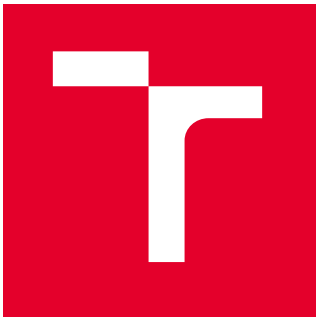
BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

SEMESTRAL THESIS

Brno, 2024 Vojtěch Pavlík



BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

GENOMIC PREDICTION IN PLANTS USING CONVOLUTIONAL NEURAL NETWORKS

GENOMICKÁ PREDIKCE U ROSTLIN POMOCÍ KONVOLUČNÍCH NEURONOVÝCH SÍTÍ

SEMESTRAL THESIS
SEMESTRÁLNÍ PRÁCE

AUTHOR — **Vojtěch Pavlík**
AUTOR PRÁCE

SUPERVISOR — **Ing. et Ing. Jana Schwarzerová, MSc**
VEDOUCÍ PRÁCE

BRNO 2024

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Semestral Thesis

Bachelor's study program **Biomedical Technology and Bioinformatics**

Department of Biomedical Engineering

Student: Vojtěch Pavlík ***ID:*** 247288

Year of study: 3 ***Academic year:*** 2024/25

TITLE OF THESIS:

Genomic prediction in plants using convolutional neural networks

INSTRUCTION:

1) Perform a literature review on techniques used for genomic prediction. 2) Study prediction methods based on linear and nonlinear approaches. 3) Pre-process available genomic data so that it can be appropriately applied to genomic prediction analysis. Test its suitability on at least two prediction methods. 4) Expand the work to include a prediction approach based on deep learning using convolutional neural networks. 5) Focus on specific parameters and optimize the prediction method accordingly. 6) Discuss the results.
To complete the semester project, it is necessary to fulfill points 1) to 3).

RECOMMENDED LITERATURE:

[1] Weiszmann, Jakob, et al. Metabolome plasticity in 241 Arabidopsis thaliana accessions reveals evolutionary cold adaptation processes. Plant Physiology 193.2 (2023): 980-1000.

[2] Fu, Yabo, et al. Deep learning in medical image registration: a review. Physics in Medicine & Biology 65.20 (2020): 20TR01.

[3] Sidak, D., Schwarzerová, J., Weckwerth, W., & Waldherr, S. Interpretable machine learning methods for predictions in systems biology from omics data. Frontiers in Molecular Biosciences, 9, 926623 (2022).

Date of project specification: 16.9.2024 ***Deadline for submission:*** 3.1.2025

Supervisor: Ing. et Ing. Jana Schwarzerová, MSc
Consultant: Univ.-Prof. Dr. Wolfram Weckwerth

doc. Ing. Jana Kolářová, Ph.D.
Chair of study program board

ABSTRACT

This semestral thesis focuses on genomic prediction in plants, using dynamic prediction based on genomic data consisting of single nucleotide polymorphism datasets. It contains four parts. The first part focuses on basic genomic prediction terminology, the second one is a comprehensive overview of the prediction methods used for genomic prediction, the third part describes the data sets used in this thesis and the fourth part follows up with the implementation of the data sets pre-processing for selected machine learning models.

KEYWORDS

Genomic prediction, polymorphism, machine learning, ridge regression, random forest, deep learning, metabolomics

ABSTRAKT

Tato semestrální práce se zaměřuje na genomickou predikci u rostlin pomocí dynamické predikce založené na genomových datech sestávajících z datových sad jednonukleotidových polymorfismů. Obsahuje čtyři části. První část se zabývá základní terminologií genomické predikce, druhá je přehledem predikčních metod používaných pro predikci genomu, třetí část popisuje datové sady použité v této práci a čtvrtá část navazuje na implementaci předzpracování datových sad pro vybrané modely strojového učení.

KLÍČOVÁ SLOVA

Genomická predikce, polymorfismus, strojové učení, hřebenová regrese, náhodný les, hluboké učení, metabolomika

PAVLÍK, Vojtěch. *Genomic prediction in plants using convolutional neural networks*. Semestral Project. Brno: Brno University of Technology, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2025. Advised by Ing. et Ing. Jana Schwarzerová, MSc.

Author's Declaration

Author:	Vojtěch Pavlík
Author's ID:	247288
Paper type:	Semestral Project
Academic year:	2024/25
Topic:	Genomic prediction in plants using convolutional neural networks

I declare that I have written this paper independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the paper and listed in the comprehensive bibliography at the end of the paper.

As the author, I furthermore declare that, with respect to the creation of this paper, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll. of the Czech Republic, Section 2, Head VI, Part 4.

Brno ....

author's signature*

*The author signs only in the printed version.

ACKNOWLEDGEMENT

I would like to thank my supervisor Ing. et Ing. Jana Schwarzerová, MSc for professional supervision, support, patience and explanation of the genomic prediction problematics.

Contents

List of Figures

List of Tables

Introduction

Genomic prediction, the process of predicting phenotypic traits from genotypic data, has revolutionized fields such as agriculture, plant breeding, and personalized medicine. This powerful approach enables researchers to estimate complex traits, such as crop yield or disease resistance, using genetic markers. With advancements in high-throughput genotyping technologies, vast amounts of genomic data are now available, necessitating the use of sophisticated computational tools for effective analysis.

This thesis focuses on machine learning and deep learning methods, modeling complex nonlinear relationships, and uncovering patterns in high-dimensional data.

The goal of the thesis is to compare the performance of linear machine learning methods with deep learning CNN method and compare their suitability on two plant data sets.

1 Genomics prediction

Genomic Prediction (GP) is used to forecast phenotypes and breeding values of an individual based on their genomic data. GP allows researchers to predict traits and disease predispositions by understanding complex interactions between genetic markers in DNA and phenotypic traits [7].

The genotype-phenotype relationship is often distorted by a multitude of factors, including environmental influences and epigenetic modifications. The environmental effects can be suppressed by incorporating not only genome-wide, but also environment-wide studies [8]. Models that account for these factors have proven to be more precise. However, the dynamic interaction between genetic predispositions and environmental exposures is quite difficult to predict [9].

1.1 Genome-wide association studies

Genome-Wide Association Studies (GWAS) are seeking fully understanding genotype-phenotype correlations. GWAS are based on comparing individuals who differ phenotypically but share common ancestors in terms of allele frequency of genetic variants [10]. GWAS explore relationships between common genome sequence variations and disease predisposition on a genome-wide scale.

GWAS involve scanning the genomes of a large number of individuals to identify genetic variants associated with specific diseases or traits. By comparing DNA samples of two groups that have different traits of interest, it can be identified which specific genetic variants are probably related to the observed trait. The advantage of GWAS lies in their ability to detect associations with little effect.

Throughout time GWAS have contributed to a substantial variety of associations for particular traits, diseases, or condition in the individual's genome. Identifying genetic variants associated with diseases such as coronary artery disease [11], inflammatory bowel disease [12], type 2 diabetes mellitus [13] and major depression [14], GWAS provide insights into the biological pathways involved in these conditions [15]. This knowledge can lead to the development of new therapeutic strategies, agricultural breeding, and personalized medicine approaches.

Sample sizes in GWAS have grown from several thousand individuals to hundreds of thousands, enhancing the statistical power to detect genetic variants with smaller effect sizes [15]. The number of diseases and traits studied has also expanded. This growth has led to the identification of numerous genetic variants associated with complex human diseases and traits.

To predict phenotypes from genomic data, the main focus of GP is to find within

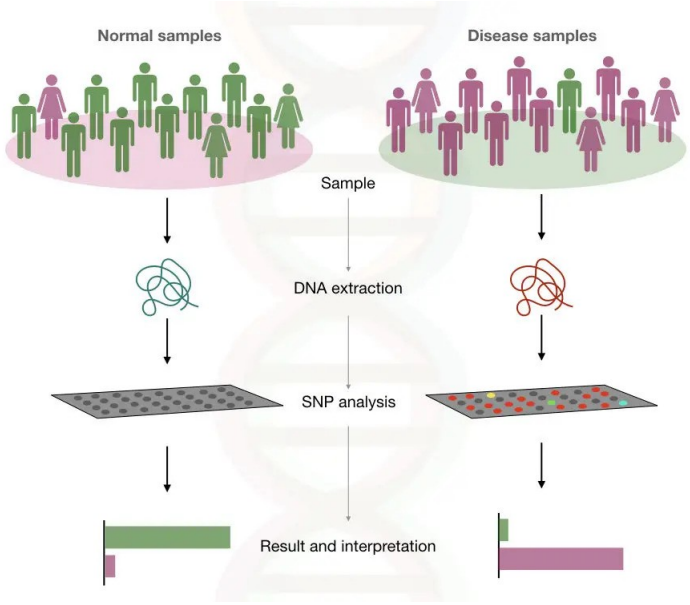


Fig. 1.1: Genome-wide association [1]

polymorphisms. The most common polymorphisms are Single Nucleotide Polymorphisms (SNP) [16].

1.2 Single nucleotide polymorphisms

SNPs are among the most common type of sequence differences between alleles. SNP is a one-base position where the genomes of some individuals differ from the genome sequences of others. Human Genome Project, showed that these single-nucleotide changes in our genetic code are placed all across individuals' genomes [16]. These variations are significant because of their influence on disease developments, response to pathogens, and traits display. SNPs are not necessarily pathological mutations. They can be benign and contribute to genetic diversity within a population.

SNPs are also used in the detection of associations between allele forms of a gene and phenotypes for multifactorial genetics considering Linkage disequilibrium (LD) blocks. LD is the non-random association of alleles at neighboring loci. The closer SNPs are to each other, the more it is probable to be a region inheritable as a block.

Although the human genome is vast and complex, various plant genomes and their phenotype oscillations are practically controlled by SNP. This gives great advantage into SNP based GP research [17].

Overall, there is a clear relation between genotype and phenotype, and even if not yet fully decoded, many complex traits have been linked to changes in specific genome locations.

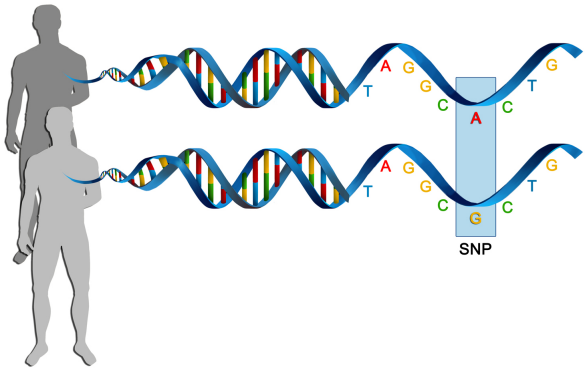


Fig. 1.2: Case of a single nucleotide polymorphism, where two individuals' genomes differ at single base position [2]

1.3 Genomic prediction in practise

The success rate of GP models depends heavily on the statistical methods employed. While linear statistical models like best linear unbiased prediction (BLUP; [18]) have traditionally been used for GP, their limitations by ignoring non-linear statistical effects are becoming evident [19]. Consequently, Machine Learning (ML) methods are gaining dominance due to their ability to capture non-linear relations between prediction and response. On the other hand, ML methods are significantly more demanding in terms of computational resources. However, advancements in technologies and ML algorithms are enabling use of ML methods [20].

Recent study by Wang et al. [21] discusses perfecting GP methods using Deep neural network genomic prediction (DNNGP) that takes into account multi-omics data from plants. This model is particularly promising in multiple ways. Not only it has potential of being more accurate thanks to multi-omics data. It is still competitive with linear models even when small data sets are given, outperforming other DL methods that need large data sets. Even the computational time was somewhat comparable to other methods, being multiple times shorter than DL genomic selection.

2 Prediction methods

ML implementations are broadly classified into several categories based on the nature of the learning signal or feedback available to the system. The primary types include *supervised learning*, *unsupervised learning* and *reinforced learning* [22].

This thesis is focused on supervised algorithms due to their prevalence in general and in the context of predictive systems biology [23]. Supervised algorithms have predetermined classes. These classes are created in a manner of finite set, defined by the human, which in practice means that a certain segment of data will be labeled with these classifications. The task of the machine learning algorithm is to find patterns and construct mathematical models [24]. The supervised algorithms include both conventional methods (regular regression methods and Random Forest (RF)) and DL models.

2.1 Conventional supervised learning based methods

Supervised learning has two main models - *classification models* and *regression models*. Regression models map the input space into a real-value domain. The classifiers map the input space into predefined classes. We will examine the regression model, specifically Regularized regression methods and RF.

2.1.1 Regularized regression methods

Linear regression models, while straightforward and interpretable, often suffer from issues like overfitting, especially when dealing with high-dimensional data sets. Continuous penalization methods have emerged as a robust solution to mitigate these issues by introducing penalties on the model coefficients. Continuous penalization methods contain Ridge Regression (RR) and Least Absolute Shrinkage and Selection Operator (LASSO) [25].

RR (L2 regularization) and LASSO (L1 regularization) are types regularization for linear regression models [26]. The purpose of these regularizations is to reduce overfitting in machine learning models. Both L1 and L2 regularization add a penalty term to the loss function in a machine learning model. This reduces overall model complexity by giving weight to specific parameters that deter the model from fitting training data too closely. However, both regularization methods have different penalty terms and specific use cases [26].

Both penalty terms for L1 and L2 regularization are controlled by the model hyperparameter λ, which determines the tradeoffs between bias and variance in coefficients. RR can reduce coefficient values toward zero but never exactly to zero.

This means that L2 regularization cannot perform feature selection [26]. The cost function for Ridge regression is defined as:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \tag{2.1}$$

Where; y=response vector; β=regression coefficients; x=predictor variables; n=number of the biological samples; p=number of variables [25].

LASSO introduces a penalty based on the absolute values of the coefficients. This method not only helps reduce overfitting but also performs feature selection by driving some coefficients to zero, effectively removing them from the model [26]. The LASSO cost function is given by:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \tag{2.2}$$

2.1.2 Random forest

Random Forest (RF) is an ensemble of tree predictors. It is highly data adaptive, applies to large number of variables and small number of biological samples. RF can take into account both correlation and feature interactions [27]. RF also allows measuring the relative importance of each predictor (independent variable) for prediction [28]. This makes RF significantly appealing for high-dimensional genomic data analysis [27].

The basic unit of RF is Decision Tree (DT) . DT is a prediction model with given set of data. DT manufactures logic construction diagrams similar to rule-based prediction, which express and classify a series of conditions that occur in succession to solve a problem. Each DT has several nodes, branches and leaf nodes. Every node tests a particular independent variable, often value compared to a constant. Each branch represents the outcome of the test. Each leaf node represents a class label in a classification problem. Each individual classification is therefore based on going down the tree according to the values of the independent variables tested in the successive nodes, until a leaf is reached, classifying the individual accordingly to the class assigned to the leaf (figure 2.1) [28].

In RF, each tree votes for the most popular class (majority voting / averaging). The variability of every DT in RF is ensured by generating random vectors that lead the growth of each DT. One of the main and most popular techniques is bootstrap aggregating (bagging). Bagging means that for each DT a random selection is made from the examples of the training data set [28] [29]. This means that in the bootstrap

sample, some training samples will be included more often than others and about one third will not be included at all. It is usually about one third of the training samples, later labeled as out-of-bag samples [25].

The performance of RF is influenced by several hyperparameters that require tuning. Large number of DTs generally leads to better performance, but with diminishing returns and increased computational costs. Too deep DTs can capture more information but may lead to overfitting [30].

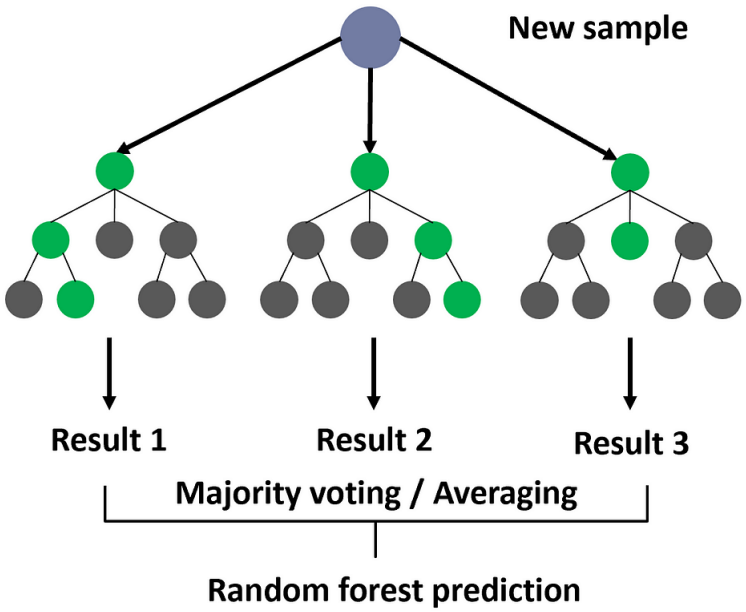


Fig. 2.1: Random forest architecture, processing new sample into prediction [3]

2.2 Deep learning methods

DL is a class of ML algorithms based on deep multi-layered neural network architecture with outstanding flexibility. The ability of DL models to process large data sets and capture complex patterns makes them particularly suited for GP [31].

Although DL has no proven performance superiority, there are clear evidences that DL algorithms capture non-linear patterns better than classical ML methods [32]. Unlike classical ML methods, DL models can operate on the sequence directly, therefore not requiring pre-defying features extracted from the sequence. This means that DL can automatically learn features and patterns with less expert handcrafting [5].

A generic DL architecture consists of multiple hidden neuron layers. Each layer takes the output of previous layer as its input. Single neuron is represented by activation function (f) where input is a list of variable values (x) multiplied by weights (w). Output is a non-linear transformation $f(wx + b)$, where(b) is the bias. Weights and bias are adaptive for each neuron and their values defy loss function [33].

All DL methods share common principle of using stacked neuron layers. The architectures however, differ. The most popular DL methods are Feed-forward neural network (FFNN) , Convolutional neural network (CNN) and Recurrent neural network (RNN) .

2.2.1 Feed forward neural network

FFNN, also known as MLP (Multi-layer perceptron) (figure 2.2, b) consists of fully connected layers (input, multiple hidden layers, output layer). Each layer's output is a weighted non-linear function of each previous layer's output, plus the bias [20, 33].

2.2.2 Recurrent neural network

RNNs (figure 2.2, a) are designed to model space-temporal structures. In this network all the neurons in one hidden layer have not only incoming connections from previous layer, but also ongoing connections to subsequent layer and recurrent connections to propagate information between neurons of the same layer [32, 33]. In RNN has at least one feedback loop since the signals travel in both directions. The model has short memory or delay, since it has a bias based on the past [32].

2.2.3 Convolutional neural network

CNNs are particularly suited for analyzing spatial and temporal dependencies of the input, good for one-dimension (GP), and two-or three dimensions (images) [33]. In each convolutional layer, a convolutional operation (kernel) is performed along input width and strides. After each convolution, an activation function produces the output, making matrix called feature map [33]. The pooling layer (figure 2.3) is then applied to reduce the spatial size of the representation to reduce computation in the network and to smooth out the result [32]. It consists of merging the kernel outputs of different feature map positions by taking the mean, maximum, or minimum of all values of those positions [33]. In the end, after the convolutional layers, the input is flattened.

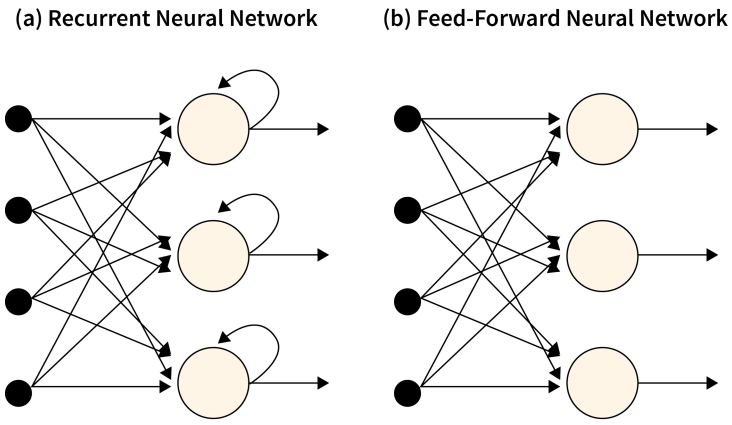


Fig. 2.2: Difference between recurrent neural network and feed-forward neural network [4]

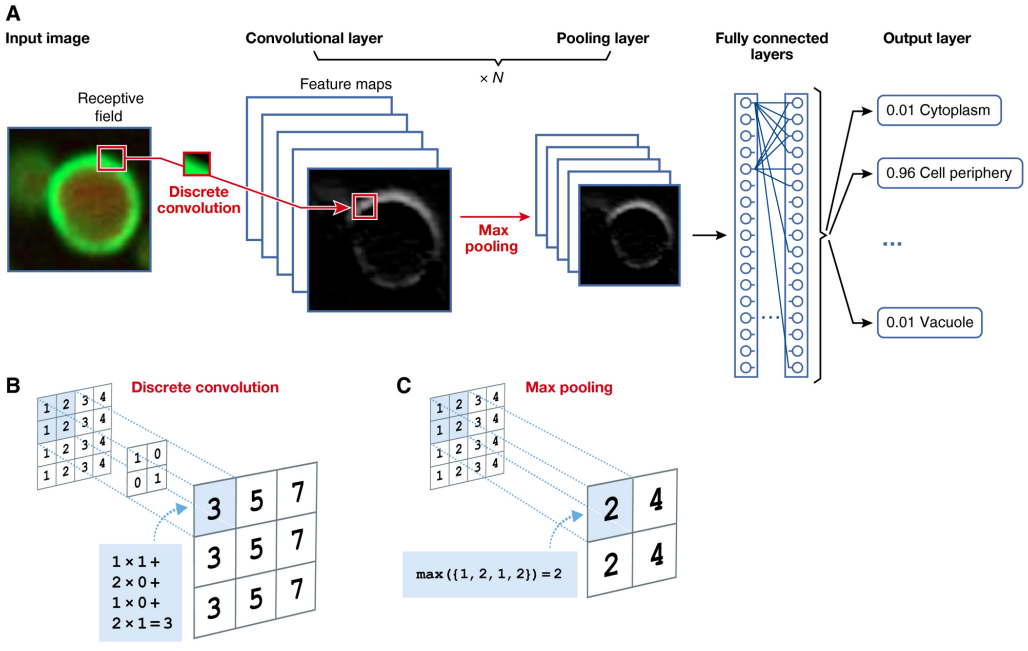


Fig. 2.3: Convolutional neural network architecture [5]

3 Materials

This chapter is organized into two main sections. The first section introduces the data sets used in this thesis. Since the focus of this work is on GP in plants, the data sets are derived from two representative species - *Arabidopsis thaliana* and *Hordeum vulgare*.

These data sets contain genomic and metabolomic information. Genomic data consist of SNP markers that capture genetic variation throughout the genome. Metabolomic data include measurements of key metabolites that play crucial roles in plant traits, such as plant physiology and development. Together, the data sets enable comprehensive analysis of the relationships between genome and metabolome. The performance of GP is evaluated and compared between both data sets.

The second part addresses data preprocessing steps for GP implementation.

3.1 *Arabidopsis thaliana*

Arabidopsis is a small plant in the *Brassicaceae* family, being a model of choice for research in plant biology. Working with *Arabidopsis* offers several benefits, including low maintenance requirements, high reproductivity, and a relatively short genome [34]. Majority of the genes in the *Arabidopsis* genome have been identified thanks to its excellent annotation through several projects such as The Arabidopsis Information Resource (TAIR) , mainly TAIR10, which facilitates GP research [35].

In this study, the *Arabidopsis* data set of genomic and metabolomic information is based on the study by Weiszmann et al. [36]. This study offers 241 natural accessions of *Arabidopsis thaliana*. These accessions were grown under two different temperature regimes (6 ° C and 16 ° C), allowing the analysis of genome-environment interactions and their effects on metabolome variation.

3.1.1 Genomic information

The genomic data set from [36] consists of high density SNP data, capturing genetic variation across the genome. This data set provides comprehensive coverage of genetic variation, with 1,756,214 bialellic SNPs identified across the genome, giving a high resolution set of approximately one SNP per 77 bp used for GWAS computations [36]. For GP, the data set has been altered. No SNP profiles repeat across the 241 accessions, and the data set consists of vectors of length 241 with elements "1" and "-1". SNPs selected from the original data set are only those that required at most one missing allele to be imputed, making 16,544 biallelic SNPs with resolution of approximately one SNP per 8 kb [36].

3.1.2 Metabolomic information

The metabolomic data set features quantified levels of 37 primary metabolites measured across all 241 natural accessions for both temperature regimes. The metabolites were extracted and measured by gas chromatography combined with mass spectrometry [36]. These metabolites include sugars, organic acids, and amino acids, which play crucial roles in physiological responses to environmental changes (figure 3.1).

Metabolites concentration data are normalized to render their distribution Gaussian and to reduce batch and repeat effects [36]. Later, data were median-centered, and all metabolites were standardized, therefore mean = 0, variance = 1. Resulting, for each of the 37 metabolites a single value was obtained for all 241 accessions in two temperature regimes [36].

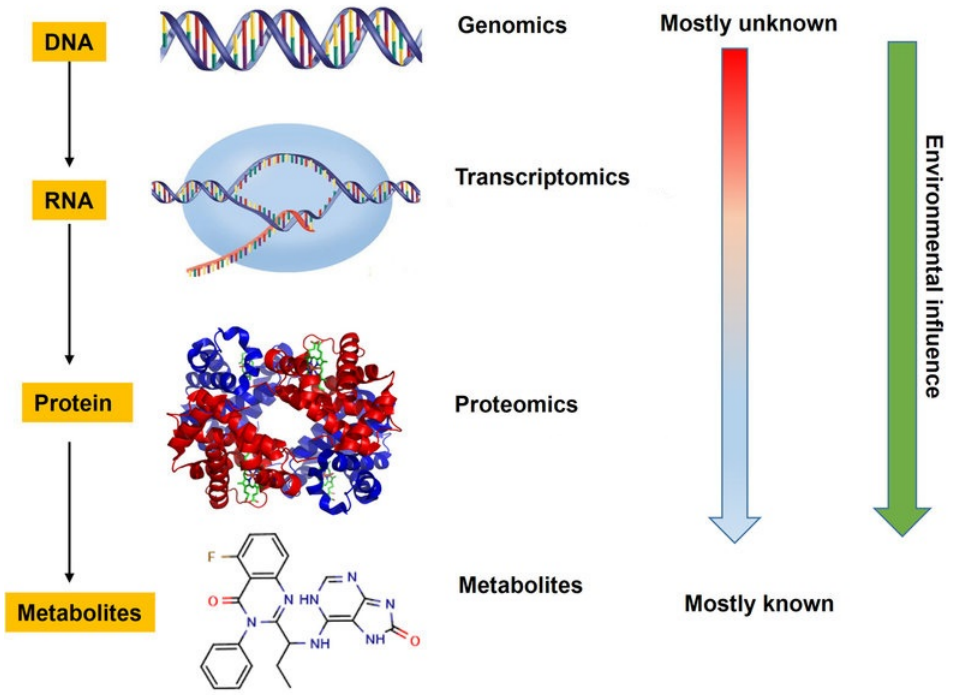


Fig. 3.1: Relationship between metabolomics and other omics with their relevance to environmental influence [6].

3.2 *Hordeum vulgare*

Barley (*Hordeum vulgare*) is the fourth most important cereal crop worldwide. Its importance is due not only to its agricultural value, but also to its potential as a

model organism for genomic research in the Triticeae tribe, along with wheat and rye [37]. The genomic exploration of barley provides critical insights into complex trait prediction, essential for improving crop yields and quality. Barley has a relatively large genome size, approximately 5 Gb, which is well characterized due to the release of high-quality reference genome by Mascher et al. [38].

In barley, traits such as grain yield [39], disease resistance [40], and drought stress tolerance [41] have been successfully targeted using GP.

In this study, the barley data set is based on the study by Gemmer et al. [42], using the first Nested association mapping (NAM) population, HEB-25, by crossing 25 wild barley accessions (24 *Hordeum vulgare* ssp. *spontaneum* and one *Hordeum vulgare* ssp. *agriocrithon*) with the German elite spring barley cultivar Barke (*Hordeum vulgare* ssp. *vulgare*) [42].

3.2.1 Genomic information

Like the genomic information of *arabidopsis*, the barley genomic information data set from [42], features high density SNP data.

DNA was extracted using two methods, the Biosprint 96 DNA Plant Kit and a Biosprint workstation (Qiagen, Hilden, Germany). The extracted DNA was dissolved in distilled water at a concentration of approximately 50 ng /µl [42]. Genotyping was performed with the recently developed barley Infinium iSelect 50K chip at TraitGenetics, Gatersleben, Germany. SNP markers that did not meet the quality criteria (polymorphic in at least one HEB family, $< 10\%$ failure rate, $< 12.5\%$ heterozygous calls) were removed from the data set [42].

Ultimately, the 33,005 SNPs that met the criteria were analyzed. Based on the Barke reference genotype, the state of each allele in the data set was classified as follows. Homozygous Barke alleles were coded as 0, homozygous wild barley alleles were coded as 2 and heterozygous alleles were coded as 1. If a SNP was monomorphic in one HEB family but polymorphic in another, the lines from the monomorphic family were coded as 0, consistent with their similarity to the Barke allele [42]. Gaps made from missing geeenotypes, approximately 0.84% were estimated by applying the mean imputation approach [42, 43].

3.2.2 Metabolomic information

The metabolomic data set from [42] contains fully developed leaf tissue samples collected from NAM population HEB-25. A total of 128 metabolites were profiled with sampling at two time points in one year, therefore two different stages of development. The first sampling was on 22 May 2017 between nine and ten O'clock. The second sampling was on 22 June 2017, under the same circumstances (clear

sky, equal time of day, same sampling methods) [42]. All samples along with 20% quality controls were analyzed with a combination of gas chromatography and mass spectrometry, determining all metabolites and their quantity. The raw data were processed using MassHunter Qualitative Analysis software. As a result, 158 metabolites concentrations of 1419 lines were obtained [42].

3.3 Data pre-processing

The data pre-processing stage is a crucial step in ensuring the quality and compatibility of data used for downstream analyses. This section describes specific pre-processing steps applied to each dataset. focuses on the pre-processing of genomic and metabolomic datasets from *arabidopsis*, while the second part addresses barley data sets.

All codes are written in Python 3.13 [44]. For pre-processing, Pandas [45] and Numpy [46] libraries are used. With Pandas, all the data sets are set to one format - Pandas [45] dataframes. Numpy [46] is used for data type oversight.

The barley data are altered by deleting not necessary rows and are reduced by removing non-matching HEB lines and uniting them in genomic and metabolomic data sets. Resulting 1307 accessions across different HEB lines in both data sets. The genomic data set is also altered by scaling -1 to 1 instead of 0 to 2, optimizing the input for predictions. The missing data marked as NaN were replaced with 0. These missing values may have been caused by low metabolite concentrations or insufficient signal strength at certain loci. The decimal values in the genomic data set were cassified as 1 (SNP detected). Lastly, the final pre-processed records are reduced into 10% of their size to shorten the training time cost. The reduction is performed by random selection of records.

The genomic and metabolomic data sets of *arabidopsis* are lined by accession numbers. The genomic accessions are unique, but for metabolomic data, there are multiple records for one accession. In order to secure uniform number of rows in both datasets, each row (accession) in genomic data set was multiplied by the number of records in one accession in metabolomic data. The NaN values in both *arabidopsis* data sets are set to 0. Since the metabolomic data are measured in two temperature conditions, data set is preprocessing splits it into 3 sub-data sets: For 6 °C only, for 16 °C only and both temperature regimes in one dataset. The comparison of these three data sets is mentioned in the following chapter. Other alternations were performed in the primer study [36].

4 Implementation

This chapter summarizes further practical operations for implementing the prediction models previously addressed in the thesis. That covers the steps involved in setting up the models, optimizing their hyperparameters, and training them for GP. The chapter is divided into two main sections: regularized linear regression and RF implementations, and the development of CNN. For the train/ test data split, the Scikit-learn [47] library is used.

4.1 Regularized linear regression and random forest

This section discusses the implementation of regularized linear regression methods, including RR and LASSO, and RF. All *arabidopsis* metabolomic data sets are normalized with MinMaxScaler in the Skicit-learn library [47]. All these models are set up, trained, and tested using the Scikit-learn library [47], too.

For *arabidopsis* data, the RR hyperparameters are set as follows: regularization strength = 0.001, maximum of iterations = 10,000. For barley, the regularization strength = 0.01 and maximum of iterations = 10,000. The LASSO regression hyperparameters for *arabidopsis* and barley are for each the same as in RR. In RF, the number of estimators is for both plants set to 100.

4.2 Convolutional neural network

This section describes the implementation of CNN for GP. It includes details on the architecture design, training procedures, and hyperparameter tuning. For the work with CNN, Jax library [48] is used. Jax library [48] is chosen for its compatibility with python 3.13 [44]. It works with numpy [46] arrays only, so the preprocessed pandas [45] dataframes are converted back into arrays.

Both *arabidopsis* and barley data are not reduced. For *arabidopsis* data, the number of epochs used in CNN is set to 10. The batch size is 32. For barley data, the number of epochs is 10 and the batch size is 22.

5 Predictive models in plants

This chapter examines the application of predictive models in plants GP, showing the results of custom implementations, concluding with a discussion of the results in the context of GP.

5.1 Own implementation for genomic prediction

This section highlights own implementations of GP models. It describes how the previously introduced and implemented methods are applied to plant data sets.

To evaluate the model, the used metrics are Mean square error (MSE) and Correlation coefficient (CC) . MSE is used to measure accuracy of the model by calculating the averages of the squared differences between the predicted and the actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.1}$$

Where n=number of data points, y_i = actual value of the i-th observation, y_i^2=predicted value of the i-th observation.

CC quantifies strength and direction of the relationship between two variables. It ranges from -1 to 1. 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship and 0 indicates no linear relationship between variables.

$$CC = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{5.2}$$

Where (x_i, y_i)=data points, ($\bar{x}$, $\bar{y}$)=means of the (x, y) variables.

To better understand the performance of each prediction model, there are following tables with columns: Model, Mean square error and Corelation coefficient.

Model	MSE [6 °C]	CC
RR	0.0243	0.279
LASSO	0.0146	0.350
RF	0.0159	0.310
CNN	0.0152	0.328

Table 5.1: Comparison of RR, LASSO, RF and CNN models performance through their computed evaluation metrics on *arabidopsis* data in 6 °C only.

Model	MSE [16 °C]	CC
RR	0.0111	0.310
LASSO	0.0087	0.369
RF	0.0095	0.341
CNN	0.0084	0.336

Table 5.2: Comparison of RR, LASSO, RF and CNN models performance through their computed evaluation metrics on *arabidopsis* data in 16 °C only.

Model	MSE [both]	CC
RR	0.0104	0.269
LASSO	0.0091	0.339
RF	0.0101	0.268
CNN	0.0091	0.312

Table 5.3: Comparison of RR, LASSO, RF and CNN models performance through their computed evaluation metrics on complete *arabidopsis* data in both temperature regimes.

Model	MSE	CC
RR	0.0074	0.700
LASSO	0.0077	0.735
RF	0.0093	0.637
CNN	0.0001	0.392

Table 5.4: Comparison of RR, LASSO, RF and CNN models performance through their computed evaluation metrics on barley data.

The tables allow us to compare the suitability of each model for different data sets. Overall, it is obvious that the LASSO prediction model is the most performing. For comparison, two figures of correlation function are shown. One is LASSO (figure 5.1) and the other is CNN (figure 5.2) with *arabidopsis* 16 °C data.

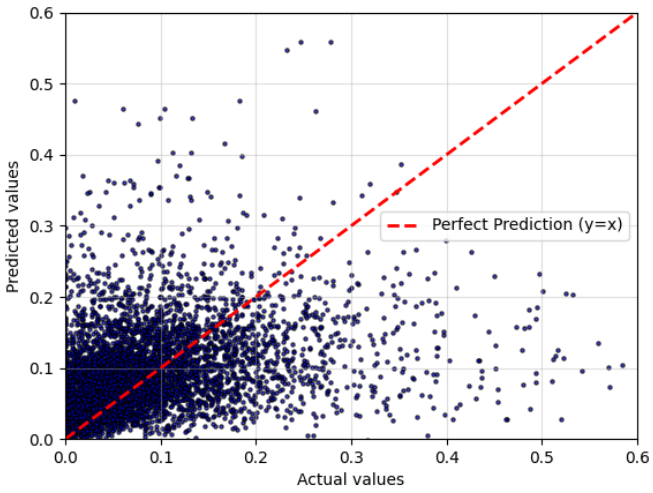


Fig. 5.1: Scatter plot of predicted vs. actual values using LASSO regression on *arabidopsis* 16 °C data.

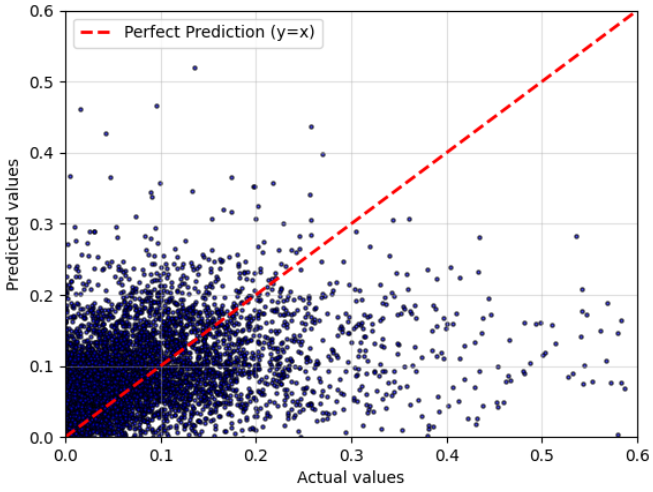


Fig. 5.2: Scatter plot of predicted vs. actual values using CNN on *arabidopsis* 16 °C data.

5.2 Discussion

The final section of this thesis is dedicated to discussion about the insights gained throughout the thesis. It evaluates the performance of the different prediction models and reflects on their strength and limitations. Then it compares the used GP between *arabidopsis* data set and the barley data set, describing the differences projected on the GP performance.

As revealed before, the LASSO model was overall the most performing with the chosen evaluation metrics. The data are lacking good normalization and the hyperparameters of all prediction models need to be further examined and tuned. This is why CNN, even though theoretically most promising, was over exceeded by linear a regression model that is more stable for not enough scaled data. However the difference is small enough to assume better performance after more time of model and data tuning.

Since the *arabidopsis* data are problematic for the different temperatures regimes, the data difference can now be analyzed. All the models predicted best on the data obtained in 16 °C only. This suggests better data consistency and better metabolite activity in the plant cells. The reference growth temperature of *arabidopsis thaliana* is 16 °C and the 6 °C is a stressed condition, so the derived metabolic activity is quite appropriate [36].

The Jax library [48] for CNN is not ideal for more complex model operations, therefore more suitable libraries shall replace Jax [48] in the further study.

Conclusion

This semestral thesis provides exploration of genomic prediction in plants using convolutional neural networks. The goal was to address genomic prediction terminology, prediction methods overview and practical implementations of genomic predictions.

The first part is focused on genomic prediction terminology. It analyzes genotype phenotype relationship and phenotype influence factors. It also describes genome-wide association studies and single nucleotide polymorphisms.

The second part describes various genomic prediction methods, from linear regressions to deep learning. The chapter contains regularized regression, lasso regression, random forest and deep learning models, mentioning feed forward and recurrent neural networks, and importantly convolutional neural network.

The third part comments on the data sets used in the thesis. First data set being *arabidopsis thaliana* and the second *hordeum vulgare*, both obtained from different studies.

The fourth part is focused implementation of chosen prediction models, evaluating their performance with means square error and correlation coefficient. Resulting in choosing the best genomic prediction model at the time and aleborating on further work in the future. It also analyses *arabidopsis* data inconsistency caused by metabolite concentrations measured in two different time and temperature regimes.

The subject of the following bachelor's thesis will be expanding the work to make a solid performing model using convolutional neural network and to elaborate on the differences between two plant data and their impact on prediction performance and model choice.

Bibliography

[1] Tushar Chauhan. An introduction to genome-wide association study. *Genetic Education*, 2019.

[2] Annalisa Lonetti, Maria Chiara Fontana, Giovanni Martinelli, and Ilaria Iacobucci. Single nucleotide polymorphisms as genomic markers for high-throughput pharmacogenomic studies. *Microarray Technology: Methods and Applications*, pages 143–159, 2016.

[3] Pooja Sharma, Nainisha Sharma, Disha Bendale, Mayur Shinde, and Yaminee Patil. Mentalwellness compass for engineering student. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 385–393. Springer, 2024.

[4] Edwin Dong. Reccurent neural networks, revisited. *AI, But Simple*, 2024.

[5] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[6] Li Yu, Kefeng Li, and Xiaoye Zhang. Next-generation metabolomics in lung cancer diagnosis, treatment and precision medicine: mini review. *Oncotarget*, 8(70):115774, 2017.

[7] Suzanne E McGaugh, Aaron J Lorenz, and Lex E Flagel. The utility of genomic prediction models in evolutionary genetics. *Proceedings of the Royal Society B*, 288(1956):20210693, 2021.

[8] Jonathan Gallion, Amanda Koire, Panagiotis Katsonis, Anne-Marie Schoenegge, Michel Bouvier, and Olivier Lichtarge. Predicting phenotype from genotype: Improving accuracy through more robust experimental and computational modeling. *Human mutation*, 38(5):569–580, 2017.

[9] V Orgogozo, B Morizot, and A Martin. The differential view of genotype–phenotype relationships. front genet 6: 179, 2015.

[10] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021.

[11] A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121–1130, 2015.

[12] Katrina M De Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256–261, 2017.

[13] Wei Zhao, Asif Rasheed, Emmi Tikkanen, Jung-Jin Lee, Adam S Butterworth, Joanna MM Howson, Themistocles L Assimes, Rajiv Chowdhury, Marju Orho-Melander, Scott Damrauer, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nature genetics*, 49(10):1450–1457, 2017.

[14] Craig L Hyde, Michael W Nagle, Chao Tian, Xing Chen, Sara A Paciga, Jens R Wendland, Joyce Y Tung, David A Hinds, Roy H Perlis, and Ashley R Winslow. Identification of 15 genetic loci associated with risk of major depression in individuals of european descent. *Nature genetics*, 48(9):1031–1036, 2016.

[15] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.

[16] Anton A Komar et al. Single nucleotide polymorphisms. *Methods in Molecular Biology*, 578(10.1007):978–1, 2009.

[17] Antoni Rafalski. Applications of single nucleotide polymorphisms in crop genetics. *Current opinion in plant biology*, 5(2):94–100, 2002.

[18] Charles R Henderson et al. *Applications of linear models in animal breeding*, volume 462. University of Guelph Guelph, 1984.

[19] Ciaran Michael Kelly and Russell Lewis McLaughlin. Comparison of machine learning methods for genomic prediction of selected arabidopsis thaliana traits. *Plos one*, 19(8):e0308962, 2024.

[20] Vanda M Lourenço, Joseph O Ogutu, Rui AP Rodrigues, Alexandra Posekany, and Hans-Peter Piepho. Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. *BMC genomics*, 25(1):152, 2024.

[21] Huihui Li. Dnngp, a deep neural network-based method for genomic prediction using multi-omics data in plants. In *ASA, CSSA, SSSA International Annual Meeting*. ASA-CSSA-SSSA, 2023.

[22] Nils J Nilson. Introduction to machine learning. *AN EARLY DRAFT OF A PROPOSED TEXTBOOK. Robotics Laboratory. Department of Computer Science Stanford University. Stanford. USA. Recuperado de: http://ai. stanford. edu/˜ nilsson/MLBOOK. pdf*, 1998.

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[24] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4(51-62):56, 2017.

[25] Animesh Acharjee, Richard Finkers, RG Visser, and Chris Maliepaard. Comparison of regularized regression methods for˜ omics data. *Metabolomics*, 3(3):1–9, 2013.

[26] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[27] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

[28] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Random forest for genomic prediction. In *Multivariate statistical machine learning methods for genomic prediction*, pages 633–681. Springer, 2022.

[29] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[30] Osval Antonio Montesinos López, Abelardo Montesinos López, and José Crossa. *Multivariate statistical machine learning methods for genomic prediction*. Springer Nature, 2022.

[31] Arno van Hilten, Steven A Kushner, Manfred Kayser, M Arfan Ikram, Hieab HH Adams, Caroline CW Klaver, Wiro J Niessen, and Gennady V Roshchupkin. Gennet framework: interpretable deep learning for predicting phenotypes from genetic data. *Communications biology*, 4(1):1094, 2021.

[32] Osval Antonio Montesinos-López, Abelardo Montesinos-López, Paulino Pérez-Rodríguez, José Alberto Barrón-López, Johannes WR Martini, Silvia Berenice Fajardo-Flores, Laura S Gaytan-Lugo, Pedro C Santana-Mancilla, and José Crossa. A review of deep learning applications for genomic selection. *BMC genomics*, 22:1–23, 2021.

[33] Miguel Pérez-Enciso and Laura M Zingaretti. A guide on deep learning for complex trait genomic prediction. *Genes*, 10(7):553, 2019.

[34] Daniele Raimondi, Antoine Passemiers, Nora Verplaetse, Massimiliano Corso, Ángel Ferrero-Serrano, Nelson Nazzicari, Filippo Biscarini, Piero Fariselli, and Yves Moreau. Biologically meaningful genome interpretation models to address data underdetermination for the leaf and seed ionome prediction in arabidopsis thaliana. *Scientific Reports*, 14(1):13188, 2024.

[35] Philippe Lamesch, Tanya Z Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L Alexander, Margarita Garcia-Hernandez, et al. The arabidopsis information resource (tair): improved gene annotation and new tools. *Nucleic acids research*, 40(D1):D1202–D1210, 2012.

[36] Jakob Weiszmann, Dirk Walther, Pieter Clauw, Georg Back, Joanna Gunis, Ilka Reichardt, Stefanie Koemeda, Jakub Jez, Magnus Nordborg, Jana Schwarzerova, et al. Metabolome plasticity in 241 arabidopsis thaliana accessions reveals evolutionary cold adaptation processes. *Plant physiology*, 193(2):980–1000, 2023.

[37] Martin Mascher, Todd A Richmond, Daniel J Gerhardt, Axel Himmelbach, Leah Clissold, Dharanya Sampath, Sarah Ayling, Burkhard Steuernagel, Matthias Pfeifer, Mark D'Ascenzo, et al. Barley whole exome capture: a tool for genomic research in the genus hordeum and beyond. *The Plant Journal*, 76(3):494–505, 2013.

[38] Martin Mascher, Heidrun Gundlach, Axel Himmelbach, Sebastian Beier, Sven O Twardziok, Thomas Wicker, Volodymyr Radchuk, Christoph Dockter, Pete E Hedley, Joanne Russell, et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature*, 544(7651):427–433, 2017.

[39] Salvatore Ceccarelli, Stefania Grando, and John Hamblin. Relationship between barley grain yield measured in low-and high-yielding environments. *Euphytica*, 64:49–58, 1992.

[40] Frank Waller, Beate Achatz, Helmut Baltruschat, József Fodor, Katja Becker, Marina Fischer, Tobias Heier, Ralph Hückelhoven, Christina Neumann, Diter von Wettstein, et al. The endophytic fungus piriformospora indica reprograms barley to salt-stress tolerance, disease resistance, and higher yield. *Proceedings of the National Academy of Sciences*, 102(38):13386–13391, 2005.

[41] Ahmed Sallam, Ahmad M Alqudah, Mona FA Dawood, P Stephen Baenziger, and Andreas Börner. Drought stress tolerance in wheat and barley: advances in

physiology, breeding and genetics research. *International journal of molecular sciences*, 20(13):3137, 2019.

[42] Mathias Ruben Gemmer, Chris Richter, Yong Jiang, Thomas Schmutzer, Manish L Raorane, Björn Junker, Klaus Pillen, and Andreas Maurer. Can metabolic prediction be an alternative to genomic prediction in barley? *PLoS One*, 15(6):e0234052, 2020.

[43] Jessica E Rutkoski, Jesse Poland, Jean-Luc Jannink, and Mark E Sorrells. Imputation of unordered markers and the impact on genomic selection accuracy. *G3: Genes, Genomes, Genetics*, 3(3):427–439, 2013.

[44] Python Software Foundation. Python 3.13.0, 2025. URL: `https://www.python.org/downloads/release/python-3130/`.

[45] Wes McKinney et al. pandas: Python data analysis library. 2010. URL: `https://pandas.pydata.org/`.

[46] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. 2020. URL: `https://numpy.org/`.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. 2007-2023. URL: `https://scikit-learn.org/stable/index.html`.

[48] Peter Hawkins Matthew James Johnson Chris Leary Dougal Maclaurin Skye Wanderman-Milne James Bradbury, Roy Frostig. Jax: Composable transformations of python+numpy programs, 2018. URL: `https://github.com/google/jax`.

Symbols and abbreviations

Ath *Arabidopsis thaliana*

GP	Genomic Prediction
DL	Deep learning
ML	Machine Learning
GWAS	Genome-Wide Association Studies
SNP	Single Nucleotide Polymorphisms
DNA	Deoxyribonucleic acid
TAIR	The Arabidopsis Information Resource
BLUP	Best Linear Unbiased Prediction
RR	Ridge Regression
LASSO	Least Absolute Shrinkage and Selection Operator
RF	Random Forest
DT	Decision Tree
FFNN	Feed-forward neural network
RNN	Recurrent neural network
CNN	Convolutional neural network
DNNGP	Deep neural network genomic prediction
LD	Linkage disequilibrium
NAM	Nested association mapping
NaN	Not a Number
MSE	Mean square error
CC	Correlation coefficient

List of appendices

A Contents of attachment

This chapter lists all the attachment files, containing python codes for genomic prediction.