zingaretti, 2019).

Neural networks and deep learning (DL) are subfields of artificial intelligence that use multiple interconnected nodes, called artificial neurons, to process information and make predictions. These networks can have multiple layers, allowing for complex and non-linear analysis of data. Deep learning algorithms have demonstrated improved accuracy and speed compared to other artificial intelligence-based methods (Voulodimos et al., 2018). Deep neural networks (DNN) have been already applied for genomic prediction in several fields, using real (Papa et al., 2019; van Klompenburg et al., 2020) and simulated (Abdollahi-Arpanahi et al., 2020) SNP marker data, stimulated by the growing availability of easy-to-use DNN frameworks such as PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2015). Within the DL field of study, generative models (i.e. unsupervised models that learn from patterns to finally generate new examples that could have been derived from the original dataset) encompass a vast and complex area (Harshvardhan et al., 2020) and are living a new golden age after the publication of the generative adversarial networks (GANs) where one neural network compete to generate new synthetic data that can fool a discriminator network [13]. Even further, the recent success of stable diffusion (Rombach et al., 2021), and other models that can generate detailed images from text descriptions, are pushing the known limits of the generative networks field. These advancements forecast an unexplored landscape of possibilities for innovative applications and improvements.

Autoencoders are a well-known neural network model from the generative model family successfully applied in image processing

specially to reconstruct images. They are mainly used in different fields for nonlinear dimensionality reduction and automatic feature learning (Pinaya et al., 2020). Combined with convolutional neural networks (CNN), they can extract features from images in an unsupervised manner, preventing the descriptors bias introduced by subjective decisions. This approach has proven useful to learn complex features in a large variety of fields where images have a prominent role, such as animation (Mourot et al., 2022), neurobiology (Richards et al., 2019), and medical imagery (Song et al., 2021). In agriculture, images are increasingly used for plant and crop phenotyping (Jangra et al., 2021; Mochida et al., 2018), providing relevant information to undertake genomic analysis. For example, Wang et al. (X. Wang et al., 2019) applied DL methods in images obtained by mobile vehicles to extract high-throughput phenotypic information have served to study the genetic architecture of flowering time in wheat. Such a strategy is particularly useful for traits that are difficult to describe by other means. It is worth mentioning that most of the important agronomic traits in crops are complex and vaguely described by simple descriptors, for they can be dissected into a complex combination of characteristics. This is the case of fruit quality, which embraces visual (fruit color, shape, symmetry), organoleptic (taste, flavor, texture), and sensory (firmness) aspects. Therefore, acquiring relevant data challenges genomic studies. In the case of traits related to fruit or plant appearance, the direct use of images could defeat this difficulty (Sarić et al., 2022). However, the difficulty of translating the visual perception provided by images into data or objective parameters, or into measurable traits, challenges their use in genomic studies. Thus, feeding whole images to CNN models appears

as a good opportunity. Here we present an autoencoder framework to predict fruit shape from known molecular data (SNPs) retrieving, as output, the predicted fruit shape as an image.

The objective of this work is, rather than providing a ready-to-use model, to demonstrate that neural networks can be used to predict and retrieve image-based phenotypes from a small set of SNPs. While we have applied them here to retrieve apple images with an SNP-predicted shape, this approach could be used to predict other genetically controlled visual phenotypes in other organs and species.

3.3. Materials and Methods

3.3.1. General Approach

The general approach (Figure 3.1) consisted of training two models: 1) an image compressor model (autoencoder), trained with images of apple sections, and 2) an embedding predictor, a model that predicts the mean compressed values, also known as embeddings, of a genotype based on its single nucleotide polymorphisms (SNPs). Both models were then merged into a SNPs-to-Image model which we named GenoDrawing.

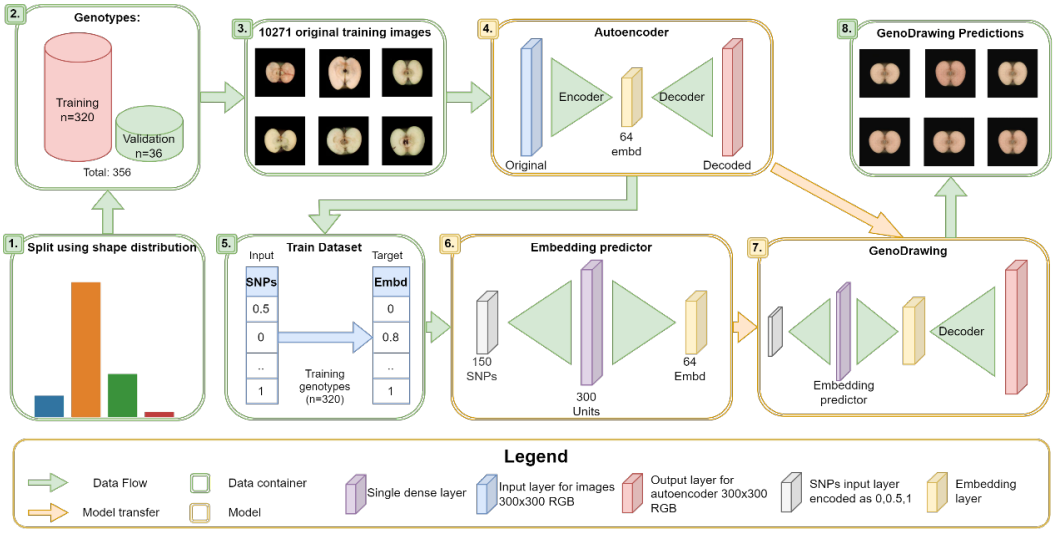


Figure 3.1: Graphical abstract of the general approach. (1) The dataset of apple genotypes was split into two subsets, the training and the validation, preserving in each the overall fruit shape distribution. (2) The SNP genotypic matrix was split into two, one per dataset. (3,4) Images from the training dataset were used to fit the autoencoder model. (5,6) Encoded values for each image (with 64 embeddings) were averaged by genotype and used together with the SNP matrix to train the embedding predictor model. (7) The resulting GenoDrawing is not a trained model but an assembly of both, autoencoders decoder, and embedding predictor. (8) Example of images predicted by GenoDrawing .

3.3.2. Plant materials, image acquisition and processing

In this study we used the apple images generated by Dujak et al. (Dujak et al., 2023). Shortly, the images contained from six to ten apple halves from 356 genotypes of the apple reference collection (Apple RefPop) (Jung et al., 2020) grown at the IRTA experimental field of Gimenells (Lleida, Spain). The apples were collected in 2019 (247 genotypes) and 2020 (the 247 genotypes of 2019 plus 109) at ripening stage. Single apple images were segmented using a fine-tuned version of Mask-RCNN (He et al., 2017) into individual images of size 300x300 RGB pixels. Apples were classified into five categories depending on their external fruit shape index (FSI): flat, flat-globose, globose, oval, and oblong (Lateur et al., 2022). FSI is defined as the ratio between the height and the width of the apple.

3.3.3. Genotypic data

Genotypic data consisted of a matrix of 303,239 SNPs from the 356 apple genotypes extracted from data produced by Jung et al. (Jung et al., 2020) and prepared for the machine learning process by encoding the dataset to a numerical scale (0, 0.5, 1 representing AA, Aa, and aa alleles, respectively).

To train the predictor model, two SNP datasets were extracted from the matrix: one of 150 size and shape-related SNPs (tSNPs) from literature (Table S1), and the one of 150 randomly selected SNPs (rSNPs) using NumPy library (Harris et al., 2020). The random selection of SNPs was renewed every time a new model was trained.

3.3.4. Neural networks models: autoencoder and embedding predictor

The neural networks were developed using PyTorch (Paszke et al., 2019). First, a convolutional autoencoder network (Goodfellow et al., 2016) was created, which reduced the number of the image parameters while keeping the relevant features related to the edges and shapes of the apple sections. The architecture used for this convolutional autoencoder network consisted of six convolutions with kernel size of 3x3, set the stride to two and used a rectified linear activation function (ReLu). The number of filters were progressively increased (from 16 to 128) through the multiple encoder layers. The architecture finished with three linear layers with a leaky ReLU activation and sizes of 8192, 4096, and 2048 and with a final linear layer of 64 units with sigmoid activation (Figure S3.1, Encoder). The decoder followed the inverse structure with a final sigmoid activation in the last layer (Figure S3.1, Decoder). The model was compiled using the perceptual loss of the Visual Geometry Group (VGG) network (R. Zhang et al., 2018) as the loss function, and the stochastic gradient descent (SGD) algorithm (Robbins & Monro, 1951) was used as optimizer. Training was performed for 35 epochs, over individual apple cuts of the training dataset.

The SNPs and embedding values of each genotype in the training dataset were used to train a two layers model to generate an embedding predictor model. This model consisted of i) a linear layer of 300 units with sigmoid activation, which received as input the 150 SNPs; and ii) a 64 units linear layer with sigmoid activation which

connected with the 300 units layer and served later as a connection to the decoder when the model was fully trained (Figure S3.1, predictor). The embedding predictor was then compiled using mean absolute error (MAE) as the loss function and the SGD algorithm as optimizer (learning rate of 0.05 with scheduler to reduce on plateau) and then trained for 1000 epochs with an early stopper watching for over-fitting symptoms (G. Raskutti et al., 2011). The resulting model was attached to the decoder part of the autoencoder model generating a SNPs-to-Image model, the GenoDrawing (Figure S3.1, GenoDrawing). No end-to-end training was performed at this stage.

3.3.5. Embedding predictor target dataset generation.

To generate a target dataset for the embedding predictor model, we utilized a trained autoencoder to reduce the dimensionality of each image to a vector of 64 values. Subsequently, each vector representing an image was grouped by genotype, and a mean vector was computed. This mean vector represented the target for the embedding predictor models and was also utilized in the decoder to produce a mean image representing the mean image for the genotype (Figure S3.2). The resulting dataset of generated images was utilized in the evaluation of the metrics by estimating their FSI and their Shoulder Ratio (SR), which was calculated as the ratio between the width at the most top and bottom points of the apple.

3.3.6. Computational specifications

All the training was developed in a HP Workstation with an Intel Xeon W-2265 CPU, 64GB of RAM and a NVIDIA Quado RTX 4000 8GB GPU. The autoencoder training took in average 5 hours while the embedding predictors with 150 inputs took minutes. The costliest in time were the embedding predictors using all the SNP dataset as it took around 4 hours each which led to around 60h in total for 15 models. Times varied in function of early stopping activating sooner or later.

3.4. Results

3.4.1. Resulting Framework

The general approach of GenoDrawing is described in Figure 3.1. Images were separated into a training and a validation dataset conserving similar shape typologies distributions (Figure 3.1.1). The training dataset contained the 90% of the images (10271 of 320 genotypes). The rest (1340 of 36 genotypes) were used for validation (Figure 3.1.2). The autoencoder model was fitted with the training dataset to ultimately produce vectors of 64 embedding values per image (Figure 3.1.3-4). These vectors were used to generate a 10271x64 matrix, which was then reduced to a 320x64 matrix using the mean embedding values per genotype. The decoder function of the autoencoder model was then used in the 320x64 matrix to retrieve the average image for each of the 320 genotypes. The same process was

followed for the validation dataset without retraining the image compressor.

Image and genotypic datasets (Figure 3.1.5) were used for training and validating the embedding predictor model. We used two genotypic datasets: one consisted in the whole matrix of 303,239 SNPs and the other in a subset of 150 trait associated SNPs (tSNPs). Random subsets of 150 SNPs, renewed in each interaction to avoid possible bias produced by a specific SNP selection, were extracted from the whole matrix (Figure 3.1.6). The decoder section of the autoencoder and the embedding predictor were assembled into the GenoDrawing to retrieve the predicted image (Figure 3.1.7-8).

3.4.2. Autoencoder and embedding predictor training.

To reduce the complexity of the apple images we used the autoencoder network, which was able to compress (encode) each into 64 dimensions (embeddings), number that was selected During experimentation we observed that when using a low number of embeddings the decoded images lost crucial information, while when using values larger than 64 the visual quality was not significantly improved and the prediction target size was expanded, which might be detrimental.. The reconstructed (decoded) images from the 64 embeddings maintained, in a great measure, the shape aspect of the original image while the flesh color and lobule depth were not reproduced (Figure 3.2). The selected dimensionality allowed for the preservation of important information in a small dimensional space,

which served as the target for prediction. During the autoencoder training, the perceptual loss, which indicates the resemblance of the reconstructed image to the original, improved consistently from around 0.15 to a 0.09 along the 35 epochs (Figure S3.3). The mean encoder values per genotype were used together with SNP data to train the embedding predictor and reconstruct the apple section images. The embedding predictor was trained through the number of epochs needed until an early stopper was activated. For all the SNP datasets tested, a slow decrement in the mean absolute error was observed through the various iterations of training on the validation dataset, with a stagnation that arrived earlier or later depending on the number of SNPs used. There was no evidence of overfitting.

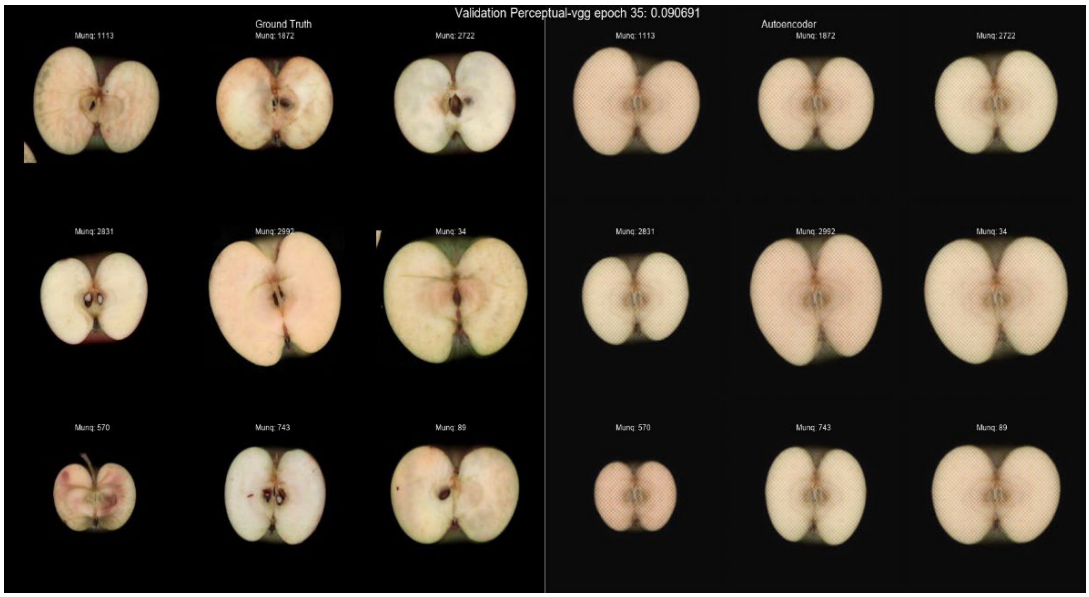


Figure 3.2: On the left side, original apple section. On the right side, output given by the autoencoder for the correspondent apple.

3.4.3. Comparing the impact of trait-related and randomly selected SNPs on embedding prediction models for GenoDrawing.

To compare the accuracy of the embedding prediction model using targeted or random SNPs, we trained 100 models for each case. Using tSNPs consistently led to lower mean absolute error (MAE) values (t-test p-value of 3.4e-108). On average, the MAE values obtained using tSNPs was 0.0864 with a standard deviation of 0.0034. When rSNPs were used, the average MAE was 0.0944 with a standard deviation of 0.0017 (Figure S3.4).

The best two embedding predictors among the 100 previously described (MAE 0.0857 and 0.0889 for the targeted and random SNP datasets, respectively) were used along with the decoder model to generate the predicted images and, eventually, test the accuracy of GenoDrawing. Mean decoded and predicted images were analyzed to determine the Fruit Shape Index (FSI), and the Shoulder Ratio (SR) (Figure S3.5). Deviation of FSI and SR values with the tSNP dataset was lower, with MAE of 0.0386 and 0.0317 for the FSI and SR, respectively, compared to MAE of 0.0651 and 0.0493 for the same to metrics in the rSNPs dataset. To determine the quality threshold for the MAE values, the mean (μ) and standard deviation (σ^2) of the FSI and SR of each of the 36 genotypes of the validation dataset were used to produce 10,000 datasets of 36 samples each, following a $N(\mu,\sigma^2)$ distribution. FSI and SR values of such samples were compared with the FSI and SR

values of the images produced by GenoDrawing. The resulting MAE superior boundary was established at 0.075 with a standard deviation of 0.0084 for the FSI, and at 0.0529 with a standard deviation of 0.0054 for the SR. Given that these boundary values exceed those produced by both rSNPs and tSNPs derived models, it may be concluded that the predictions are not merely the result of random sampling from the original distribution.

Once the non-randomness of the models had been confirmed, the distributions corresponding to both the rSNPs and tSNPs based versions of the embedding prediction were compared to the decoded means and the original data distributions for each genotype in the validation dataset (Figure 3.3). Wasserstein distances between the distributions (Table S2), show that tSNPs models generate FSI distributions with values closer to the ones of the decoded and genotype distributions (distances of 0.0209 and 0.0259, respectively), than the rSNPs models, (distances of 0.0469 and 0.0331, respectively). Similarly, distances between the decoded and the means of the SR predicted distributions were lower when using targeted SNPs (0.0184 for tSNPs compared with 0.0247 for rSNPs). Additionally, a small difference between the distribution of the decoded mean image and the original images was observed. Specifically, the Wasserstein distance values between the decoded mean distribution and the original image distributions were relatively small for both the FSI attribute, 0.0144, and the SR attribute, 0.01748.

To further understand the limitations of the proposed model, the autoencoder and predicted images were classified, based on the FSI,

into five categories (flat, flat globose, globose, oblong, and oval) (Lateur et al., 2022). The success in class assignation was assessed using the accuracy and F-Score metrics and plotted in a confusion matrix (Figure S3.6). Both metrics were consistently lower for the tSNPs-based version with an accuracy of 0.61 and F-score of 0.59 against an accuracy of 0.39 and F-Score of 0.34 for the random version. Furthermore, the the rSNPs-based version only predicted two shape categories, belonging most to the flat type, while the targeted version was able to predict fruits of three fruit classes, closer to the ground truth sample. None of them was able to produce images belonging to the oblong and oval categories.

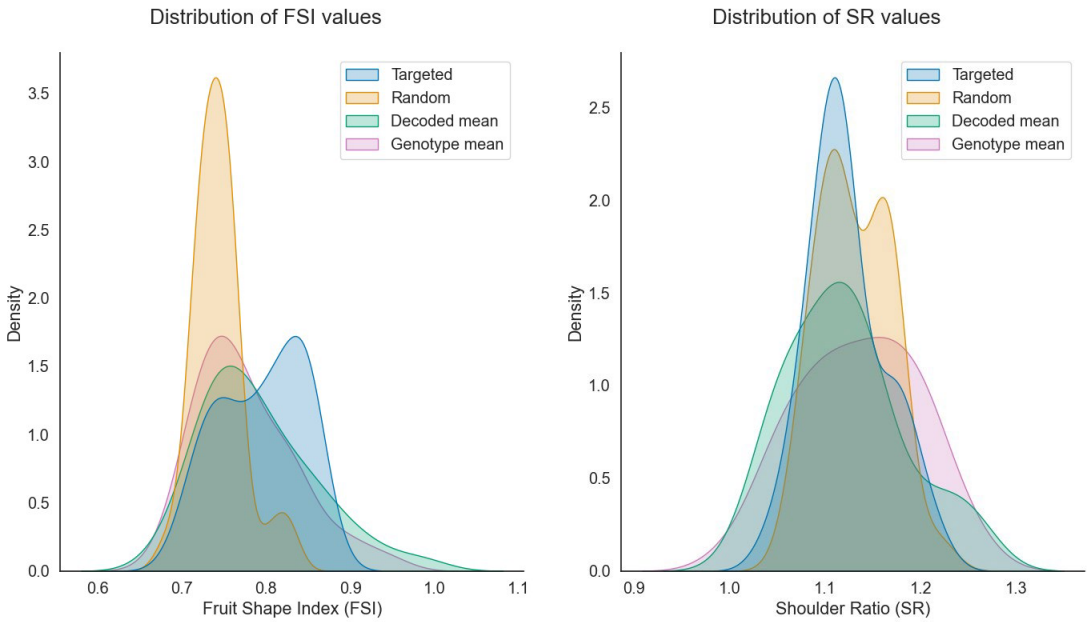


Figure 3.3: The distribution plots for two different traits: Fruit Shape Index (FSI) on the left and Shoulder Ratio (SR) on the right. The distributions were generated using the GenoDrawing approach with the best-scoring version of the embedding predictor for both trait-related SNPs (represented in blue) and random SNPs (represented in orange). For comparison, the distributions of the validation dataset were also plotted using the mean images (represented in green) and the original images (represented in purple).

3.4.4. Effect of Adding Non-Shape-Related Markers on The Embedding Predictor Model Accuracy.

To investigate the effect of increasing the number of markers selected without knowing the biological relevance, we added to the tSNP dataset an increasing number of SNPs in nine steps up to a total of: 150, 300, 1650, 4650, 10150, 20150, 50150, 100150, and 300150. We used a similar approach with the random-selected SNPs, increasing the size of the selection at each step. For each of these steps and datasets, 15 models were trained. The embedding predictor models with only 150 tSNPs were generally more effective in achieving a low mean absolute error (MAE) compared to models that include additional SNPs. The exception was observed when using 300150 targeted version, where three outliers showed lower MAEs than their random selection counterparts. Notably, a unique case was observed where a targeted version consisting of 300 SNPs achieved the lowest observed MAE of0.0835 (Figure 3.4).

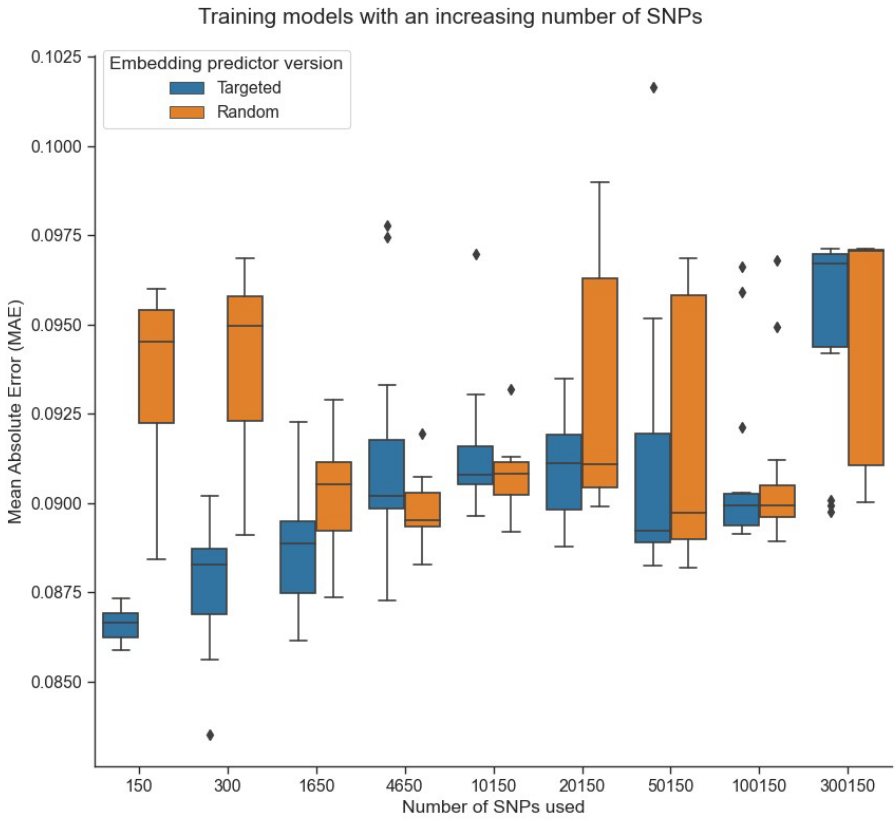


Figure 3.4: Embedding predictor models with increasing size represented by their mean absolute error. Lower in the y axis means less error and, therefore, a better model. Two versions of the models were observed: a targeted version that always included 150 SNPs relevant to shape as well as random markers, denoted by blue color, and a random version which involved a fully random selection of markers, denoted by orange color. Fifteen models were trained for each of the different sizes and versions.

3.5. Discussion

3.5.1. Evaluation of Autoencoder Performance, Loss Function Selection, and Embedding Predictor Structure.

In this study, we employed an autoencoder network to simplify the complexity of the apple images into 64 embeddings. This number of embeddings was deemed sufficient to account for small variations in the images while circumventing the complexity associated with adjusting a large number of embeddings to a reduced dataset of SNPs. Although changes in the number of embeddings can produce different outcomes, given the scope of this study, we selected the number that aligned better with our objectives. The images produced from the embeddings successfully captured a reliable and straightforward representation of the apples, effectively reproducing the original image structure. Nevertheless, certain features, such as lobule depth and flesh color, were not accurately represented, likely due to the sparsity of the embedding space and the difficulty faced by a simple autoencoder network in generating photorealistic images. As the primary focus of this study was on shape, the autoencoder structure was deemed appropriate for the task. However, future studies that aim to capture a wider range of attributes may benefit from utilizing a different structure to better accomplish the reconstruction task.

With respect to the loss function utilized in our study, we evaluated the Mean Square Error (MSE), Structural Similarity Index Measure

(SSIM) based metrics, and perceptual loss (Figure S3.7). We selected the perceptual loss given the well proved adequacy to the reconstruction task (R. Zhang et al., 2018) and the results provided. Although SSIM-based metrics, such as MS-SSIM and CW-SSIM, are robust in terms of comparing image structures, they often overlook details (Temel & Alregib, 2014). The images reconstructed were similar to the original ones in terms of shape, while they were less alike regarding other visual aspects as fruit conicity. Furthermore, MSE tendency to generate blurry boundaries between figures (Z. Wang & Bovik, 2009) proved to be highly detrimental to the comparison of image shapes, preventing a proper examination of the data.

The embedding predictor structure employed in this study was relatively simple, consisting of two single layers with 300 units each and sigmoid activation. This approach was selected following an examination of several architectures, including deeper models with skip connections, convolutions, and direct models, on the 150 SNP version predictions. The simplicity of this approach allowed for the exploration of different input sizes while minimizing training time. For larger datasets, however, more complex architectures capable of studying sequences while retaining position relevance could prove beneficial. Nevertheless, the primary goal of this study was to assess the feasibility of this approach and the impact of selecting the appropriate SNPs relevant to the features studied.

3.5.2. Using the mean embedding as a representation of the phenotype

One of the primary challenges associated with utilizing an autoencoder is determining which information is being encoded in the embedding space. In our study, a relevant question that arose was whether the mean of the embeddings represented the mean of the images corresponding to a specific genotype. To evaluate this, we compared the distribution of both for the FSI and SR metrics. We found out that the FSI distributions were highly similar both in distributions and images; however, differences were observed in the SR(Figure 3.3) with Wasserstein distances larger on the SR than in the FSI (0.0178 vs 0.1444). While no significant difference were obtained with a Kolmogorov-Smirnov test, p-values were lower for the SR (0.34) than for the FSI (0.88) which might hint to what can be observed in the validation dataset, a total loss of the symmetry in the decoding and the averaging process. An asymmetric apple that is cut in two halves will produce two asymmetric mirror images, which will be then merged in a symmetric mean image. This issue could be tackled in the future by removing from the dataset one of the halves of each apple.

3.5.3. Relevance of SNP selection in genome prediction models

Genome prediction models are typically founded on dense SNP data. Nevertheless, generating and managing large SNP datasets remains a challenge for small to medium-sized breeding companies, despite the relatively low cost of high-throughput genotyping methods. Strategies such as genotyping-by-sequencing (GBS) have been developed to reduce genotyping costs (Elshire et al., 2011); however, the use of smaller arrays would be more conducive to the implementation of this methodology. To this end, we reduced the genotyping array to a small array of 150 SNPs that we found to be significant for size and shape traits. In this work we discarded predicting fruit size since the images to feed the model were an average of all the images of a genotype collected over three years, and Jung et al. (Jung et al., 2022) found low between year correlation of some of the attributes. To evaluate the adequacy of reducing the number of SNPs for the predictions, we tested the model with both random and selected SNPs and consistently observed improved results in predicting shape when using the selected SNPs. This confirms the association of most of the selected SNPs with apple shape (Figure S3.4). However, we also identified a rare case in which the addition of random SNPs to the selected SNPs led to better results, leading us to believe that our selection may be missing some shape-relevant SNPs. The inclusion of relevant SNPs is crucial for accurate predictions, which is consistent with the biological interpretation of SNP information. In theory, the

presence of the relevant SNPs should be sufficient for making predictions, regardless of the addition of random SNPs. Our results demonstrate a consistent improvement in performance when the relevant SNPs are included in the input, with a decrease in performance as the input size increases. This decrease in performance could be attributed to the simplicity of the embedding prediction structure used in our study. As previously mentioned, a more complex architecture may be necessary for larger inputs to better extract relevant information, which can be addressed through architecture improvements.

3.5.4. Evaluating embedding predictor performance using FSI and SR metrics in GenoDrawing.

Two metrics were employed to evaluate the similarity in fruit shape between the mean image generated by encoding all images for a particular genotype and their prediction. FSI was used as a reliable measure of the general shape of the fruit, while the shoulder ratio was used to capture the conicity of the apple. However, it was noted that the shoulder ratio was not adequately captured by using the mean embedding, limiting the interpretability of the results obtained, but still providing valuable insight into the challenges of using this model to predict asymmetric metrics. Conversely, the FSI metric was found

to be well-suited to the task. The results, as shown in Figure 3.3, revealed that the random embedding predictor was highly focused on a limited range of variability for predictions. This indicated that the embedding predictor consistently produced similar images to minimize the error but did not effectively learn to solve the task. In contrast, the targeted version produced a better-distributed range of predictions, which could achieve higher FSI values and result in better approximations of the image production task. This was further supported by the categorization of predicted images into five shape categories (Figure S3.6) using FSI scores, which revealed that the random version only predicted in two categories (flat and flat-globose), while the targeted version predicted in three categories (from flat to globose). Overall, these findings suggest that GenoDrawing can learn the task effectively when relevant markers for shape are provided.

3.5.5. Limitations

By estimating the mean shape appearance for a genotype, we aimed to cover the variability within genotype, but apples can be very influenced by environmental effect, leading to a wide range of phenotypes within the same individual (Jung et al., 2022). Even though most of the genotypes counted with multiple biological replicates and two-year imaging, this is one of the biggest challenges faced in genomic prediction. Furthermore, the overrepresentation of flat-globose apples limits the capacity of the model trained, forcing it to be biased toward this category (Figure S3.8). Removing the number of examples from this category has been considered, but this considerably would limit the size of the dataset in term of number of genotypes which might hurt the learning process more than helping it. For traits with such a low separation between classes a larger genotype dataset, including more extreme phenotypes may help to improve accuracy in future studies.

3.5.6. Data Availability:

The code repository including notebooks, and models with their trained weights can be found in the following github repository: https://github.com/Fedjurrui/GenoDrawing. The images were produced in (Dujak et al., 2023), from where the images can be obtained under request.

Chapter 4: Generating images from SNP data through artificial intelligence

4.1. Abstract

Generative artificial intelligence (AI) has been the subject of multiple transformative efforts in the past eight years. The advent of transformer-based models for text generation, and diffusion models for image generation has initiated a veritable “gold rush” to uncover novel applications of these technologies. Two prevailing research directions have emerged: the ChatGPT-like systems that generate literature for text-based tasks, and the image generator models that create representative examples for different research situations. Despite some applications in image generation, the plant science community has primarily focused on the artificial augmentation of databases to train detection models. Here we introduce an innovative approach by leveraging neural networks for phenotype image prediction based on genomic markers. We propose the combination of a genetic model with an image generator to produce an array of predictive outcomes for a given genotype. Our research seeks to illustrate the vast potential of integrating advanced generative AI technologies into plant breeding. This combination aims to improve trait prediction accuracy and efficiency, thus enhancing breeding strategies’ overall effectiveness. By bridging the gap between genomics and image generation, we anticipate significant advancements in plant breeding methodologies, paving the way for more robust AI-driven plant science.

4.2. Introduction

Generative Artificial Intelligence (AI), characterized by its ability to generate new content from learned data, has found its way into multiple disciplines, with transformer-based models applied to text and diffusion models employed for image creation. Transformers are models built around the attention mechanism that highlights the relationship between inputs. This is extremely important when the context (as for example the order of the words in the text or of the K-mers in the DNA sequence) can significantly alter the meaning of the output. These technologies have been extensively utilized in fields such as natural language processing (NLP)(Brown et al., 2020; Touvron et al., 2023) and digital art (Ramesh et al., 2022b; Rombach et al., 2021). To date, plant science's encounter with generative AI has primarily focused on augmenting databases artificially to improve detection models (Douarre et al., 2019; Min et al., 2023). This is proving to be a relevant application as it largely reduce the time and cost of producing large datasets with minimal costs. However, there remains a largely untapped potential for generative AI applied to genomic prediction. Genomic prediction has been extensively used in breeding and is largely based on linear and logistic regression models (Uffelmann et al., 2021), which are powerful association methods for continuous and discrete traits. Nonetheless, when searching for associations between genetic markers and complex traits assessed by visual evaluation, which could be well captured by images, the only way to approach the task with these models would be measuring specific attributes of the images, i.e. colour, height or specific ratios. Assuming that the chosen trait is distinctly represented in the genome,

this may not pose a problem. However, there's always potential risk of observer bias when deciding on the specific traits to quantify. Merging genomics and image generation has the potential to enhance the precision of trait prediction and the effectiveness of breeding strategies. The feasibility of this approach has been proved in Chapter 5, yet there is still much to explore. This study seeks to test different models on the applications of generative AI and its possible integration with genetic models, adding a new dimension to the existing AI applications in life science.

The GenoDrawing study (Jurado-Ruiz et al., 2023, Chapter 5) served as a conceptual platform, where we explored the feasibility of linking Single Nucleotide Polymorphisms (SNPs) to image-based phenotypes. Our objective was to generate images from SNPs, and two specific metrics were used to measure the shape of an apple: the Fruit Shape Index (FSI), indicating the ratio between height and width, and the Shoulder Ratio (SR), expressing the ratio between top and bottom amplitude of the apple. The principal methodology involved reducing the dimensionality of the images to 64 values using an autoencoder, which subsequently served as the basis for a predictive model. This model aimed to reconstruct the images based on genetic markers. Despite the inherent complexity of the task and the simplicity of the model, we achieved considerable success in reconstructing the shape. The autoencoder's simplicity, however, led to a loss of substantial information during the reconstruction, including the color, which is an element crucial to the completeness of the image prediction. Additionally, the input of the SNPs through two linear layers of neurons might have restricted the model's precision in predicting the

64 values derived from the genetic data. While these limitations highlight areas for improvement, the model validated the potential of the concept.

Motivated by the insights from GenoDrawing, we set out to find novel tools that could address its limitations. Studies have shown that the compression of intermediate information is more efficient when employing vector quantization (VQ) (van den Oord et al., 2017). Thus, incorporating VQ into the process could not only improve image quality but also enhance the extraction of information from SNP vectors. One of the disadvantages of using VQ is determinism. No randomness is involved in the output and, therefore, it cannot be used to produce a range of examples. This limitation reduces its ability to reflect the reality and diversity of phenotypes we might find in a fruit tree. Luckily, this issue was addressed in by (Esser et al., 2021), who used a transformer to take advantage of the discretization of the information and create a predictive model that uses as ”context” the information given by the genetic markers.

In this chapter we investigate the implementation of VQ and transformers in the GenoDrawing approach as a possibility to overcome current limitations. Also, we use the whole SNP matrix (with 303,239 SNPs) instead of a subset of shape-related SNPs for its use in traits and species for which the previous knowledge of associated SNPs may be missing.

4.3. Methods

4.3.1. Dataset

The dataset used in this study is the same used in Chapter 5 of this thesis. It is compound of 356 genotypes from the reference population of apple (AppleRefPop). For each of these genotypes, we used 6 to 10 images corresponding to apple fruit sections and the corresponding set of SNPs for the genotype (303,239 SNP). A deeper description of the dataset can be found at (Dujak et al., 2023). The dataset used the same partition as in Chapter 5 of this thesis, with a 90% (322 genotypes) of the examples for training and 10% (34 genotypes) for testing.

4.3.2. Neural networks architecture and training.

Vector quantization for images.

The image reconstruction model was a CNN based on the architecture detailed in GenoDrawing (Jurado-Ruiz et al., 2023, Chapter 5) without the final linear layers, giving as an output 12x12x512 matrices. The architecture is based on vector quantization(VQ) as described by van den Oord et al. (van den Oord et al., 2017). The output given by the modified GenoDrawing encoder was divided into 144 vectors of 512 length. These vectors were compared using Euclidian distance with the Codebook entries and the most similar index was retrieved. The codebook was built with 64 entries and 512 features per entry. The retrieved indices for each of the 144 vectors were used to build a new

matrix which served as input for the decoder. The decoder goal was to rebuild the original matrix from this matrix (Figure 4.1). The model was trained using the Adam algorithm (Kingma & Ba, 2014) and the sum of four different losses as optimization goal. Three losses represent the quality of the image reconstruction, being them mean squared error (MSE), the mean absolute error (L1) and the perceptual loss using the Visual Geometry Group 16 (VGG16) network (R. Zhang et al., 2018). The fourth loss was the commitment loss defined in the original manuscript, used to improve the usage of the codebook. The learning rate was adjusted following a warmup technique, with a rapid increase in the learning rate during the ten percent of the epochs and slow decrement for the remaining. The learning rate started at 1e-5, rising to 1e-3 and decaying to 1e-7.

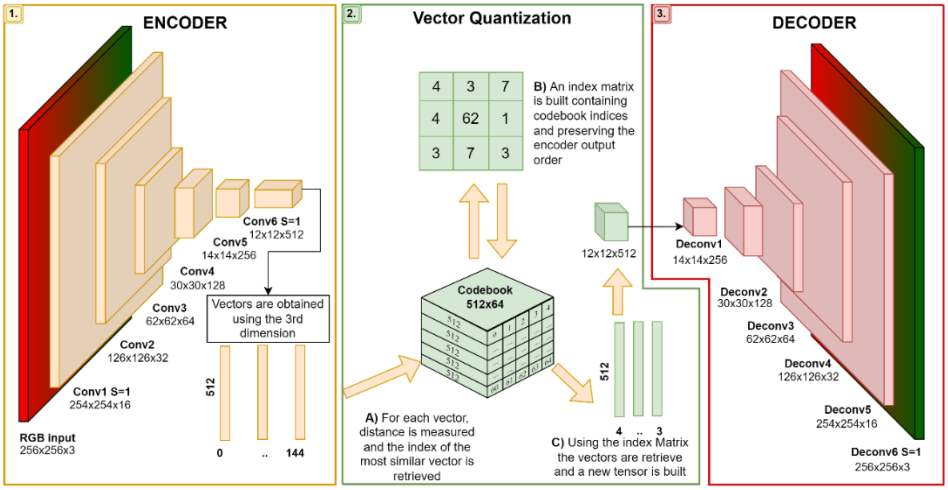


Figure 4.1: Vector quantization illustrated. **1)** Original images, in 256x256 pixels were reduced through a Convolutional Neural Network (CNN) to a 3D matrix of size 12x12 (total of 144 vectors) with a depth of 512. **2.A)** Each of these vectors with 512 values was compared using Euclidian distance with a codebook containing 64 vectors of 512 length. **2.B)** This allows the discretization of the information into a new matrix of 12x12 (144) positions containing the indices. **2.C)** These indices could be used to retrieve the codebook values for those positions and construct a new matrix of 12x12 with depth 512. **3)** The new matrix was then computed through a decoder CNN to reconstruct the original image.

Vector quantization for SNPs.

The SNPs were first set in a range from 0 to 1. In this range 0 represented a missing input, 0.33 homozygote for allele A, 0.66 heterozygote and 1 homozygote for allele B. The SNPs vectors were padded to 303.616 by adding zeros so a square matrix of 593x512 was formed. For the SNPs quantization, a similar vector quantization was employed with a key difference. Following the work developed by Dosovitskiy (Dosovitskiy et al., 2021) in vision transformers (ViT), the CNN architecture was substituted by a transformer encoder in the encoder step, and by second transformer encoder in the decoder (Figure 4.2). The positional encoders and decoders were modified respecting the original ViT on the positional information. The patches definition and positional information added was defined by the chromosomes, providing information related to the relative position of the SNP in the genomic space. As the SNP arrays are one dimensional, a convolution of one dimension was used in the linear projection. The codebook defined contained 64 entries with 512 features. The redimension from 303,616 was made to a total of 593 entries of the codebook. Three losses were used during the reconstruction, mean absolute error (L1), mean squared error (MSE) and commitment loss. To determine the reconstruction accuracy, and not for training, the outputs were rounded to the closest value 0, 0.33, 0.66 and 1. This allowed the retrieval of the predicted categories and the determination of true positives.

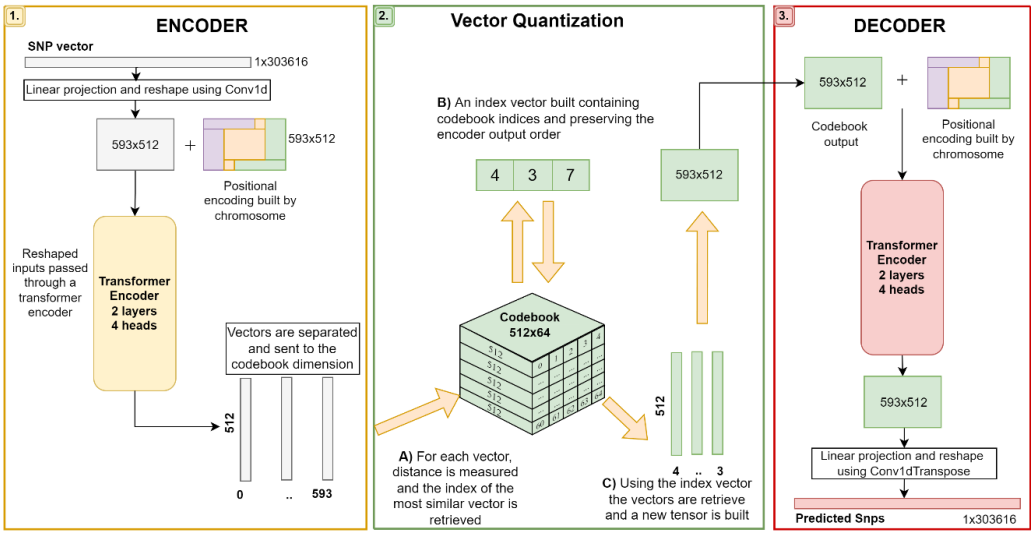


Figure 4.2: Vector quantization for SNPs illustrated. **1)** An SNP vector of size 303,616 is linearly projected and reshaped, using a convolutional layer, into a 593x512 matrix. A corresponding matrix was prepared with the position information by chromosome and added to the SNPs matrix. This final matrix was fed to a transformer encoder with 2 layers and 4 heads giving as output the processed matrix. **2.A)** From the SNP matrix, 593 vectors were obtained with 512 each. These vectors were compared using Euclidian distance with a codebook containing 64 vectors of 512 length. **2.B)** This allowed the discretization of the information into a new vector of 593 length containing the indices. **2.C)** These indices could be used to retrieve the codebook values for those positions and reconstruct the SNP matrix with a shape of 593x512. **3)** The SNP matrix was fed to a transformer encoder with 2 layers and 4 heads tasked to reconstruct the original inputs.

Connection through a transformer.

To connect both vector quantiziers a transformer decoder was used. With the quantiziers trained all the information was transformed into index vectors using their correspondent model. This generated a conditioning vector of 593 tokens with a target of 144 tokens. The transformer was built following the model defined by OpenAI as GPT2 in (Brown et al., 2020; Radford et al., 2019). Three different versions of GPT were tested in this work with different architecture. For each of the tested versions, two different trainings were performed following the same learning strategy for a different number of epochs. In every case the strategy followed a learning rate with a starting value of 1e-5 and with a linear increment up to 1e-3 for half of the training. The second half of the training would see cosine decrease up to 1e-7 by the end of the process. Two GPT were built with 8 layers, 8 heads and depth of 768 in the codebook, one trained through 100 epochs and other for 150 epochs. Two additional GPT were built with 12 layers, 12 heads, 768 depth, one trained through 100 and other for 150 epochs. Last, two GPTs were built with 16 layers, 12 heads, 768 depth, one trained through 100 epochs and the other for 200 epochs.

4.3.3. Evaluation of the predicted images.

Each GPT was evaluated comparing the similarity of the images produced with different shape measurements produced by the Shape Analyzer. For each genotype in the evaluation dataset we measured up to 9 of the original apple cuts and generated the same number of examples to be measured. As the evaluation could not be performed

one to one, the datasets were compared using Wilcoxon test and distribution density plots.

4.3.4. Evaluation of SNP reconstruction accuracy.

To evaluate the quality of the SNPs VQ we used the frequency of the most frequent allele in each of the positions as base reference. If the ANN is unable to extract any useful information it will default to predict the most frequent allele for each position to reduce the error rate. This would be showcased with a reconstruction mean accuracy equal to the frequency of the most common allele for that position.

4.3.5. Code and libraries used.

The neural networks were developed fully in Python, utilizing the PyTorch and PyTorch-lightning libraries. Additionally, we incorporated code from the repository published alongside the original article (Esser et al., 2021).

4.4. Results

4.4.1. Image reconstruction

After the effective vector quantization (VQ) into 12x12 indices, the image reconstruction preserved great resemblance to the originals. The loss compound of mean squared error, mean absolute error and perceptual loss was effectively reduced in train and test, and clear similarities can be observed between originals and reconstructions. One common artifact was the smoothness of the decoded images, which minimized or removed features such as water core or defects of the apple (Figure 4.3).

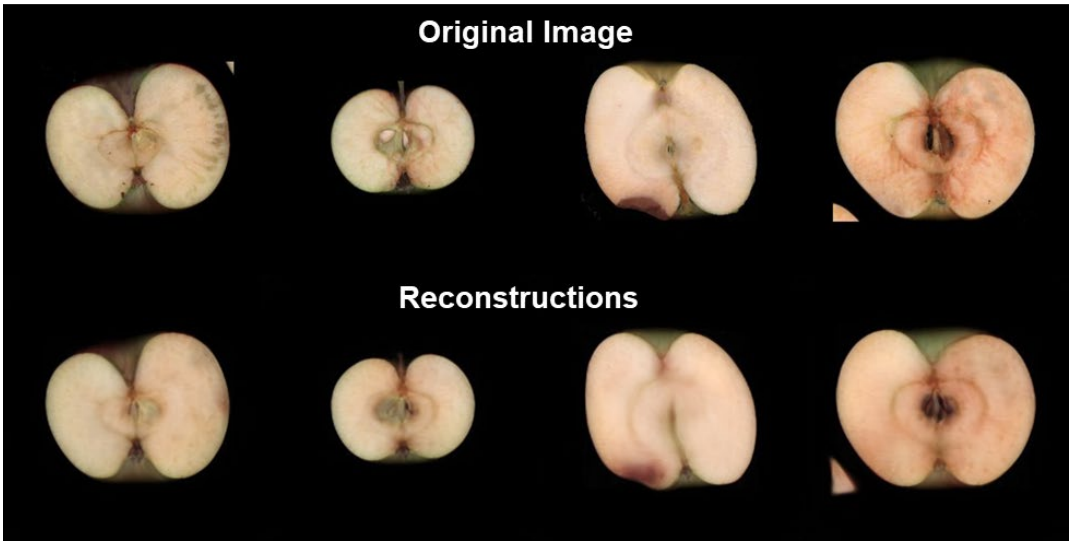


Figure 4.3: Vector quantization for apple cuts. Pairs top bottom display the original images with their corresponding reconstructions.

4.4.2. SNP reconstruction

In this approach, we used the whole dataset of SNPs instead of a subset of trait-associated markers. The information embedded within the SNP matrix was discretized developing a new VQ autoencoder with the reconstruction task. The mean accuracy rate for the reconstructions was 85,65% across all chromosomes, surpassing a base random reconstruction rate of 64,61%, which was obtained using the average frequency of the most frequent allele for each position. The minimum accuracy was found in chromosome 17 with 81.75% while the maximum accuracy was 88,97% in chromosome 15 (Table 4.1).

The accuracy was evaluated as a general value per chromosome as well as per position. The visual representation of the global accuracy in Figure 4.4 shows how the accuracy in most of the cases surpasses the most frequent allele for that position.

Table 4.1: Accuracy of the reconstruction obtained by the vector quantization. The base rate of error in a random reconstruction is 64,61%. The maximum obtained 88,23% while the minimum 81.8%.

Chromosome	**1**	**2**	**3**	**4**	**5**	**6**
Accuracy	88,23%	83,31%	86,27%	84,35%	85,47%	88,14%
Frequent allele	65,20%	63,98%	63,40%	65,44%	63,15%	67,31%
	7	**8**	**9**	**10**	**11**	**12**
	87,32%	86,02%	84,02%	85,17%	83,62%	83,53%
	65,09%	64,12%	64,89%	65,91%	63,78%	64,44%
	13	**14**	**15**	**16**	**17**	
	88,53%	86,18%	88,97%	85,60%	81,75%	
	63,47%	64,55%	61,89%	65,16%	66,86%	

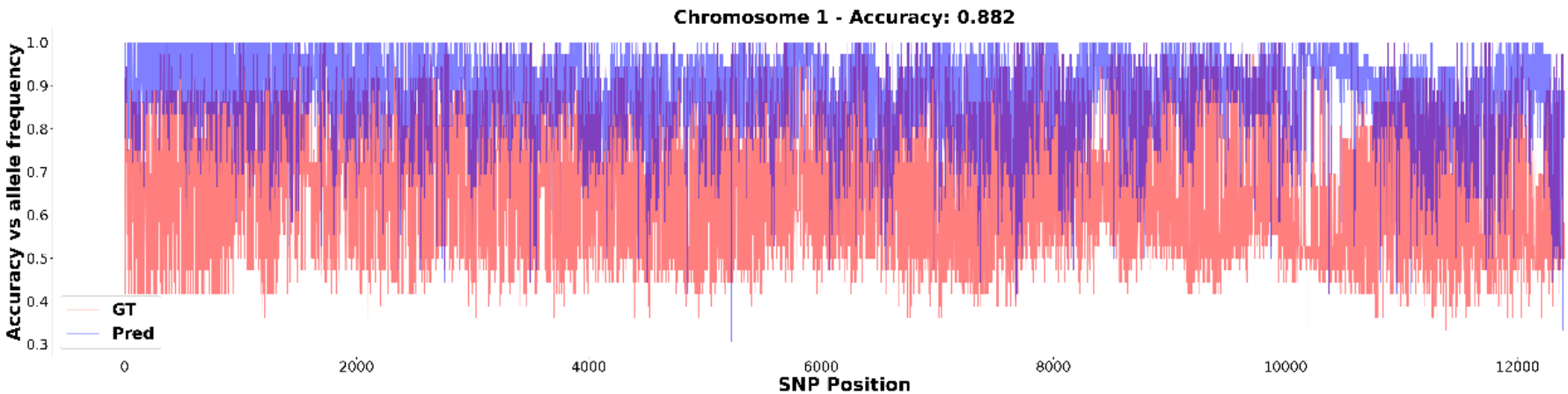


Figure 4.4: Vector quantization of SNPs. The red line, representing the frequency of the most probable allele in that region, is lower than the blue line, which symbolizes the correctly predicted allele for each of these positions.

4.4.3. Connection through transformer

Once both models were trained, a connection was established through a transformer. Given a 14% random base error, all six models tested achieved a higher accuracy, with all of them peaking at around 44% accuracy (Figure S4.1). All models were tested by comparing the shapes similitude to the original dataset using Shape Analyzer. Figure 4.5 presents a series of distribution comparison plots between the ground truth and the predictions from one of the generative models, but all displayed a very similar pattern. For height (px), the ground truth distribution was broad, centered around 130-150 pixels, while the prediction was narrowly centered around 140 pixels. The width (px) distribution of the ground truth ranged from 125 to 250 pixels, whereas the prediction was centered around 175 pixels. The ground truth and predicted distributions for the FSI (Fruit Shape Index) were similar, though the latter was more peaked. The box ratio distribution for the prediction was more peaked and centered around 0.8, with the ground truth showing a broader range. In terms of roundness, the prediction was more peaked and slightly shifted to the left compared to the ground truth. The solidity distribution was narrower for the prediction. The compactness distribution for the prediction is sharply peaked around 0.95, while the ground truth has a wider distribution. The cone feature's prediction distribution was more peaked and centered around 0.6, with the ground truth showing more variability. Both symmetry and symmetry Jaccard features showed more peaked prediction distributions compared to the broader ground truth

distributions. The top angle distributions were similar in shape and spread, indicating that the model captures this feature well. Similarly, the bot angle distributions overlap, showing that the model captures both the central tendency and variability well for this feature. Overall, the generative model captures the central tendencies of the features reasonably well but generally produces distributions with lower variance compared to the ground truth.

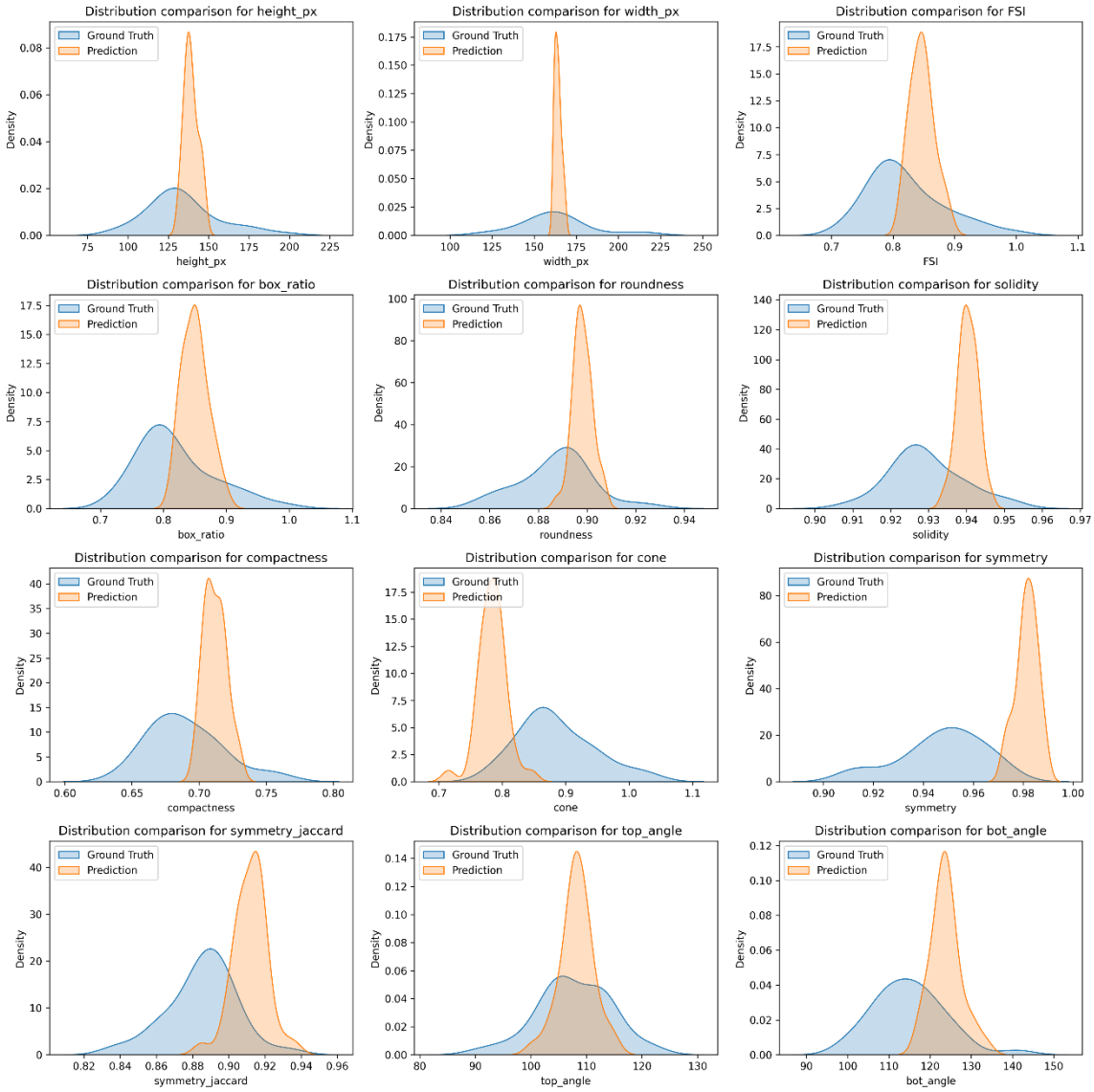


Figure 4.5: Comparison of the distribution of the ground truth and prediction values of apple fruit shape and size attributes using the Transformer model with 8 heads, 8 layers, trained for 98 epochs. Each plot shows the density distribution for a specific feature, including height (px), width (px), FSI, box ratio, roundness, solidity, compactness, cone, symmetry, symmetry Jaccard, top angle, and bot

angle. The ground truth distributions are depicted in blue, while the prediction distributions are shown in orange.

4.5. Discussion

4.5.1. Quality of the quantized images.

The vector quantization approach represents a significant advancement in image reconstruction over the previous work developed in GenoDrawing (Jurado-Ruiz et al., 2023, Chapter 5). It excels at maintaining a high degree of resemblance to the original images, preserving features such as flesh color and finer details like lobule depth better than a traditional autoencoder. However, the smoothness observed in the reconstructed images remains a known limitation of this method, underscoring the need for future enhancements. Integrating generative adversarial networks (GANs) could potentially address this issue by providing more detailed and accurate image reconstructions (Esser et al., 2021).

4.5.2. SNP vector quantizer effectiveness.

One of the key goals of this work was to effectively represent the information provided by a large set of SNPs into a lower number of features while maintaining the relevant information about both their position and their relationship to other neighboring SNPs. To accomplish this task, a mix between a transformer encoder and vector quantization has been proposed. The mechanism of attention in transformers is ideal as it captures the relevance of a given input in

their full context. This has been showcased both in vision transformers (ViT) and in large language models (LLMs). As the proposed approach has been so far unexplored, we decided on the autoregression task, which allows to directly study the efficiency of the compression by studying the reconstruction accuracy. The achieved mean reconstruction accuracy of 85.6% across all chromosomes significantly surpasses the base random reconstruction accuracy of 64,6%, demonstrating the model's fitness to the task. In other experiments (not showcased in this work) we observed that linear layers and convolutional layers are unable to surpass the base random accuracy. Therefore, we can infer that different usages of the same model will be able to extract an adequate amount of the input information, surpassing the issues traditionally found with linear and convolutional layers. This work opens the door to further research in the line of genomic prediction through transformers which would enable to use whole SNPs arrays and not only informative trait-associated SNPs.

4.5.3. Predictions evaluation.

Integrating both vector quantizers with a transformer model to predict phenotypic traits from genetic data has yielded an encouraging accuracy of around 44% when predicting image indices. Eventhough this is a substantial improvement over the random base error rate of 14% the accuracy of this predictions does not guarantee a good result in the image prediction. One of the challenges exposed in GenoDrawing was to avoid falling in the pit of global mean representation for the model. This is what happened when non-

relevant SNPs were used to create images. In the results this is what it seems to happen here. While the central tendencies of the predicted traits are well captured, the variance in the predicted trait distributions is generally lower than the ground truth. This suggests that while the model is good at capturing average trait values, it struggles with capturing the full variability present in the actual phenotypes. This raises a difficult situation as discerning in this case the root of the problem is complex. The full process could be failing at different stages of development. It could be 1) that the SNP quantization even if effective, is not good enough to translate this information to the image generation. 2) The relevant SNP information does not represent a large enough part of the conditioning to make the image generation focused on such part. 3) The dataset is not big enough to fit such a complex model.

4.6. Conclusions

This study presents significant advancements in image reconstruction and SNP representation through the use of vector quantization and transformer models. The vector quantization approach markedly improves image quality, maintaining high fidelity to original images and preserving fine details better than traditional autoencoders, although it still faces limitations in achieving image smoothness. The SNP vector quantizer, combined with a transformer encoder, has proven effective in reducing the dimensionality of large SNP datasets while retaining essential positional and relational information. Achieving a mean reconstruction accuracy of 85.6% across all chromosomes highlights the superiority of this method over baseline

approaches, indicating its potential for various applications beyond the scope of this study. Predicting phenotypic traits from genetic data using the combined vector quantizer and transformer model has shown interesting results. However, the model's current inability to fully capture the variance in predicted trait distributions suggests areas for improvement. Potential issues could stem from the SNP quantization process, the extent of relevant SNP information, or the dataset's size and complexity. Overall, this work demonstrates the potential of advanced machine learning techniques in genomic prediction and image reconstruction, paving the way for future research to enhance model performance and address identified challenges.

General Discussion.

Discussion

In this thesis, we explored various artificial intelligence (AI) methods for phenotyping and genomic prediction. The accurate and efficient analysis of phenotypic and genotypic data is crucial for advancements in the fields of breeding and agriculture. This work aimed to showcase different AI approaches to solve current and emerging challenges, evaluating their effectiveness. Through all four core chapters, we have shown applications of detection, segmentation, tracking, and generative models covering some of the largest and most relevant fields in AI.

Synthesis of the findings and highlights.

Chapter 3 demonstrated how contemporary detection and segmentation methods can revitalize existing software, enabling it to perform intricate segmentation tasks. A key highlight of this chapter is the pressing need for open-source, user-friendly software tools accessible to the scientific community. Many potential users lack the skills to code and manipulate images through programming. By providing an editable open-source tool that is also easy to use, we enhance accessibility for both bioinformaticians and researchers. This dual approach ensures that sophisticated segmentation tasks can be performed without requiring extensive programming knowledge, thus broadening the tool's usability and impact.

In Chapter 4, we explored the combination of detection and segmentation methods with simple algorithms to create innovative solutions to existing challenges. This work forms the foundation for

future, more refined approaches and provides new insights into leaf photosynthetic behavior. By tracking leaves, we identified different growth and photosynthesis patterns between two genotypes. While further research is necessary, this work paves the way for ongoing advancements in understanding plant physiology and behavior.

Chapter 5 introduced a novel approach to predicting complex traits in plants using generative AI. The research leveraged advancements in generative AI to address the difficulty of describing phenotypes, such as fruit shapes, with traditional metrics. For example, describing an apple's shape involves multiple traits—circularity, height, width, and lobulations—none of which alone capture the complete shape. Our approach aimed to generate images from genetic information, reflecting the complex polygenic nature of these traits. This work aligns with recent studies identifying shape traits (Dujak et al., 2024) and suggests that traits encoded in the genome can be visualized and analyzed through AI-generated images. This approach demonstrated the feasibility of generating understandable images from genetic data, providing a new avenue for phenotypic analysis.

Last, in Chapter 6, we advanced from the basic concept of image generation from SNPs to developing a single model aimed at improving the issues identified in Chapter 5. Although the final model did not perform as intended, the insights gained during this research are valuable. We identified ways to process SNPs using transformers and improved image quality through vector quantization. These insights lay the groundwork for future development, suggesting paths

to overcome current limitations and enhance the effectiveness of AI in genomic prediction.

As a summary, throughout this thesis, we demonstrated the potential of AI methods to address various challenges in phenotyping and genomic prediction. Each chapter highlighted different aspects of AI applications, from enhancing image segmentation tools to developing novel approaches for trait prediction. The integration of detection, segmentation, tracking, and generative models offers a comprehensive toolkit for future research and practical applications in breeding and agriculture

Current state, future prospectives and personal views.

The field of AI is evolving at an impressive pace. When I was developing the first chapter of this thesis and learning on CNNs, Mask R-CNN was the most used model architecture, for instance segmentation. However, since 2020 we have seen an explosion of different alternatives and methods that could not only equal but outperform this option. The most used method now is YOLO (You Only Look Once), which in 2020 could only perform instance detection without segmentation. Right now, YOLO, in version eight, is used for every possible vision task, including instance segmentation, human pose detection, key point detection and more. This is only one example of fast evolution in general architecture. If we move to the concrete blocks that compound this architecture (also known as backend) imagination is the only limit. CNNs were, and still are the main option

for the backend on computer vision. However, with the groundbreaking publication on transformers in 2019 and the subsequent release of ChatGPT in 2022, new alternatives are emerging. The usage of transformers as a substitute in vision tasks is consolidating as shown by architecture such as the Vision Transformer (ViT). This kind of architecture is noteworthy as transformers base models are aware of the context and the relationship between the inputs, which in some cases is determinant for a positive outcome. This awareness has been shown to be utterly important in image generation where the goal is to provide a coherent image.

Generative AI is the field which is living the greatest of all changes. Transformers revolutionized the text generation field while the concept of diffusion has transformed the image generation field. We have evolved from variational autoencoders and generative adversarial networks (GANs) being the way to go in 2016 to the current state using denoising and diffusion which provides a superior result. Furthermore, before 2020 most of the conditioning on generative models was extremely limited, with ideas such as CycleGANs allowing limited modifications of images into different variations, such as changing a horse for a zebra or a landscape from winter to summer. Currently, as we have seen in DALL-E2, diffusion is capable of transforming text embeddings into image embeddings. This allows us to represent the idea contained in a paragraph into an actual image. But if this was not impressive enough, recently multimodal models have made an appearance. The approach to multimodality was coupling different models with different capabilities and attempting

to make a connection later. Now we are witnessing generative models that can natively generate text, images and audio and use as inputs any of those. These achievements have wide implications in other fields of research. More precisely, in plant biology, it means that we can develop models which integrate multiple types of information. This could be translated into models using genetics, soil, climate and textual descriptions to make predictions of the plant production and development. More importantly, these models have the potential to capture and emulate the emergent properties produced by the interaction of complex systems. Following this line of research, we can come closer to a true in silico reproduction of breeding programs and agriculture, consolidating the field of digital twins to plant science. These digitals twins will allow for rapid response and adaptation which in a climate change scenario is utterly important to secure food supplies and food variety.

The mayor goal of this thesis was to test and prove different applications of deep learning to plant research. This has been achieved but there is still a mayor concern, the pace of adaptation. As we were proving that these applications did work as expected, we were also watching the field advance and leave the methodology antiquated and serving just as a demonstration of the feasibility. Adapting to the new landscape of AI requires an enormous effort of active research and investment which might be overwhelming for small research teams. Academic research is more transparent than industry research, but also, as consequence of this transparency is inherently slower in the publication of results. Both the development and publication paces have led to a situation where the AI industry seems to be years ahead

of AI academic publications. But at this point, it is unclear how ahead of the industry actually is. The lack of transparency with the drift of companies such as OpenAI from open-source to completely closed-source put an enormous pressure on academics to decipher how achievements such as multimodality are achieved. As a last personal view, I would like to state that Academia and Industry interests are by no means in opposition. These kind of advances in AI cannot be achieved by most companies, and current companies developing the methods are too general to be applied to the interests of all fields. AI research is a field which has natural applications to many other fields and by cooperating with companies to understand their interest academics can find new applications and achieve access to interesting resources and materials.

Conclusions

Reflecting on the journey of this thesis, it becomes evident that AI holds transformative potential for phenotyping and genomic prediction in plant research. This thesis has evaluated various AI methodologies, uncovering their strengths and limitations within the context of plant science.

1- Detection and segmentation methods based on convolutional neural networks can be used to renovate software with precise methodology for fruit measurement.
2- The development of open source-based solutions provides new opportunities for researchers and companies in breeding, facilitating the access to useful and editable tools.
3- Detection and segmentation methods coupled with association methods can be used to study growth in Arabidopsis Thaliana at a leaf level, allowing for the study of plant development behavior and architecture.
4- Differential photosynthetic behavior was found on two different genotypes revealing different patterns on photosynthesis at leaf level.
5- The application of autoencoders allows the production of images from genetic markers as long as the genetic markers are relevant to such images.
6- On the other hand, the usage of markers without the certainty that such information is contained, biased generative models to the average image.
7- Incrementing the number of markers used for prediction deteriorates the model performance if the added information is not guaranteed to be relevant to image production.

8- While vector quantization is effective in SNPs, different approaches must be tested to avoid SNPs preselection when conditioning generative models from genetic information.
9- Generative models based on vector quantization and transformers are effective at producing high quality images from genetic information, but the conditioning using genetic markers requires further research to avoid biases towards the average image.

Supplementary Materials

Figures for Chapter 1:

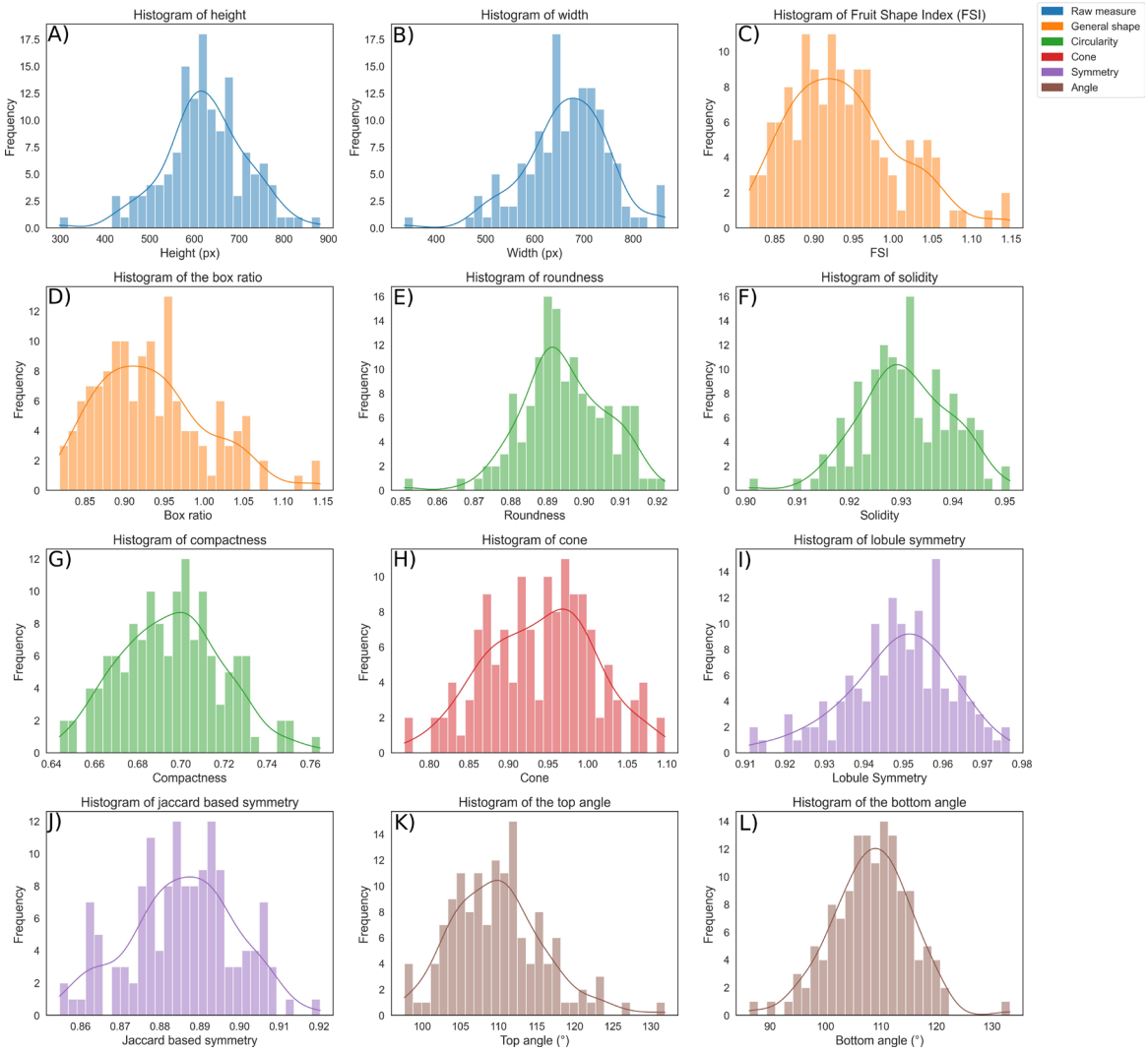


Figure S1.1: Histograms for each of the measures extracted for the 139 genotypes

Figures for Chapter 3:

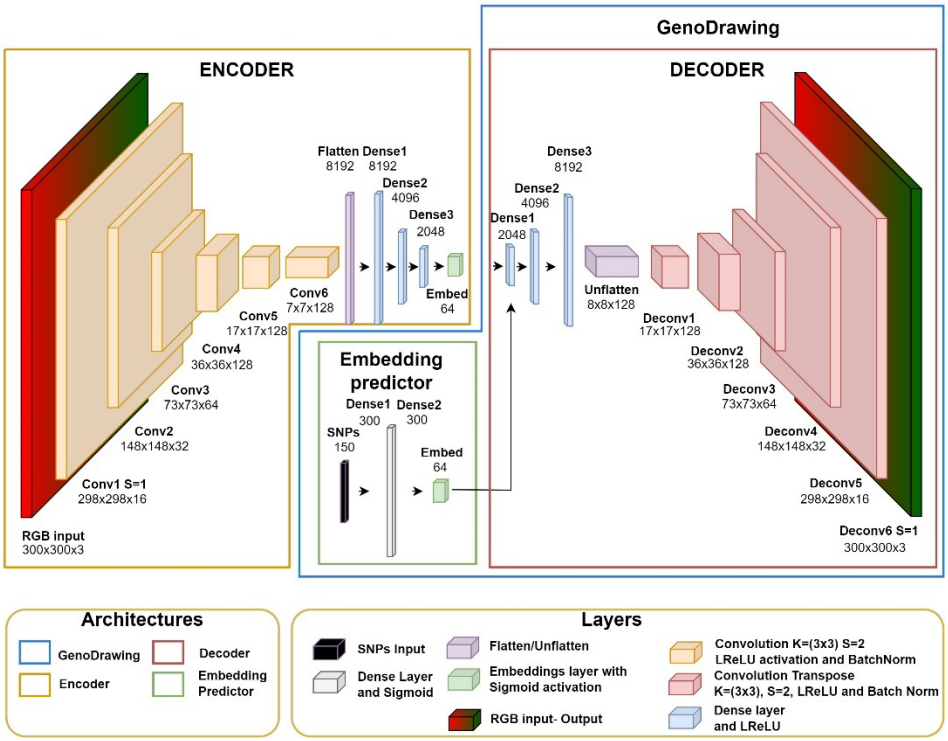


Figure S3.1: Neural networks architecture schema. The autoencoder is split into encoder and decoder. The encoder, orange box, is compound of six convolutions with three final dense layers to encode the image into twelve embeddings. The decoder, red box, uses three dense layers and six deconvolutions to recover the image from the embeddings. The embedding predictor, green box, estimates the embeddings values for every genotype. Together, embedding predictor and encoder, compose GenoDrawing, blue box.

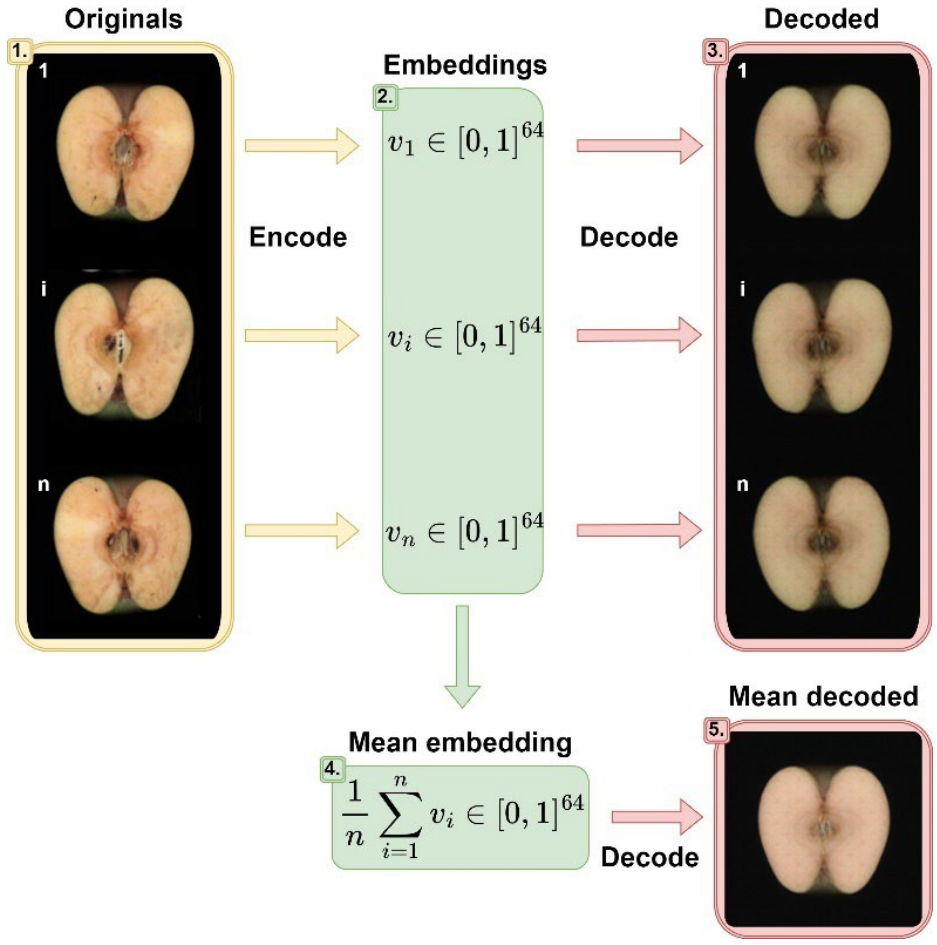


Figure S3.2: The original images (1) were encoded into vectors of size 64 (2) using the encoder. These vectors could be decoded into images (3). Also, the vectors were used to estimate a mean embedding value per genotype (4). The mean embeddings were used to produce a mean image per genotype (5). These mean embedding values were used to train the embedding predictor, and the mean images were used to evaluate the resulting GenoDrawing.

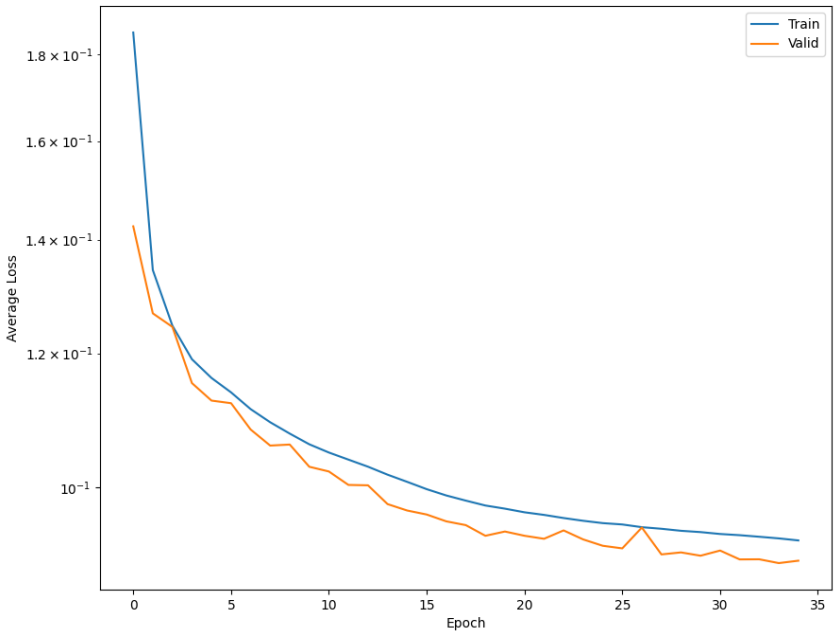


Figure S3.3: Autoencoder perceptual loss during training and validation.

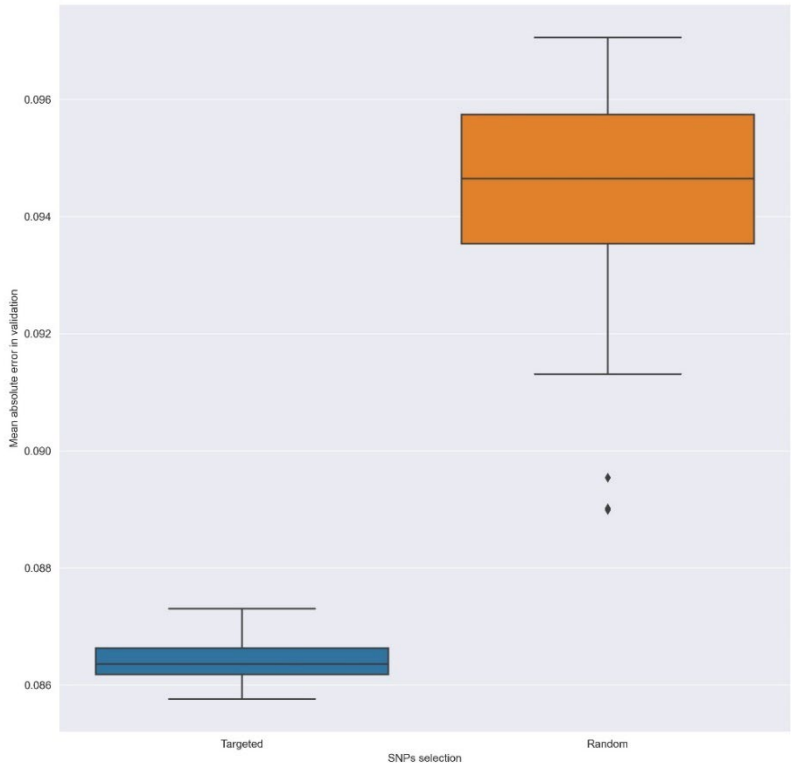


Figure S3.4: Training Losses for both versions of the SNP to embedding predictor. The training was performed one hundred times for each.

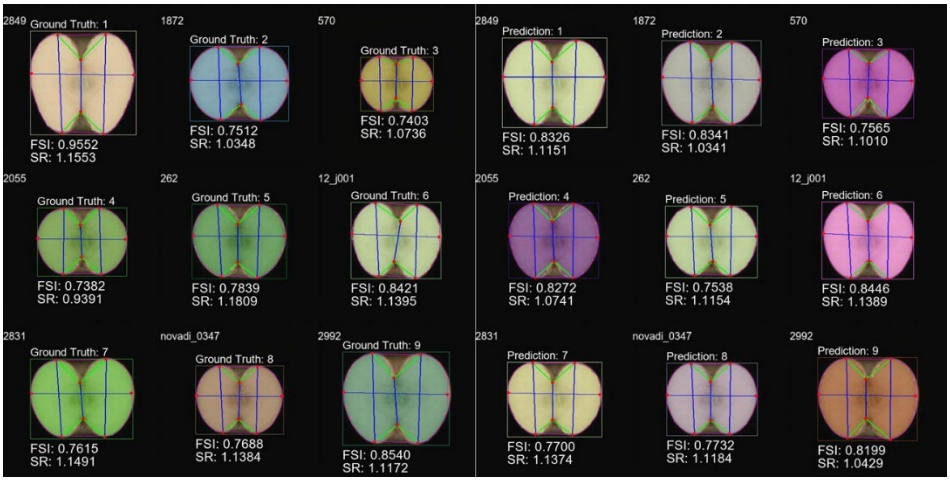


Figure S3.5: Left, nine examples produced through the mean embedding values for the genotype in the validation dataset. Right, the nine corresponding predictions for the GenoDrawing model with relevant-to-shape SNPs. The inner contour detection is displayed in colors; the blue lines are the measures used to calculate to form Fruit Shape Index External (FSI) and Shoulder ratios (SR).

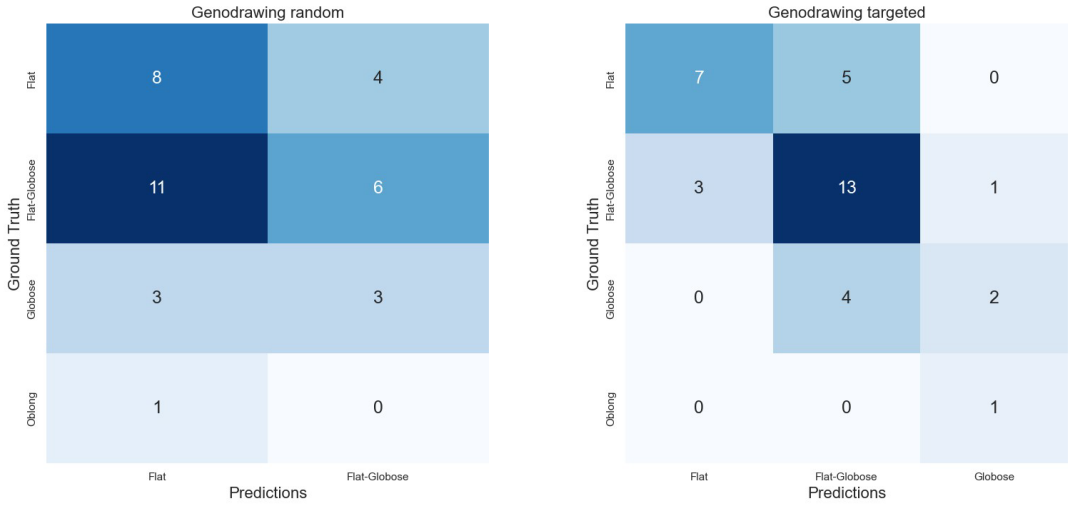


Figure S3.6: Left, the confusion matrix for the classification in the European Cooperative Program for plant Genetic Resource (ECPGR) categories using the predictions generated by the model with a random selection of genomic markers. Right the same classification but using the model with a selection of genomic markers based on a shape GWAS.

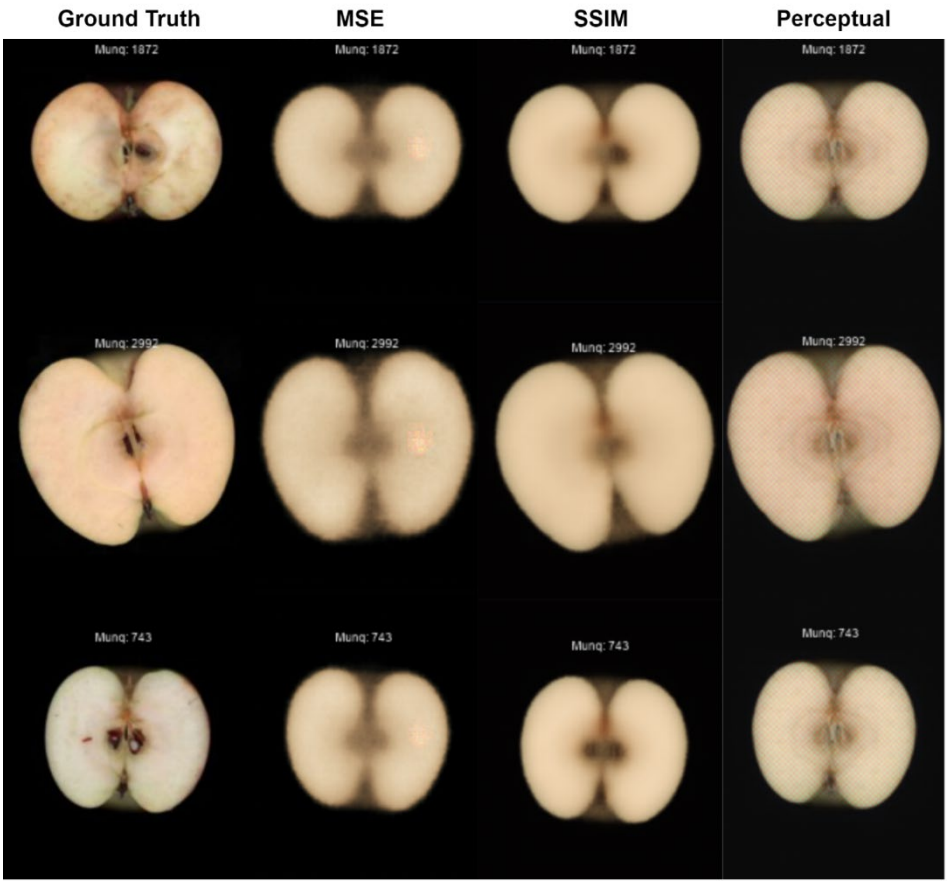


Figure S3.7: Effects of using different losses in the autoencoder training process. The perceptual loss, images at right, allowed for a better capture of patterns and appearance of the images.

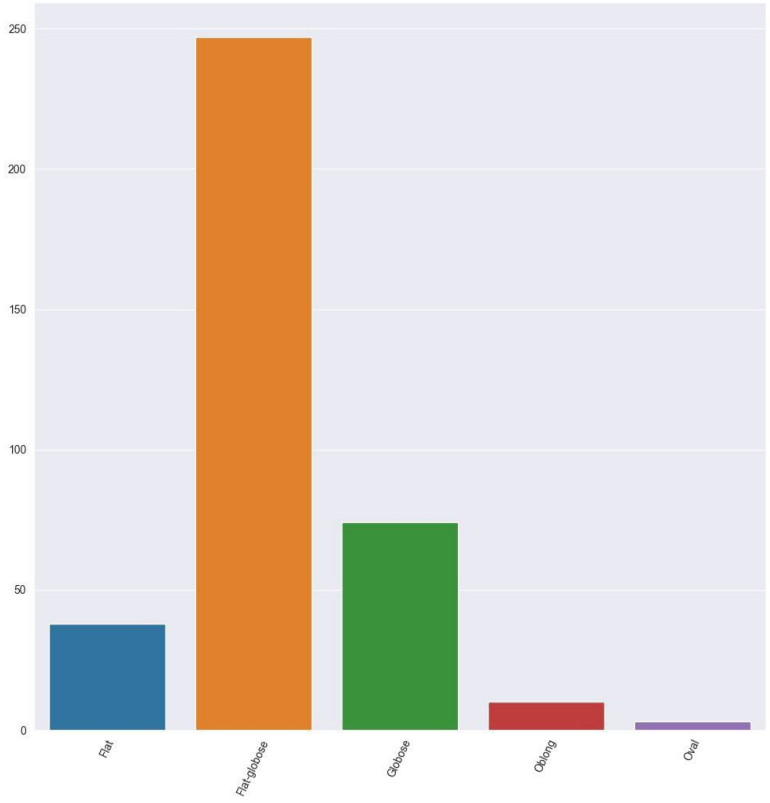


Figure S3.8: Dataset shape categories distribution.

Figures for Chapter 4:

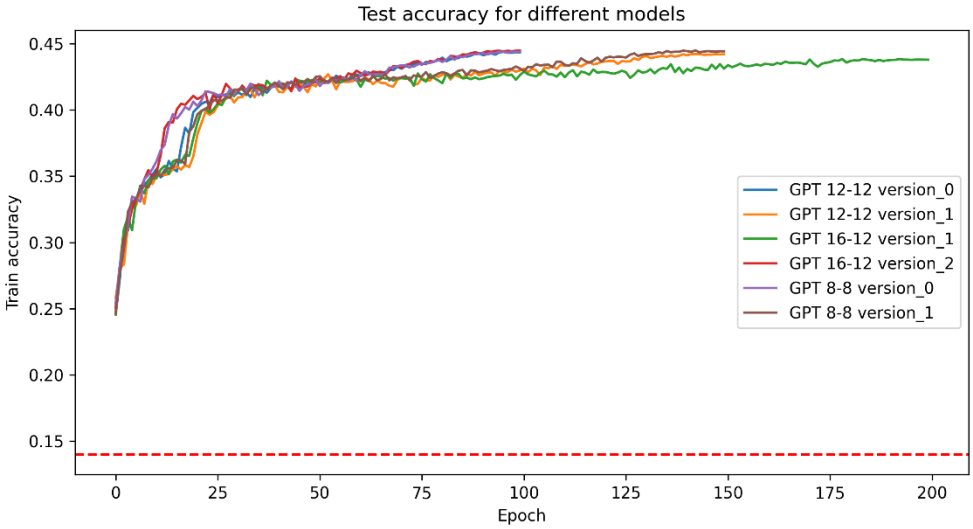


Figure S4.1: Test accuracy across epochs for different all 6 different models. The x-axis represents the number of epochs, and the y-axis represents the test accuracy. The models are compared over 200 epochs, reaching similar accuracy levels with different convergence behaviors

Tables for Chapter 4:

Table S4.1: List of size and shape associated SNPs, extracted from the GWAS analysis of Jung et al. 2022 (https://doi.org/10.1093/hr/uhac028), Dujak et al. 2023. (doi: https://doi.org/10.1101/2023.08.21.554124), and Keller et al. 2023 (https://doi.org/10.1101/2023.08.25.554767)

SNP	Category	Trait	Chromosome	Position	p-value	Reference
AX-115444766	size	Av_SSD_Pol_Abs_B.L_SD	1	15580877	1,87E-10	Keller et al. (2023)
		Av_SSD_X_Abs_A_Q25			7,40E-11	
AX-115625798	size	Av_SSD_X_Abs_B_Q25	1	15868916	3,24E-12	Keller et al. (2023)
AX-115538412	shape	PC5_shape	1	15918343	3,09E-08	Keller et al. (2023)
AX-115489880	shape	DFB	1	17141808	4,35E-08	Dujak et al (2023)
AX-115221494	size	Av_SSD_X_Abs_A_Q25	1	17152361	9,64E-10	Keller et al. (2023)
AX-115341839	size	Av_SAv_Pol_SD_A.LR_Min	1	20696409	1,54E-07	Keller et al. (2023)
AX-115183496	shape	Av_SSD_Pol_SD_B.LR_Q25	1	23578695	1,50E-07	Keller et al. (2023)
AX-115240273	size	Fruit diameter	1	27698606	1,76E-08	Jung et al (2022)
		Fruit length			1,47E-08	
AX-115380250	size	PC1_shape	1	31575453	1,19E-07	Keller et al. (2023)
AX-115378613	size	Av_SAv_Pol_SD_A.LR_Min	2	7626325	1,55E-07	Keller et al. (2023)
AX-115655116	size	Av_SSD_Pol_SD_B.LR_Q25	2	7872781	1,10E-08	Keller et al. (2023)

AX-115490525	shape	PC2_shape	2	7904059	1,20E-07	Keller et al. (2023)
AX-115203615	size	MH	2	12023424	2,74E-08	Dujak et al (2023)
AX-115534322	size	PC1_shape	2	12130252	3,75E-08	Keller et al. (2023)
AX-115314222	size	MH	2	18679518	1,83E-13	Dujak et al (2023)
AX-115240432	size	Maximum fruit size	2	19141996	9,45E-12	Jung et al (2022)
AX-115487044	size	Fruit length	2	19162755	2,17E-08	Jung et al (2022)
AX-115482211	size	A	2	21759081	3,88E-09	Dujak et al (2023)
		MW	2		1,39E-08	
		WMH			9,69E-09	
AX-115353634	size	Maximum fruit size	2	30721980	6,93E-09	Jung et al (2022)
AX-115629265	shape	Av_SSD_Pol_Rat.Av_AB.LR_Max	2	35251284	1,31E-07	Keller et al. (2023)
AX-115388893	size	MH	3	47566	3,09E-08	Dujak et al (2023)
AX-115251897	size	Av_SSD_Pol_Abs_B.L_SD	3	2544800	1,30E-08	Keller et al. (2023)
AX-115451488	shape	DFB	3	6354436	9,41E-09	Dujak et al (2023)
AX-115278645	size	Av_SSD_Pol_Rat.Av_AB.LR_Min	3	8512354	3,20E-08	Keller et al. (2023)
AX-115451037	undefinable	Av_SSD_Pol_Abs_AB_RatM	3	29264257	2,57E-09	Keller et al. (2023)
AX-115481999	size	MW	3	30093454	3,25E-08	Dujak et al (2023)
		WMH	3		1,73E-09	
AX-115520624	size	Av_SSD_Pol_SD_B.LR_Q25	3	32236169	7,39E-08	Keller et al. (2023)
AX-115232080	shape	FSIINT	3	36538297	1,76E-08	Dujak et al (2023)
AX-115194800	size	MH	4	30131460	5,64E-09	Dujak et al (2023)

AX-115506351	size	Av_SAv_Pol_SD_A.LR_Min	4	30194434	5,96E-08	Keller et al. (2023)
AX-115392975	shape	Fruit firmness	5	4949273	3,78E-09	Jung et al (2022)
AX-115459245	shape	Av_SSD_Pol_Rat.Av_AB.LR_Max	5	5031575	7,60E-09	Keller et al. (2023)
AX-115638603	size	MW	5	19798861	2,46E-09	Dujak et al (2023)
AX-115436710	size	WMH	5	19900796	1,42E-07	Dujak et al (2023)
AX-115429431	shape	Av_SSD_Pol_SD_B.LR_Max	5	25209696	2,13E-08	Keller et al. (2023)
		PC2_shape			7,00E-08	
AX-115467342	shape	C	5	36952289	1,36E-07	Dujak et al (2023)
AX-115435503	size	A	5	43940933	2,57E-08	Dujak et al (2023)
AX-115248476	size	WMH	5	44022794	5,20E-08	Dujak et al (2023)
AX-115257677	shape	Av_SAv_XY_Sum_A_MaxR	5	45121659	5,11E-08	Keller et al. (2023)
AX-115644112	shape	Av_SSD_Pol_Rat.Av_AB.LR_SD	5	46356956	1,28E-09	Keller et al. (2023)
AX-115448691	size	MH	6	1063912	2,47E-08	Dujak et al (2023)
AX-115513701	shape	FSIINT	6	1098614	1,70E-07	Dujak et al (2023)
AX-115319264	size	Av_SSD_X_Abs_B_Q25	6	4761377	1,05E-07	Keller et al. (2023)
AX-115378718	shape	C	6	5160566	1,03E-07	Dujak et al (2023)
AX-115556718	size	Maximum fruit size	6	7115267	1,46E-10	Jung et al (2022)
AX-115375194	size	Av_SAv_XY_Sum_A_MaxR	6	8743643	6,44E-14	Keller et al. (2023)
AX-115492426	shape	Av_SSD_Pol_SD_B.LR_Max	6	10584742	9,90E-09	Keller et al. (2023)
AX-115318835	size	PC1_shape	6	20393412	9,86E-09	Keller et al. (2023)
AX-115474007	size	MH	6	23046692	9,99E-10	Dujak et al (2023)

AX-115473970	size	Av_SSD_X_Abs_B_Q25	6	23085308	9,03E-08	Keller et al. (2023)
AX-115378078	size	A	6	35186920	3,64E-10	Dujak et al (2023)
		WMH			3,74E-08	
AX-115430402	size	WMH	6	35271054	8,76E-10	Dujak et al (2023)
AX-115460130	size	PC1_shape	7	3024174	1,52E-07	Keller et al. (2023)
AX-115419204	shape	DFB	7	4975573	1,58E-07	Dujak et al (2023)
AX-115254491	size	Av_SAv_XY_Sum_A_MaxR	7	8648220	3,66E-09	Keller et al. (2023)
AX-115497172	shape	C	7	9057598	1,13E-10	Dujak et al (2023)
AX-115254276	shape	PC3_shape	7	21705520	2,58E-11	Keller et al. (2023)
AX-115277101	shape	R	7	25833427	7,97E-08	Dujak et al (2023)
AX-115397806	shape	PAMa	7	25852956	9,39E-14	Dujak et al (2023)
AX-115395087	size	Fruit diameter	7	30364920	9,09E-12	Jung et al (2022)
AX-115303803	shape	FSII	7	31227220	1,08E-07	Dujak et al (2023)
		CAT-own			8,65E-11	
AX-115457086	shape	Av_SSD_Pol_Abs_AB_RatM	8	537872	1,82E-08	Keller et al. (2023)
AX-115362374	size	Fruit volume	8	2005502	2,59E-08	Jung et al (2022)
AX-115445494	shape	Av_SAv_Pol_Rat.Av_AB.LR_Q50	8	9562443	7,66E-08	Keller et al. (2023)
AX-115629269	size	PC5_shape	8	21161287	3,25E-08	Keller et al. (2023)
AX-115518043	shape	Fruit firmness	8	26876759	4,92E-10	Jung et al (2022)
AX-115359216	size	MW	8	31580496	2,28E-08	Dujak et al (2023)
AX-115524603	shape	C	9	3091269	3,58E-08	Dujak et al (2023)

AX-115650703	shape	FSII	9	4052611	1,45E-09	Dujak et al (2023)
AX-115289134	size	Fruit diameter	9	4504072	1,06E-09	Jung et al (2022)
AX-115442384	size	Av_SSD_X_Abs_A_Q25	9	4698417	1,19E-08	Keller et al. (2023)
AX-115521252	size	Av_SSD_X_Abs_B_Q25	9	5170431	2,84E-08	Keller et al. (2023)
AX-115391247	shape	Av_SSD_Pol_SD_B.LR_Max	9	31184244	7,35E-08	Keller et al. (2023)
AX-115500353	shape	Av_SSD_X_Abs_B_Q25	9	33000755	6,82E-08	Keller et al. (2023)
AX-115507212	shape	PC3_shape	9	33255077	2,48E-08	Keller et al. (2023)
AX-115349165	shape	Av_SSD_Pol_Rat.Av_AB.LR_Min	10	15436508	1,59E-10	Keller et al. (2023)
AX-115419379	shape	PC4_shape	10	17661953	4,24E-08	Keller et al. (2023)
AX-115404241	shape	PC3_shape	10	19586733	2,20E-08	Keller et al. (2023)
AX-115494202	size	Fruit length	10	26445213	9,47E-09	Jung et al (2022)
AX-115552368	shape	PC2_shape	10	33469975	5,50E-09	Keller et al. (2023)
AX-115184639	shape	Av_SSD_Pol_SD_B.LR_Q25	10	33907766	6,40E-08	Keller et al. (2023)
AX-115325098	size	Fruit diameter	10	34563543	1,11E-08	Jung et al (2022)
AX-115299520	size	Maximum fruit size	10	35060486	1,43E-07	Jung et al (2022)
AX-115227580	size	PC1_shape	10	35210847	1,62E-07	Keller et al. (2023)
AX-115302154	size	Fruit diameter	10	37712000	1,25E-08	Jung et al (2022)
AX-115330695	shape	PC2_shape	11	3241944	1,93E-09	Keller et al. (2023)
AX-115421714	shape	Av_SSD_Pol_SD_B.LR_Q25	11	3408716	1,06E-08	Keller et al. (2023)
AX-115327898	shape	FSIINT	11	4699021	3,16E-11	Dujak et al (2023)
AX-115327900	shape	CAT-own	11	4703926	7,08E-13	Dujak et al (2023)

AX-115327911	size	Av_SSD_Pol_Abs_AB_RatM	11	4706860	4,19E-15	Keller et al. (2023)
AX-115335214	shape	FSII	11	4947462	1,40E-11	Dujak et al (2023)
AX-105204673	shape	FSIINT	11	5934614	5,97E-13	Dujak et al (2023)
AX-115216348	shape	Fruit firmness	11	6195430	8,02E-08	Jung et al (2022)
AX-115449727	shape	Av_SSD_Pol_SD_B.LR_Max	11	6367847	1,83E-08	Keller et al. (2023)
AX-105211625	size	MH	11	9396684	5,61E-09	Dujak et al (2023)
AX-115659654	shape	PC4_shape	11	12416234	1,35E-07	Keller et al. (2023)
AX-115510709	shape	C	11	12642154	8,06E-10	Dujak et al (2023)
AX-115479723	size	MW	11	13039227	6,84E-08	Dujak et al (2023)
AX-115451348	size	WMH	11	13196378	1,35E-08	Dujak et al (2023)
AX-115464400	size	WMH	11	14369160	1,65E-08	Dujak et al (2023)
AX-115380060	size	Fruit length	11	14585569	2,41E-09	Jung et al (2022)
AX-115436422	shape	Av_SSD_Pol_Abs_B.L_SD	11	15841357	9,29E-10	Keller et al. (2023)
AX-115414638	size	Maximum fruit size	11	17878463	8,13E-13	Jung et al (2022)
AX-115394557	size	Fruit volume	11	18521895	1,62E-12	Jung et al (2022)
		Fruit diameter			6,94E-09	
AX-115482171	shape	Av_SSD_Pol_Rat.Av_AB.LR_Min	11	33111593	1,77E-09	Keller et al. (2023)
AX-115440676	size	Av_SSD_Pol_Abs_B.L_SD	11	34816193	2,22E-10	Keller et al. (2023)
AX-105213957	shape	FSII	11	37649389	6,43E-09	Dujak et al (2023)
AX-115300506	shape	Av_SAv_Pol_Rat.Av_AB.LR_Q50	11	38069491	9,69E-10	Keller et al. (2023)
AX-115553037	shape	FSII	11	38174120	3,80E-08	Dujak et al (2023)

AX-115265526	shape	PC4_shape	11	41448132	1,26E-11	Keller et al. (2023)
AX-115476093	size	PC4_shape	12	9098303	3,16E-08	Keller et al. (2023)
AX-115517804	shape	FSII	12	11952215	1,51E-08	Dujak et al (2023)
AX-115378066	shape	DFB	12	26084416	5,57E-12	Dujak et al (2023)
AX-115647469	shape	PAMa	13	470683	3,71E-10	Dujak et al (2023)
AX-115250714	size	Maximum fruit size	13	835641	1,16E-08	Jung et al (2022)
AX-115355048	shape	C	13	5186503	4,02E-09	Dujak et al (2023)
AX-115548370	shape	PC2_shape	13	5411038	9,11E-09	Keller et al. (2023)
AX-115634660	shape	Fruit firmness	13	10004542	6,86E-08	Jung et al (2022)
AX-115241968	size	A	13	10603957	5,19E-10	Dujak et al (2023)
AX-115389297	shape	Av_SSD_Pol_SD_R_Sym.rel	13	13849729	9,64E-10	Keller et al. (2023)
AX-115287563	shape	Av_SSD_Pol_Rat.Av_AB.LR_Max	13	14542354	6,22E-08	Keller et al. (2023)
AX-115604905	size	WMH	13	17503469	1,07E-07	Dujak et al (2023)
AX-115336086	shape	FSII	14	291347	4,91E-08	Dujak et al (2023)
AX-115487829	shape	PAMa	14	1446777	3,32E-09	Dujak et al (2023)
AX-115628310	size	PC1_shape	14	1478099	3,03E-09	Keller et al. (2023)
AX-115295220	size	MH	14	1587955	1,88E-08	Dujak et al (2023)
AX-115286628	size	Fruit diameter	14	3500707	5,25E-09	Jung et al (2022)
AX-115363036	size	Av_SSD_X_Abs_A_Q25	14	15862103	4,55E-08	Keller et al. (2023)
AX-115311704	size	Av_SSD_Pol_Rat.Av_AB.LR_Min	14	21571993	7,58E-10	Keller et al. (2023)
AX-115295642	size	MW	14	23185565	2,37E-08	Dujak et al (2023)

		WMH			5,79E-10	
AX-115345243	size	PC1_shape	14	24020772	2,87E-09	Keller et al. (2023)
AX-115517142	shape	FSII	14	24516722	1,11E-09	Dujak et al (2023)
AX-115189205	size	Fruit diameter	14	29606920	1,91E-08	Jung et al (2022)
AX-115250754	size	A	14	30906712	1,75E-08	Dujak et al (2023)
AX-115461813	shape	Av_SSD_Pol_SD_R_Sym.rel	15	10680808	3,28E-10	Keller et al. (2023)
AX-115646188	size	Fruit diameter	15	17136643	2,09E-08	Jung et al (2022)
AX-115631329	shape	Av_SAv_Pol_Rat.Av_AB.LR_Q50	15	47545121	7,77E-08	Keller et al. (2023)
AX-115283778	shape	Av_SSD_Pol_SD_B.LR_Max	15	48233842	2,43E-08	Keller et al. (2023)
AX-115419024	size	Maximum fruit size	15	52885708	1,21E-07	Jung et al (2022)
AX-115344575	shape	FSIINT	16	1246746	6,38E-10	Dujak et al (2023)
AX-115600806	shape	R	16	3161613	3,11E-08	Dujak et al (2023)
AX-115235649	size	Av_SAv_Pol_Rat.Av_AB.LR_Q50	16	3230048	1,93E-10	Keller et al. (2023)
AX-115515574	shape	Av_SSD_Pol_SD_R_Sym.rel	16	3452384	2,33E-08	Keller et al. (2023)
AX-115317103	shape	PAMa	16	19284012	4,57E-08	Dujak et al (2023)
AX-115312607	size	MW	17	2796935	3,49E-11	Dujak et al (2023)
		WMH			6,50E-11	
AX-115659724	size	A	17	3349151	2,32E-10	Dujak et al (2023)
AX-115267295	size	A	17	9572628	7,56E-08	Dujak et al (2023)
AX-115375737	size	Av_SSD_Pol_Abs_B.L_SD	17	9806448	1,23E-07	Keller et al. (2023)
AX-115500929	shape	FSIINT	17	10745180	5,10E-08	Dujak et al (2023)

		C			1,53E-09	
AX-115332011	shape	FSII	17	25914860	9,39E-08	Dujak et al (2023)

Table S4.2: Wasserstein distance between distributions based on the Fruit Shape Index (FSI) and Shoulder Ratio (SR) for four different cases: Targeted, Random, Decoded Mean, and Genotype Mean. The diagonal shows the distance between identical distributions, and the upper/lower triangles shows the distances between different distributions.

FSI	**Targeted**	**Random**	**Decoded mean**	**Genotype mean**
Targeted	0	0.0505	0.0209	0.0259
Random		0	0.0469	0.0331
Decoded mean			0	0.0144
Genotype mean				0
SR	**Targeted**	**Random**	**Decoded mean**	**Genotype mean**
Targeted	0	0.0108	0.0184	0.0255
Random		0	0.0247	0.0233
decoded mean			0	0.0175
Genotype mean				0

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, … Xiaoqiang Zheng. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. https://www.tensorflow.org/

Abdollahi-Arpanahi, R., Gianola, D., & Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, *52*(1), 12. https://doi.org/10.1186/s12711-020-00531-z

Allaby, R. G., Stevens, C. J., Kistler, L., & Fuller, D. Q. (2022). Emerging evidence of plant domestication as a landscape-level process. *Trends in Ecology & Evolution*, *37*(3), 268-279. https://doi.org/10.1016/j.tree.2021.11.002

Ayalew, H., Tsang, P. W., Chu, C., Wang, J., Liu, S., Chen, C., & Ma, X.-F. (2019). Comparison of TaqMan, KASP and rhAmp SNP genotyping platforms in hexaploid wheat. *PLOS ONE*, *14*(5), 1-9. https://doi.org/10.1371/journal.pone.0217222

B. Ding, H. Qian, & J. Zhou. (2018). Activation functions and their characteristics in deep neural networks. *2018 Chinese Control And Decision Conference (CCDC)*, 1836-1841. https://doi.org/10.1109/CCDC.2018.8407425

Baker, N. R. (2008). Chlorophyll fluorescence: A probe of photosynthesis in vivo. *Annual Review of Plant Biology*, *59*(1), 89-113. https://doi.org/10.1146/annurev.arplant.59.032607.092759

Barbole, D. K., Jadhav, P. M., & Patil, S. B. (2022). A Review on Fruit Detection and Segmentation Techniques in Agricultural Field. En J. I.-Z. Chen, J. M. R. S. Tavares, A. M. Iliyasu, & K.-L. Du (Eds.), *Second International Conference on Image Processing and Capsule Networks* (pp. 269-288). Springer International Publishing.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bell, J., & Dee, H. M. (2019). Leaf segmentation through the

classification of edges. *arXiv*. https://doi.org/10.48550/arxiv.1904.03124

Benfenati, A., Bolzi, D., Causin, P., & Oberti, R. (2022). A deep learning generative model approach for image synthesis of plant leaves. *PLOS ONE*, *17*(11), e0276972. https://doi.org/10.1371/journal.pone.0276972

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53-59. https://doi.org/10.1038/nature07517

Bernardo P. de Almeida, Hugo Dalla-Torre, Guillaume Richard, Christopher Blum, Lorenz Hexemer, Maxence Gelard, Priyanka Pandey, Stefan Laurent, Alexandre Laterre, Maren Lang, Ugur Sahin, Karim Beguir, & Thomas Pierrot. (2024, marzo 14). *SegmentNT: annotating the genome at single-nucleotide resolution with DNA foundation models*. https://www.instadeep.com/research/paper/segmentnt-annotating-the-genome-at-single-nucleotide-resolution-with-dna-foundation-models/

Bharati, P., & Pramanik, A. (2020). Deep learning techniques, r-CNN to mask r-CNN: a survey. *Computational Intelligence in Pattern Recognition*, 657-668.

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P., & Prabhu, K. V. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Frontiers in Genetics*, *7*(DEC), 221. https://doi.org/10.3389/FGENE.2016.00221/BIBTEX

Bielczynski, L. W., Lacki, M. K., Hoefnagels, I., Gambin, A., & Croce, R. (2017). Leaf and plant age affects photosynthetic performance and photoprotective capacity. *Plant Physiology*, *175*(4), 1634-1648. https://doi.org/10.1104/pp.17.00904

Bloemen, M., Rector, A., Swinnen, J., Ranst, M. V., Maes, P., Vanmechelen, B., & Wollants, E. (2022). Fast detection of SARS-CoV-2 variants including Omicron using one-step RT-PCR and Sanger sequencing. *Journal of Virological Methods*, *304*, 114512. https://doi.org/10.1016/j.jviromet.2022.114512

Boursianis, A. D., Papadopoulou, M. S., Diamantoulakis, P., Liopa-Tsakalidi, A., Barouchas, P., Salahas, G., Karagiannidis, G., Wan, S., & Goudos, S. K. (2022). Internet of Things (IoT) and Agricultural Unmanned Aerial Vehicles (UAVs) in smart farming: A comprehensive review. *Internet of Things*, *18*, 100187. https://doi.org/10.1016/j.iot.2020.100187

Bowles, S., & Choi, J.-K. (2019). The Neolithic Agricultural Revolution and the Origins of Private Property. *Journal of Political Economy*, *127*(5), 2186-2228. https://doi.org/10.1086/701789

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/2005.14165v4

Chawade, A., van Ham, J., Blomquist, H., Bagge, O., Alexandersson, E., & Ortiz, R. (2019). High-Throughput Field-Phenotyping Tools for Plant Breeding and Precision Agriculture. *Agronomy*, *9*(5). https://doi.org/10.3390/agronomy9050258

Chen, Y.-E., Zhang, C.-M., Su, Y.-Q., Ma, J., Zhang, Z.-W., Yuan, M., Zhang, H.-Y., & Yuan, S. (2017). Responses of photosystem II and antioxidative systems to high light and high temperature co-stress in wheat. *Environmental and Experimental Botany*, *135*, 45-55. https://doi.org/10.1016/j.envexpbot.2016.12.001

Crossley, B. M., Bai, J., Glaser, A., Maes, R., Porter, E., Killian, M. L., Clement, T., & Toohey-Kurth, K. (2020). Guidelines for Sanger sequencing and molecular assay monitoring. *J Vet Diagn Invest*, *32*(6), 767-775. https://doi.org/10.1177/1040638720905833

Czibula, G., Codre, C., & Teletin, M. (2021). AnomalP: An approach for detecting anomalous protein conformations using deep autoencoders. *Expert Systems with Applications*, *166*, 114070. https://doi.org/10.1016/j.eswa.2020.114070

D'Agostino, N., & Tripodi, P. (2017). NGS-Based Genotyping, High-Throughput Phenotyping and Genome-Wide Association Studies Laid the Foundations for Next-Generation Breeding in Horticultural Crops. *Diversity*, *9*(3). https://doi.org/10.3390/d9030038

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L.,

Grzywaczewski, A. H., Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, Guillaume Richard, Marcin Skwark, Karim Beguir, Marie Lopez, & Thomas Pierrot. (2023). The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, 2023.01.11.523679. https://doi.org/10.1101/2023.01.11.523679

Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., Ramsay, L. D., Halpin, C., Mascher, M., Fleury, D. L., Langridge, P., Stein, N., & Waugh, R. (2019). A Comparison of Mainstream Genotyping Platforms for the Evaluation and Use of Barley Genetic Resources. *Frontiers in Plant Science*, *10*, 544. https://doi.org/10.3389/fpls.2019.00544

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PloS One*, *16*(10), e0257521. https://doi.org/10.1371/journal.pone.0257521

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241-258. https://doi.org/10.1007/s11704-019-8208-z

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. https://doi.org/10.48550/arXiv.2010.11929

Douarre, C., Crispim-Junior, C. F., Gelibert, A., Tougne, L., & Rousseau, D. (2019). Novel data augmentation strategies to boost supervised segmentation of plant disease. *Computers and Electronics in Agriculture*, *165*, 104967. https://doi.org/10.1016/j.compag.2019.104967

Dujak, C., Coleto-Alcudia, V., & Aranzana, M. J. (2024). Genomic analysis of fruit size and shape traits in apple: Unveiling candidate genes through GWAS analysis. *Horticulture Research*, *11*(2), uhad270. https://doi.org/10.1093/hr/uhad270

Dujak, C., Jurado-Ruiz, F., & Aranzana, M. J. (2023). Comprehensive Morphometric Analysis of Apple Fruits and Weighted Class Assignation using Machine Learning. *Research Square*. https://doi.org/10.21203/rs.3.rs-2860631/v1

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., deWinter, A., Dixon, J.,

… Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, *323*(5910), 133-138. https://doi.org/10.1126/science.1162986

El-Lithy, M. E., Rodrigues, G. C., van Rensen, J. J. S., Snel, J. F. H., Dassen, H. J. H. A., Koornneef, M., Jansen, M. A. K., Aarts, M. G. M., & Vreugdenhil, D. (2005). Altered photosynthetic performance of a natural Arabidopsis accession is associated with atrazine resistance. *Journal of Experimental Botany*, *56*(416), 1625-1634.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), 1-10. https://doi.org/10.1371/journal.pone.0019379

Ershov, D., Phan, M.-S., Pylvänäinen, J. W., Rigaud, S. U., Blanc, L. L., Charles-Orszag, A., Conway, J. R. W., Laine, R. F., Roy, N. H., Bonazzi, D., Duménil, G., Jacquemet, G., & Tinevez, J.-Y. (2022). TrackMate 7: Integrating state-of-the-art segmentation algorithms into tracking pipelines. *Nature Methods*, *19*(7), 829-832. https://doi.org/10.1038/s41592-022-01507-1

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873-12883.

Everson, E. H., & Schaller, C. W. (1955). The Genetics of Yield Differences Associated With Awn Barbing in the Barley Hybrid (Lion × Atlas10) × Atlas1. *Agronomy Journal*, *47*(6), 276-280. https://doi.org/10.2134/agronj1955.00021962004700060010x

Ferrández-Pastor, F. J., Cámara-Zapata, J. M., Alcañiz-Lucas, S., Pardo, S., & Brenes, J. A. (2023). Reinforcement Learning Model in Automated Greenhouse Control. En J. Bravo & G. Urzáiz (Eds.), *Proceedings of the 15th International Conference on Ubiquitous Computing & Ambient Intelligence (UCAmI 2023)* (pp. 3-13). Springer Nature Switzerland.

Flood, P. J., Kruijer, W., Schnabel, S. K., van der Schoor, R., Jalink, H., Snel, J. F. H., Harbinson, J., & Aarts, M. G. M. (2016). Phenomics for photosynthesis, growth and reflectance in Arabidopsis thaliana reveals circadian and long-term fluctuations in heritability. *Plant Methods*, *12*(1), 14. https://doi.org/10.1186/s13007-016-0113-y

Flood, P. J., Theeuwen, T. P. J. M., Schneeberger, K., Keizer, P., Kruijer, W., Severing, E., Kouklas, E., Hageman, J. A., Wijfjes, R., Calvo-Baltanas, V., Becker, F. F. M., Schnabel, S. K., Willems, L. A. J., Ligterink, W., van Arkel, J., Mumm, R., Gualberto, J. M., Savage, L., Kramer, D. M., … Wijnker, E. (2020). Reciprocal cybrids reveal how organellar genomes affect plant phenotypes. *Nature Plants*, *6*(1), 13-21. https://doi.org/10.1038/s41477-019-0575-9

Fonteijn, H., Afonso, M., Lensink, D., Mooij, M., Faber, N., Vroegop, A., Polder, G., & Wehrens, R. (2021). Automatic Phenotyping of Tomatoes in Production Greenhouses Using Robotics and Computer Vision: From Theory to Practice. *Agronomy*, *11*(8). https://doi.org/10.3390/agronomy11081599

Franklin, R. E., & Gosling, R. G. (1953). Molecular Configuration in Sodium Thymonucleate. *Nature*, *171*(4356), 740-741. https://doi.org/10.1038/171740a0

Frey, L. J. (2019). Artificial Intelligence and Integrated Genotype–Phenotype Identification. *Genes*, *10*(1). https://doi.org/10.3390/genes10010018

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193-202. https://doi.org/10.1007/BF00344251

G. Raskutti, M. J. Wainwright, & B. Yu. (2011). Early stopping for non-parametric regression: An optimal data-dependent stopping rule. *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1318-1325. https://doi.org/10.1109/Allerton.2011.6120320

Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. Batch normalization: An empirical study of their impact to deep learning. *Multimedia Tools and Applications*, *79*(19), 12777-12815. https://doi.org/10.1007/s11042-019-08453-9

Gené-Mola, J., Ferrer-Ferrer, M., Gregorio, E., Blok, P. M., Hemming, J., Morros, J.-R., Rosell-Polo, J. R., Vilaplana, V., & Ruiz-Hidalgo, J. (2023). Looking behind occlusions: A study on amodal segmentation for robust on-tree apple fruit size estimation. *Computers and Electronics in Agriculture*, *209*, 107854. https://doi.org/10.1016/j.compag.2023.107854

Girshick, R. (2015, diciembre). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT

Press.
Gupta, P. K., Kulwal, P. L., & Jaiswal, V. (2019). Chapter Two—Association mapping in plants in the post-GWAS genomics era. En D. Kumar (Ed.), *Advances in Genetics* (Vol. 104, pp. 75-154). Academic Press. https://doi.org/10.1016/bs.adgen.2018.12.001
H. I. Tangarife & A. E. Díaz. (2017). Robotic applications in the automation of agricultural production under greenhouse: A review. *2017 IEEE 3rd Colombian Conference on Automatic Control (CCAC)*, 1-6. https://doi.org/10.1109/CCAC.2017.8276478
Hadipour, H., Liu, C., Davis, R., Cardona, S. T., & Hu, P. (2022). Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. *BMC Bioinformatics*, *23*(4), 132. https://doi.org/10.1186/s12859-022-04667-1
Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2
Harshvardhan, G., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, *38*, 100285. https://doi.org/10.1016/J.COSREV.2020.100285
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask R-CNN*. arXiv. https://doi.org/10.48550/ARXIV.1703.06870
He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. arXiv. https://doi.org/10.48550/ARXIV.1512.03385
He, K., Zhang, X., Ren, S., & Sun, J. (2016, junio). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.
Herrera, R. J., & Garcia-Bertrand, R. (2018). Chapter 13—The Agricultural Revolutions. En R. J. Herrera & R. Garcia-Bertrand (Eds.), *Ancestral DNA, Human Origins, and Migrations* (pp. 475-509). Academic Press. https://doi.org/10.1016/B978-0-12-804124-6.00013-6
Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., & Zamir, A. (1965). Structure of a

Ribonucleic Acid. *Science*, *147*(3664), 1462-1465. https://doi.org/10.1126/science.147.3664.1462

Holley, R. W., Everett, G. A., Madison, J. T., & Zamir, A. (1965). Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid. *Journal of Biological Chemistry*, *240*(5), 2122-2128. https://doi.org/10.1016/S0021-9258(18)97435-1

I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, & N. C. Bouaynaya. (2022). Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, *39*(4), 73-84. https://doi.org/10.1109/MSP.2022.3142719

Iqbal, J., Xu, R., Sun, S., & Li, C. (2020). Simulation of an Autonomous Mobile Robot for LiDAR-Based In-Field Phenotyping and Navigation. *Robotics*, *9*(2). https://doi.org/10.3390/robotics9020046

Itakura, K., & Hosoi, F. (2018). Automatic leaf segmentation for estimating leaf area and leaf inclination angle in 3D plant images. *Sensors*, *18*(10), Article 3576. https://doi.org/10.3390/s18103576

Iwata, H., Ebana, K., Uga, Y., & Hayashi, T. (2015). Genomic Prediction of Biological Shape: Elliptic Fourier Analysis and Kernel Partial Least Squares (PLS) Regression Applied to Grain Shape Prediction in Rice (Oryza sativa L.). *PLOS ONE*, *10*(3), e0120610. https://doi.org/10.1371/journal.pone.0120610

Iwata, H., Minamikawa, M. F., Kajiya-Kanegae, H., Ishimori, M., & Hayashi, T. (2016). Genomics-assisted breeding in fruit trees. *Breeding Science*, *66*(1), 100-115. https://doi.org/10.1270/jsbbs.66.100

J. Gui, Z. Sun, Y. Wen, D. Tao, & J. Ye. (2023). A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3313-3332. https://doi.org/10.1109/TKDE.2021.3130191

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Jangra, S., Chaudhary, V., Yadav, R. C., & Yadav, N. R. (2021). High-Throughput Phenotyping: A Platform to Accelerate Crop Improvement. *Phenomics*, *1*, 31-53. https://doi.org/10.1007/s43657-020-00007-6

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics*, *135*(1), 205-211. https://doi.org/10.1093/genetics/135.1.205

Jez, J. M., Topp, C. N., Yao, L., van de Zedde, R., & Kowalchuk, G. (2021). Recent developments and potential of robotics in plant eco-phenotyping. *Emerging Topics in Life Sciences*, *5*(2), 289-300. https://doi.org/10.1042/ETLS20200275

Ji, Y., Kusuma, P., & Marcelis, L. F. M. (2023). Vertical farming. *Current Biology*, *33*(11), R471-R473. https://doi.org/10.1016/j.cub.2023.02.010

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, *37*(15), 2112-2120. https://doi.org/10.1093/bioinformatics/btab083

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo Algorithm Developments. *The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy after COVID-19*, *199*, 1066-1073. https://doi.org/10.1016/j.procs.2022.01.135

Jiang, S., Cheng, Q., Yan, J., Fu, R., & Wang, X. (2020). Genome optimization for improvement of maize breeding. *Theoretical and Applied Genetics*, *133*(5), 1491-1502. https://doi.org/10.1007/s00122-019-03493-z

Jones, P. M. (2016). Agriculture, Industry and Commerce. En *Agricultural Enlightenment: Knowledge, Technology and Nature.* (pp. 14-31). Oxford University Press.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., Al-Rifaï, M., Lewandowski, M., Sanin, N., Rymenants, M., Didelot, F., Dujak, C., Font i Forcada, C., Knauf, A., Laurens, F., Studer, B., Muranty, H., & Patocchi, A. (2022). Genetic architecture and genomic predictive ability of apple quantitative traits across environments. *Horticulture Research*, *9*.

https://doi.org/10.1093/hr/uhac028
Jung, M., Roth, M., Aranzana, M. J., Auwerkerken, A., Bink, M., Denancé, C., Dujak, C., Durel, C.-E., i Forcada, C. F., Cantin, C. M., Guerra, W., Howard, N. P., Keller, B., Lewandowski, M., Ordidge, M., Rymenants, M., Sanin, N., Studer, B., Zurawicz, E., … Muranty, H. (2020). The apple REFPOP—a reference population for genomics-assisted breeding in apple. *Horticulture Research 2020 7:1*, *7*(1), 1-16. https://doi.org/10.1038/s41438-020-00408-8
Jurado-Ruiz, F., Rousseau, D., Botía, J. A., & Aranzana, M. J. (2023). GenoDrawing: An Autoencoder Framework for Image Prediction from SNP Markers. *Plant Phenomics*, *5*, 0113. https://doi.org/10.34133/plantphenomics.0113
Kalaji, H. M., Carpentier, R., Allakhverdiev, S. I., & Bosa, K. (2012). Fluorescence parameters as early indicators of light stress in barley. *Journal of Photochemistry and Photobiology B: Biology*, *112*, 1-6. https://doi.org/10.1016/j.jphotobiol.2012.03.009
Kaynak, O. (2021). The golden age of Artificial Intelligence. *Discover Artificial Intelligence*, *1*(1), 1. https://doi.org/10.1007/s44163-021-00009-x
Keller, B., Jung, M., Bühlmann-Schütz, S., Hodel, M., Studer, B., Broggini, G. A. L., & Patocchi, A. (2024). The genetic basis of apple shape and size unraveled by digital phenotyping. *G3 Genes|Genomes|Genetics*, *14*(5), jkae045. https://doi.org/10.1093/g3journal/jkae045
Khan, S. U., Saeed, S., Khan, M. H. U., Fan, C., Ahmar, S., Arriagada, O., Shahzad, R., Branca, F., & Mora-Poblete, F. (2021). Advances and Challenges for QTL Analysis and GWAS in the Plant-Breeding of High-Yielding: A Focus on Rapeseed. *Biomolecules*, *11*(10). https://doi.org/10.3390/biom11101516
Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. https://arxiv.org/abs/1412.6980
Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. *arXiv*. https://doi.org/10.48550/arXiv.2304.02643
Kumar, J. P., & Domnic, S. (2019). Image based leaf segmentation and counting in rosette plants. *Information Processing in Agriculture*, *6*(2), 233-246. https://doi.org/10.1016/j.inpa.2018.09.005
L. Wang, X. He, & D. Luo. (2020). Deep Reinforcement Learning for Greenhouse Climate Control. *2020 IEEE International Conference*

on Knowledge Graph (ICKG), 474-480. https://doi.org/10.1109/ICBK50248.2020.00073

Lateur, M., Dapena, E., Szalatnay, D., Guyader, A., Hjalmarsson, I., Höfer, M., Militaru, M., Jiménez, C. M., Osterc, G., Rondia, A., & others. (2022). *ECPGR Characterization and Evaluation Descriptors for Apple Genetic Resources: Apple (Malus X Domestica)*. ECPGR-European Cooperative Programme for Plant Genetic Resources.

Leapman, M. (2012). *The Ingenious Mr Fairchild: The Forgotten Father of the Flower Garden*. Faber & Faber.

Li, D., Cao, Y., Shi, G., Cai, X., Chen, Y., Wang, S., & Yan, S. (2019). An overlapping-free leaf segmentation method for plant point clouds. *IEEE access : practical innovations, open solutions*, *7*, 129054-129070. https://doi.org/10.1109/ACCESS.2019.2940385

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999-7019. https://doi.org/10.1109/TNNLS.2021.3084827

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European conference on computer vision*, 740-755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. En B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 21-37). Springer International Publishing.

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics*, *12*(2), e1005767. https://doi.org/10.1371/journal.pgen.1005767

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., & Leibe, B. (2021). HOTA: a higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, *129*(2), 548-578. https://doi.org/10.1007/s11263-020-01375-2

Malusare, A., Kothandaraman, H., Tamboli, D., Lanman, N. A., & Aggarwal, V. (2024). *Understanding the Natural Language of*

DNA using Encoder-Decoder Foundation Models with Byte-level Precision.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56-61. https://doi.org/10.25080/Majora-92bf1922-00a

Mendel, Gregor. (1865). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, *Bd.4*, 3-47.

Min, B., Kim, T., Shin, D., & Shin, D. (2023). Data Augmentation Method for Plant Leaf Disease Recognition. *Applied Sciences*, *13*(3). https://doi.org/10.3390/app13031465

Mishra, P., Sadeh, R., Ryckewaert, M., Bino, E., Polder, G., Boer, M. P., Rutledge, D. N., & Herrmann, I. (2021). A generic workflow combining deep learning and chemometrics for processing close-range spectral images to detect drought stress in Arabidopsis thaliana to support digital phenotyping. *Chemometrics and Intelligent Laboratory Systems*, *216*, 104373. https://doi.org/10.1016/j.chemolab.2021.104373

Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R., & Melgani, F. (2018). *Computer vision-based phenotyping for improvement of plant productivity: A machine learning perspective*. *8*, 1-12. https://doi.org/10.1093/gigascience/giy153

Moeinizade, S., Hu, G., & Wang, L. (2022). A reinforcement Learning approach to resource allocation in genomic selection. *Intelligent Systems with Applications*, *14*, 200076. https://doi.org/10.1016/j.iswa.2022.200076

Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., & Hellier, P. (2022). A Survey on Deep Learning for Skeleton-Based Human Animation. *Computer Graphics Forum*, *41*(1), 122-157. https://doi.org/10.1111/cgf.14426

Murchie, E. H., & Lawson, T. (2013). Chlorophyll fluorescence analysis: A guide to good practice and understanding some new applications. *Journal of Experimental Botany*, *64*(13), 3983-3998.

N. Shvetsova, B. Bakker, I. Fedulova, H. Schulz, & D. V. Dylov. (2021). Anomaly Detection in Medical Imaging With Deep Perceptual Autoencoders. *IEEE Access*, *9*, 118571-118583. https://doi.org/10.1109/ACCESS.2021.3107163

N. Siddique, S. Paheding, C. P. Elkin, & V. Devabhaktuni. (2021). U-Net

and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, *9*, 82031-82057. https://doi.org/10.1109/ACCESS.2021.3086020

Naeem, M., Demirel, U., Yousaf, M. F., Caliskan, S., & Caliskan, M. E. (2021). Overview on domestication, breeding, genetic gain and improvement of tuber quality traits of potato using fast forwarding technique (GWAS): A review. *Plant Breeding*, *140*(4), 519-542. https://doi.org/10.1111/pbr.12927

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Baccus, S. A., & Ré, C. (2023). *HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution*.

Nicholls, H. L., John, C. R., Watson, D. S., Munroe, P. B., Barnes, M. R., & Cabrera, C. P. (2020). Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. *Frontiers in Genetics*, *11*. https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.00350

Nyrén, P., Pettersson, B., & Uhlén, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem*, *208*(1), 171-175. https://doi.org/10.1006/abio.1993.1024

Otto, M., Thornton, J., Rebert, C., Lauke, P. H., & Galante, A. (2024). *Bootstrap* [Software]. https://getbootstrap.com/

Pallets. (2024). *Flask* [Software]. https://palletsprojects.com/

Papa, R., Gioia, T., Torcida, E. M., Tuberosa, R., Crossa, J., Montesinos-López, O. A., Montesinos-López, A., Maccaferri, M., Sciara, G., Ammar, K., & Crossa, J. (2019). Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. Front. *Plant science : an international journal of experimental plant biology*, *10*, 1311. https://doi.org/10.3389/fpls.2019.01311

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32*, 8024-8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825-2830.

Peng, J. H., Sun, D., & Nevo, E. (2011). Domestication evolution, genetics and genomics in wheat. *Molecular Breeding*, *28*(3), 281-301. https://doi.org/10.1007/s11032-011-9608-4

Pérez-Enciso, M., & Zingaretti, L. M. (2019). A Guide on Deep Learning for Complex Trait Genomic Prediction. *Genes*, *10*(7). https://doi.org/10.3390/genes10070553

Pinaya, W. H. L., Vieira, S., Garcia-Dias, R., & Mechelli, A. (2020). Chapter 11—Autoencoders. En A. Mechelli & S. Vieira (Eds.), *Machine Learning* (pp. 193-208). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00011-0

Piperno, D. R. (2018). A model of agricultural origins. *Nature Human Behaviour*, *2*(7), 446-447. https://doi.org/10.1038/s41562-018-0390-8

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., & Ré, C. (2023). *Hyena Hierarchy: Towards Larger Convolutional Language Models*.

Prinzenberg, A. E., Campos-Dominguez, L., Kruijer, W., Harbinson, J., & Aarts, M. G. M. (2020). Natural variation of photosynthetic efficiency in Arabidopsis thaliana accessions under low temperature conditions. *Plant, cell & enviroment*, *43*(8), 2000-2013. https://doi.org/10.1111/pce.13811

R. Wei & A. Mahmood. (2021). Recent Advances in Variational Autoencoders With Representation Learning for Biomedical Informatics: A Survey. *IEEE Access*, *9*, 4939-4956. https://doi.org/10.1109/ACCESS.2020.3048309

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022a). *Hierarchical Text-Conditional Image Generation with CLIP Latents*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022b).

Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv*. https://doi.org/10.48550/arXiv.2204.06125

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Reyad, M., Sarhan, A. M., & Arafa, M. (2023). A modified Adam algorithm for deep neural network optimization. *Neural Computing and Applications*, *35*(23), 17095-17112. https://doi.org/10.1007/s00521-023-08568-z

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761-1770. https://doi.org/10.1038/s41593-019-0520-2

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, *22*(3), 400-407. JSTOR.

Rodríguez, G. R., Moyseenko, J. B., Robbins, M. D., Morejón, N. H., Francis, D. M., & van der Knaap, E. (2010). Tomato Analyzer: A useful software application to collect accurate and detailed morphological and colorimetric data from two-dimensional objects. *Journal of Visualized Experiments*, *37*.

Roitsch, T., Cabrera-Bosquet, L., Fournier, A., Ghamkhar, K., Jiménez-Berni, J., Pinto, F., & Ober, E. S. (2019). Review: New sensors and data-driven approaches—A path to next generation phenomics. *The 4th International Plant Phenotyping Symposium*, *282*, 2-10. https://doi.org/10.1016/j.plantsci.2019.01.011

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv*. https://doi.org/10.48550/arXiv.2112.10752

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, *242*(1), 84-89. https://doi.org/10.1006/abio.1996.0432

Rooijen, R. V., Kruijer, W., Boesten, R., Eeuwijk, F. A. V., Harbinson, J., & Aarts, M. G. M. (2017). Natural variation of YELLOW SEEDLING1 affects photosynthetic acclimation of Arabidopsis thaliana. *Nature communications*, *8*(1421).

https://doi.org/10.1038/s41467-017-01576-3
Rosenblatt, F. (1957). *The Perceptron, a Perceiving and Recognizing Automaton Project Para*. Cornell Aeronautical Laboratory.
Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386-408. https://doi.org/10.1037/h0042519
Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W. (2017). ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*, *18*(1), 529. https://doi.org/10.1186/s12859-017-1934-z
Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, *265*(5596), 687-695. https://doi.org/10.1038/265687a0
Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, *94*(3), 441-448. https://doi.org/10.1016/0022-2836(75)90213-2
Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, *74*(12), 5463-5467. https://doi.org/10.1073/pnas.74.12.5463
Santos, C. F. G. D., & Papa, J. P. (2022). Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks. *ACM Comput. Surv.*, *54*(10s). https://doi.org/10.1145/3510413
Sarić, R., Nguyen, V. D., Burge, T., Berkowitz, O., Trtílek, M., Whelan, J., Lewsey, M. G., & Čustović, E. (2022). Applications of Hyperspectral Imaging in plant phenotyping. *Trends in Plant Science*, *27*(3), 301-315. https://doi.org/10.1016/j.tplants.2021.12.003
Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, *12*(7). https://doi.org/10.3390/biology12070997
Sax, K. (1923). The Association of Size Differences with Seed-Coat Pattern and Pigmentation in PHASEOLUS VULGARIS. *Genetics*, *6*, 552-560. https://doi.org/10.1093/genetics/8.6.552
Scanes, C. G. (2018). Chapter 6—The Neolithic Revolution, Animal Domestication, and Early Forms of Animal Agriculture. En C. G.

Scanes & S. R. Toukhsati (Eds.), *Animals and Human Society* (pp. 103-131). Academic Press. https://doi.org/10.1016/B978-0-12-805247-1.00006-X

Scharr, H., Minervini, M., French, A. P., Klukas, C., Kramer, D. M., Liu, X., Luengo, I., Pape, J.-M., Polder, G., Vukadinovic, D., Yin, X., & Tsaftaris, S. A. (2016). Leaf segmentation in plant phenotyping: A collation study. *Machine Vision and Applications*, *27*(4), 585-606. https://doi.org/10.1007/s00138-015-0737-3

Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z., Chen, J., Wang, R., Zhao, H., Chong, Y., Shen, J., Zha, Y., & Yang, Y. (2021). Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(6), 2775-2780. https://doi.org/10.1109/TCBB.2021.3065361

Suárez, J. C., Vanegas, J. I., Contreras, A. T., Anzola, J. A., Urban, M. O., Beebe, S. E., & Rao, I. M. (2022). Chlorophyll fluorescence imaging as a tool for evaluating disease resistance of common bean lines in the western amazon region of colombia. *Plants*, *11*(10), Article 1371. https://doi.org/10.3390/plants11101371

Sun, S., Dong, B., & Zou, Q. (2021). Revisiting genome-wide association studies from statistical modelling to machine learning. *Briefings in Bioinformatics*, *22*(4), bbaa263. https://doi.org/10.1093/bib/bbaa263

Swartz, L. G., Liu, S., Dahlquist, D., Kramer, S., Walter, E. S., McInturf, S., Bucksch, A., & Mendoza-Cozatl, D. G. (2021). OPEN leaf: An open-source cloud-based phenotyping system for tracking dynamic changes at leaf-specific resolution in Arabidopsis. *bioRxiv : the preprint server for biology*, 2021.12.17.472861. https://doi.org/10.1101/2021.12.17.472861

Taghizadeh, M., & Chalechale, A. (2022). A comprehensive and systematic review on classical and deep learning based region proposal algorithms. *Expert Systems with Applications*, *189*, 116105. https://doi.org/10.1016/j.eswa.2021.116105

Temel, D., & Alregib, G. (2014). Image quality assessment and color difference. *2014 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2014*, 970-974. https://doi.org/10.1109/GLOBALSIP.2014.7032265

Thomas Stalke, H., L. Warburton, M., & R. Harlan, J. (2021). *Harlan's Crops and Man: People, Plants and Their Domestication*.

Tian, Y., Zhang, Y., & Zhang, H. (2023). Recent Advances in Stochastic

Gradient Descent in Deep Learning. *Mathematics*, *11*(3). https://doi.org/10.3390/math11030682

Tisne, S., Reymond, M., Vile, D., Fabre, J., Dauzat, M., Koornneef, M., & Granier, C. (2008). Combined genetic and modeling approaches reveal that epidermal cell area and number in leaves are controlled by leaf and plant developmental processes in arabidopsis. *Plant Physiology*, *148*(2), 1117-1127. https://doi.org/10.1104/pp.108.124271

Torkamaneh, D., Boyle, B., & Belzile, F. (2018). Efficient genome-wide genotyping strategies and data integration in crop plants. *Theoretical and Applied Genetics*, *131*(3), 499-511. https://doi.org/10.1007/s00122-018-3056-z

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2302.13971

Tsai, H.-Y., Janss, L. L., Andersen, J. R., Orabi, J., Jensen, J. D., Jahoor, A., & Jensen, J. (2020). Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. *Scientific Reports*, *10*(1), 3347. https://doi.org/10.1038/s41598-020-60203-2

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1), 59. https://doi.org/10.1038/s43586-021-00056-9

van den Oord, A., Vinyals, O., & kavukcuoglu, koray. (2017). Neural Discrete Representation Learning. En I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf

van der Veen, M. (2010). Agricultural innovation: Invention and adoption or change and adaptation? *World Archaeology*, *42*(1), 1-12. https://doi.org/10.1080/00438240903429649

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709.

https://doi.org/10.1016/j.compag.2020.105709
van Ooijen, J. W. (2009). MapQTL® 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. *Kyazma BV, Wageningen, Netherlands*, *64*.
van Ooijen, J. W. (2011). Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. *Genetics Research*, *93*(5), 343-349. https://doi.org/10.1017/S0016672311000279
Varshney, R. K., Sinha, P., Singh, V. K., Kumar, A., Zhang, Q., & Bennetzen, J. L. (2020). 5Gs for crop genetic improvement. *Current Opinion in Plant Biology*, *56*, 190-196. https://doi.org/10.1016/j.pbi.2019.12.004
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. En I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., … Viola, R. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). *Nature Genetics*, *42*(10), 833-839. https://doi.org/10.1038/ng.654
Vinod, K. K. (2011). Kosambi and the genetic mapping function. *Resonance*, *16*(6), 540-550. https://doi.org/10.1007/s12045-011-0060-x
Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, *2018*, 7068349. https://doi.org/10.1155/2018/7068349
Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, *9*(2), 187-212. https://doi.org/10.1007/s40745-020-00253-5
Wang, X., Xuan, H., Evers, B., Shrestha, S., Pless, R., & Poland, J. (2019). High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. *GigaScience*, *8*(11). https://doi.org/10.1093/gigascience/giz120

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Lot it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, *26*(1), 98-117. https://doi.org/10.1109/MSP.2008.930649

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737-738. https://doi.org/10.1038/171737a0

Wen, S., Wang, G., Yang, Z., Wang, Y., Rao, M., Lu, Q., & Hong, N. (2020). Next-Generation Sequencing Combined With Conventional Sanger Sequencing Reveals High Molecular Diversity in Actinidia Virus 1 Populations From Kiwifruit Grown in China. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.602039

Wetterstrand, K. A. (2021). The cost of sequencing a human genome. *Genome.gov*. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Wijfjes, R. Y. (2021). *Computational analysis of copy number variation in plant genomes* [Phd]. Wageningen University.

Wilkins, M. H. F., Stokes, A. R., & Wilson, H. R. (1953). Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature*, *171*(4356), 738-740. https://doi.org/10.1038/171738a0

Wingler, A., Marès, M., & Pourtau, N. (2004). Spatial patterns and metabolic regulation of photosynthetic parameters during leaf senescence. *New phytologyst*, *161*(3), 781-789. https://doi.org/10.1111/j.1469-8137.2003.00996.x

Xu, R., & Li, C. (2022). A Review of High-Throughput Field Phenotyping Systems: Focusing on Ground Robots. *Plant Phenomics*, *2022*, 9760269. https://doi.org/10.34133/2022/9760269

Yin, X., Liu, X., Chen, J., & Kramer, D. M. (2014). Multi-leaf tracking from fluorescence plant videos. *IEEE international conference on image processing (ICIP)*, 408-412. https://doi.org/10.1109/ICIP.2014.7025081

You, E., Deng, K., Linusborg, Kazupon, Singh, D., Rodrigues, C., Tepluhina, N., Hong, B., & and more. (2024). *Vue.js* [Software]. https://vuejs.org/

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple

levels of relatedness. *Nature Genetics*, *38*(2), 203-208. https://doi.org/10.1038/ng1702

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00189

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 586-595. https://doi.org/10.1109/CVPR.2018.00068

Zhao, Y., Mette, M. F., Gowda, M., Longin, C. F. H., & Reif, J. C. (2014). Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity*, *112*(6), 638-645. https://doi.org/10.1038/hdy.2014.1

Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Davuluri, R. V., Wang, Z., & Liu, H. (2024). *DNABERT-S: Learning Species-Aware DNA Embedding with Genome Foundation Models*.

Annex

In this annex it is included the references to papers such as Leaf Tracking and Genodrawing, and other papers in which I participated as co-author but were not part of the main line of this thesis.

Jurado-Ruiz, F., Nguyen, TP., Peller, J. *et al.* LeTra: a leaf tracking workflow based on convolutional neural networks and intersection over union. *Plant Methods* **20**, 11 (2024). https://doi.org/10.1186/s13007-024-01138-x

Jurado-Ruiz, F., Rousseau, D., Botía, Aranzana, M. J.. GenoDrawing: An Autoencoder Framework for Image Prediction from SNP Markers. *Plant Phenomics.* 2023;5:0113. https://doi.org/10.34133/plantphenomics.0113

Pradas, N., Jurado-Ruiz, F., Onielfa, C., Arús, P., Aranzana, M.J., PERSEUS: an interactive and intuitive web-based tool for pedigree visualization, *Bioinformatics*, Volume 40, Issue 2, February 2024, btae060, https://doi.org/10.1093/bioinformatics/btae060

Fiol, A., Jurado-Ruiz, F., García, S., Dujak, C., López-Girona, E., Pacheco, I., … Aranzana, M. J. (2023, March). An LTR retrotransposon may activate the PsMYB10.2 gene expression and promote anthocyanin biosynthesis in the fruit flesh of Japanese plum – a success story of CRISPR-Cas9 enrichment

sequencing strategy. Acta Horticulturae. International Society for Horticultural Science (ISHS). http://doi.org/10.17660/actahortic.2023.1362.3

Fiol, A., Jurado-Ruiz, F., López-Girona, E. et al. An efficient CRISPR-Cas9 enrichment sequencing strategy for characterizing complex and highly duplicated genomic regions. A case study in the Prunus salicina LG3-MYB10 genes cluster. Plant Methods 18, 105 (2022). https://doi.org/10.1186/s13007-022-00937-4