MSc in Bioinformatics - UAB

Genomic Prediction of Complex Traits in Rice:

Leveraging Structural Variation and Applying Deep Learning Methods

Alejandro Navarro Martínez

Tutor: Miguel Pérez Enciso

14th September 2022

CENTRE FOR RESEARCH
IN AGRICULTURAL GENOMICS

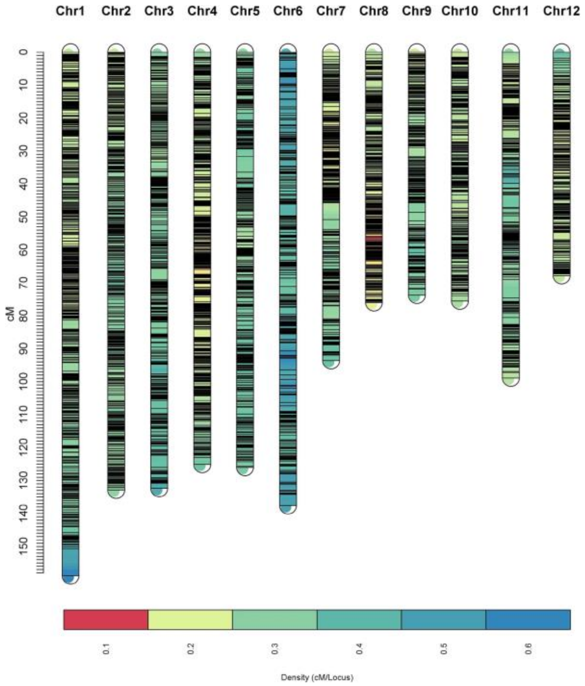
Universitat Autònoma
de Barcelona

Complex traits are phenotypes controlled by many genes and environmental factors, and they often present a continuous variation

↓

Important plant breeding targets!

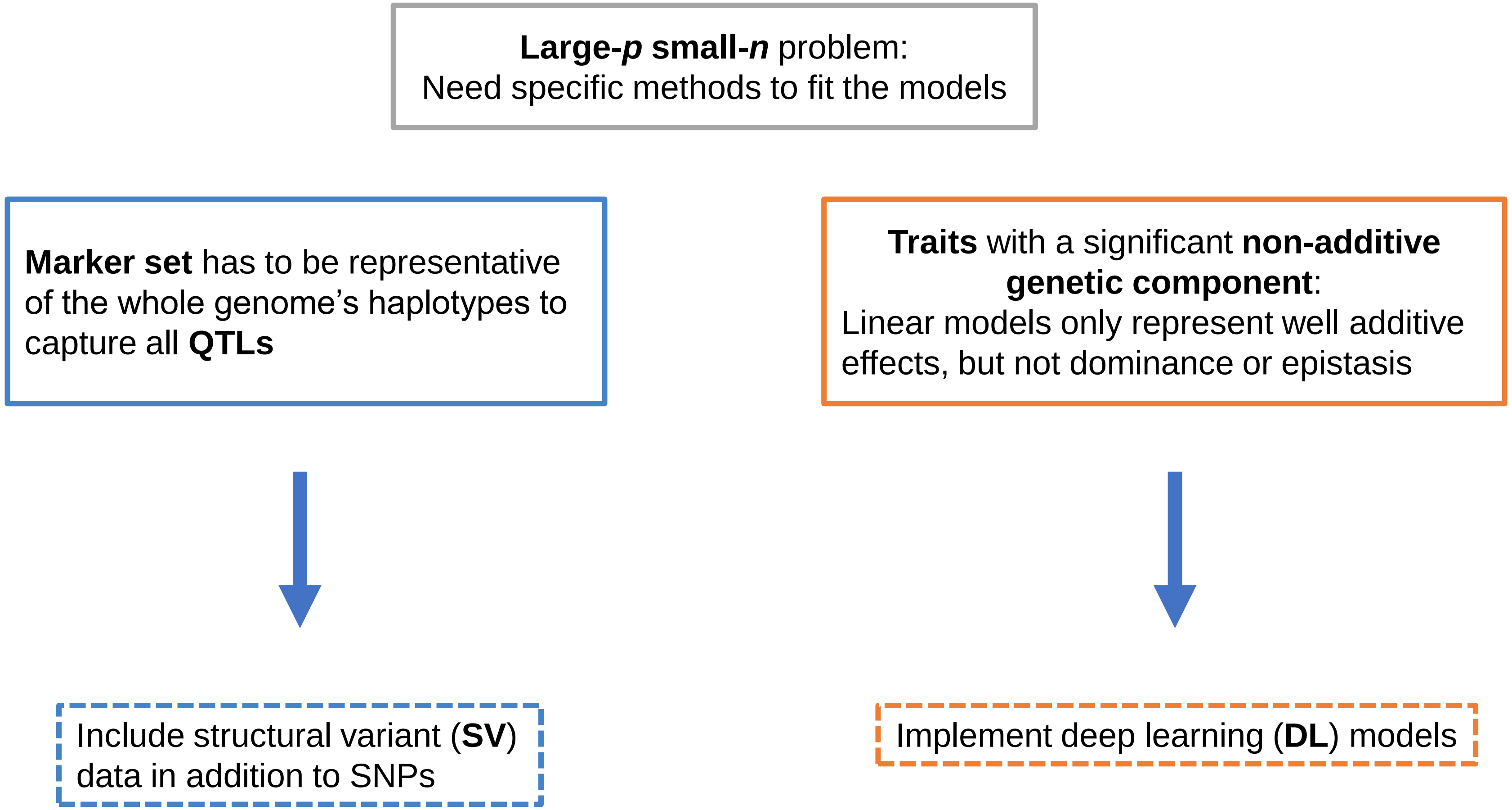
Genomic prediction (GP): to use genome-wide molecular marker data to predict a given trait, exploiting the LD between markers and QTLs



Example of a genetic linkage map of rice

[1] IRRI image collection: https://bit.ly/3qol0TB ; [2] L.M. Salazar/Crop Trust: https://bit.ly/3KYPZiJ ; [3] Qu et al. (2020)

Challenges of genomic prediction:



Large-*p* small-*n* problem:
Need specific methods to fit the models

Marker set has to be representative of the whole genome’s haplotypes to capture all **QTLs**

Traits with a significant **non-additive genetic component**:
Linear models only represent well additive effects, but not dominance or epistasis

Proposed solutions:

Include structural variant (**SV**) data in addition to SNPs

Implement deep learning (**DL**) models

1. Assess if the addition of **structural variation** data improves the prediction of complex traits in comparison to using only SNPs.

SVs
▼
QTL?

2. Evaluate how well **deep learning** performs for genomic prediction in comparison to **linear model methods**.

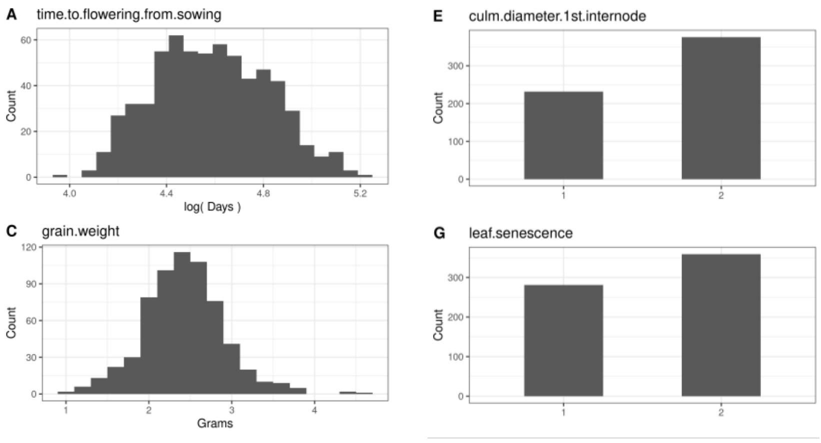
Linear
vs.
DL

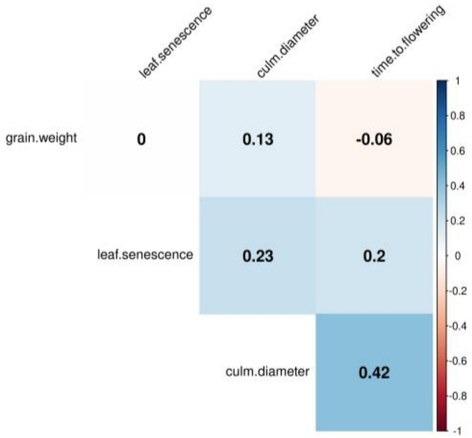
3. Compare the performance of different **deep learning implementations**, considering different input options of genotype data, network architectures and single or multiple output models.

DL
⚙

1. Targets and Features

Targets
(traits)

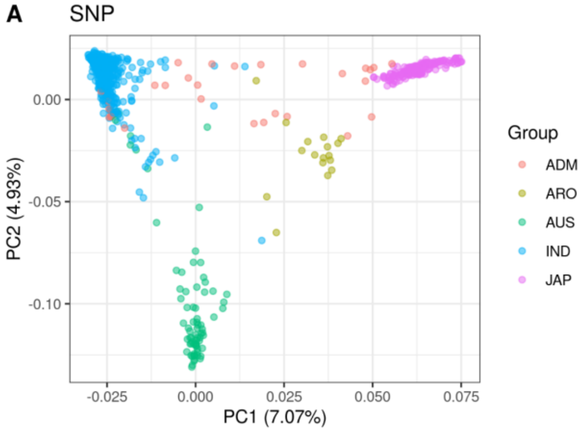


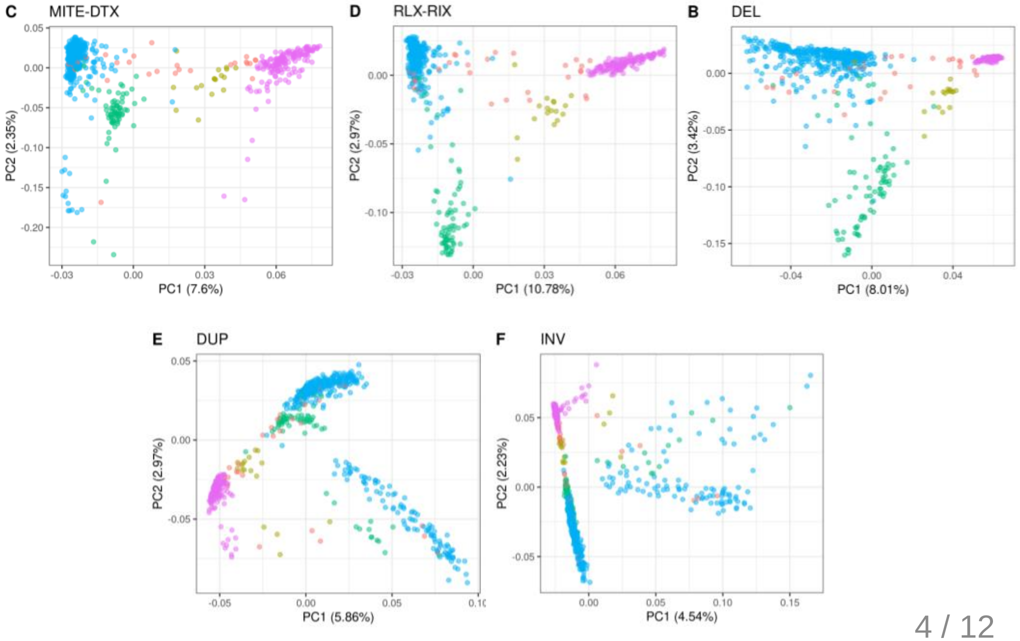


Features
(marker sets)

Marker set	No. of markers
SNP	228,871
MITE-DTX	52,120
RLX-RIX	21,571
DEL	139,229
DUP	14,638
INV	6,083

*MAF > 1%





2. Genomic Variance Inference

Estimate the fraction of phenotypic variance explained by SVs compared to SNPs

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u}_{\text{SNP}} + \mathbf{u}_{\text{SV}} + \mathbf{e}$$

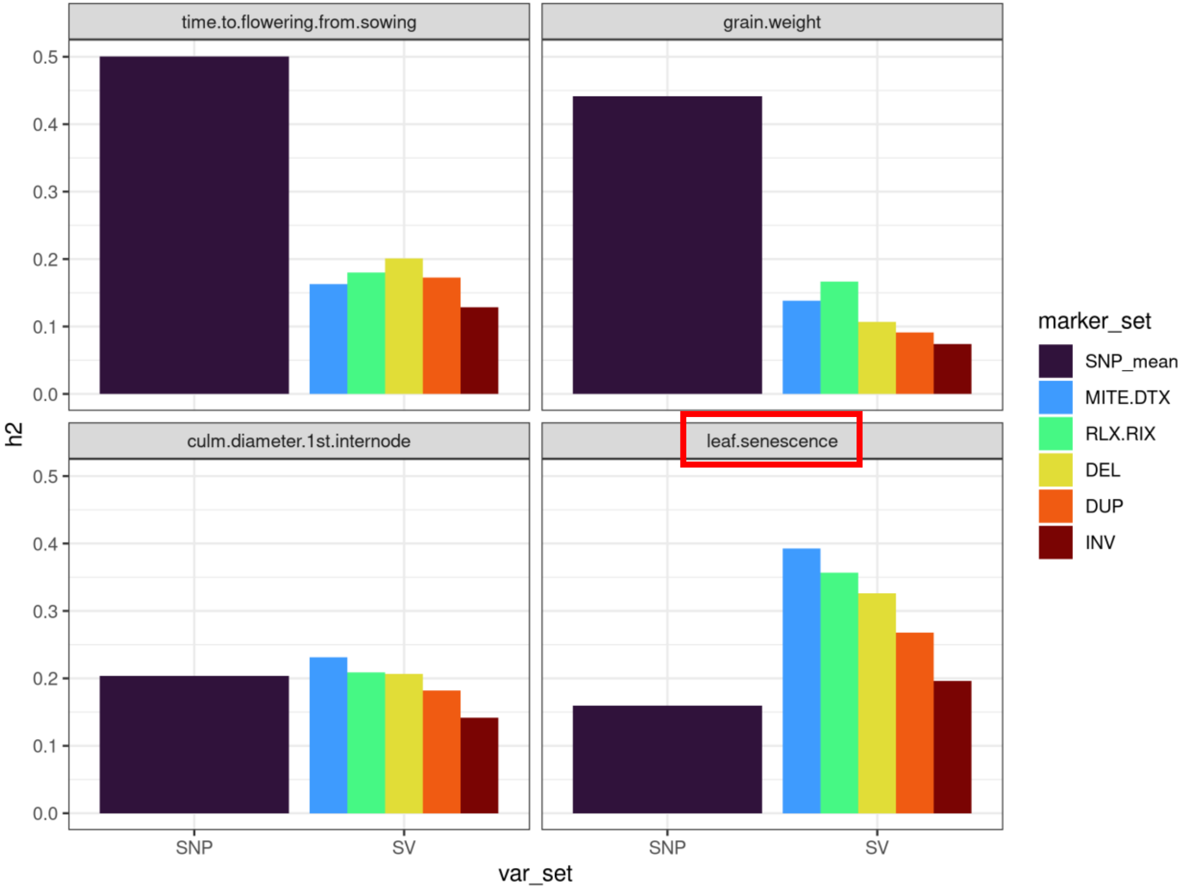
$\mathbf{u}_{\text{SNP}} \sim N(\mathbf{0}, \mathbf{G}_{\text{SNP}}\ \underline{\sigma^2_{\text{SNP}}})$ **G**: Genomic relationship matrix

$\mathbf{u}_{\text{SV}} \sim N(\mathbf{0}, \mathbf{G}_{\text{SV}}\ \underline{\sigma^2_{\text{SV}}})$ SV = {MITE-DTX, RLX-RIX, DEL, DUP, INV}

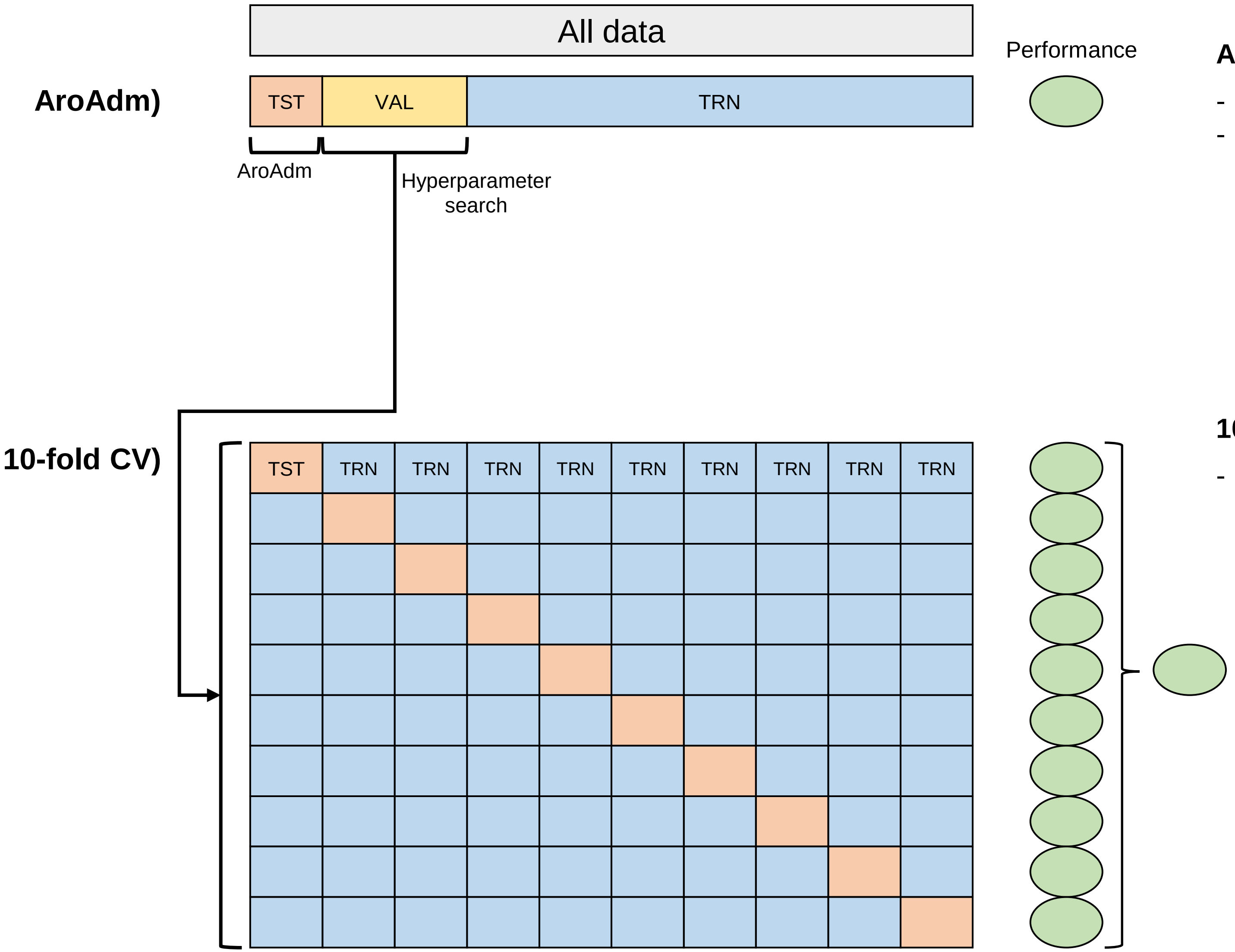
$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\ \underline{\sigma^2_{\text{e}}})$

Bayesian RKHS regression

$$h^2_{\text{marker}} = \frac{\sigma^2_{\text{marker}}}{\sigma^2_{\text{y}}} \quad \text{where} \quad \sigma^2_{\text{y}} = \sigma^2_{\text{SNP}} + \sigma^2_{\text{sv}} + \sigma^2_{\text{e}}$$



1. Frameworks



All data

Performance

AroAdm)

TST | VAL | TRN

AroAdm

Hyperparameter search

10-fold CV)

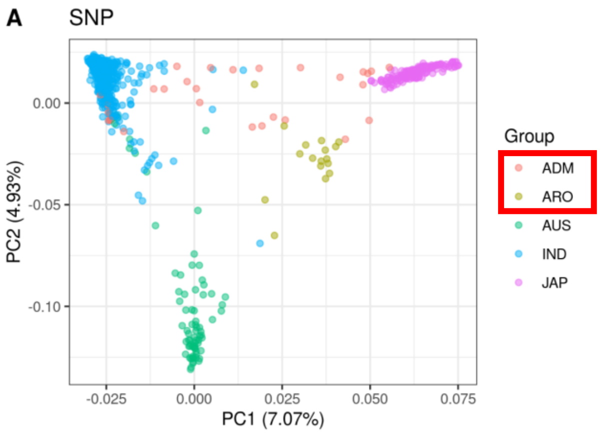
TST | TRN | TRN | TRN | TRN | TRN | TRN | TRN | TRN | TRN

Across populations scenario

- Challenging
- Good estimate of generalization error

10-fold cross-validation

- Estimate of performance less biased by data partitioning



2. Linear Models

Bayesian RKHS

RKHS_SNP $\mathbf{y} = \mathbf{1}\mu + \mathbf{u}_{\mathrm{SNP}} + \mathbf{e}$

RKHS_all $\mathbf{y} = \mathbf{1}\mu + \mathbf{u}_{\mathrm{SNP}} + \mathbf{u}_{\mathrm{MD}} + \mathbf{u}_{\mathrm{RR}} + \mathbf{u}_{\mathrm{DEL}} + \mathbf{u}_{\mathrm{DUP}} + \mathbf{u}_{\mathrm{INV}} + \mathbf{e}$

Bayes C $\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

$\mathbf{X}$: genotypes of the 10k most associated markers

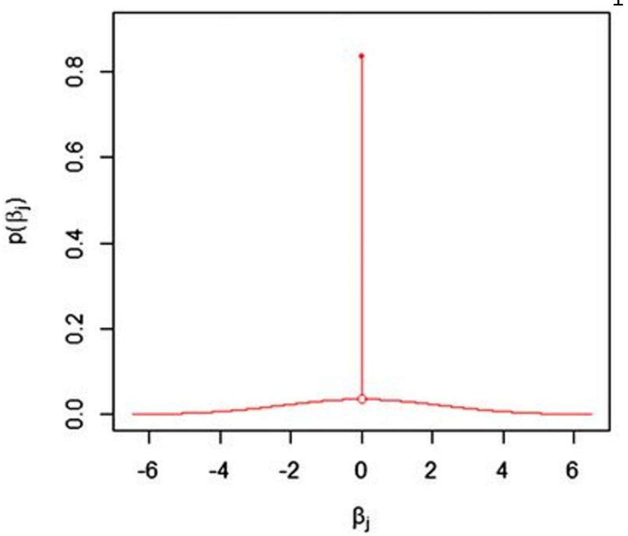
*Different GWAS for each train/test split

BayesC_SNP: top 10k SNPs

BayesC_all: top 10k markers irrespective of type

Shrinkage and selection of marker effects ($\boldsymbol{\beta}$)

$\boldsymbol{\beta} \sim \pi\, N(\mathbf{0},\ \mathbf{I}\sigma^2_{\beta}) + (1-\pi)\,\mathrm{DG}(0)$



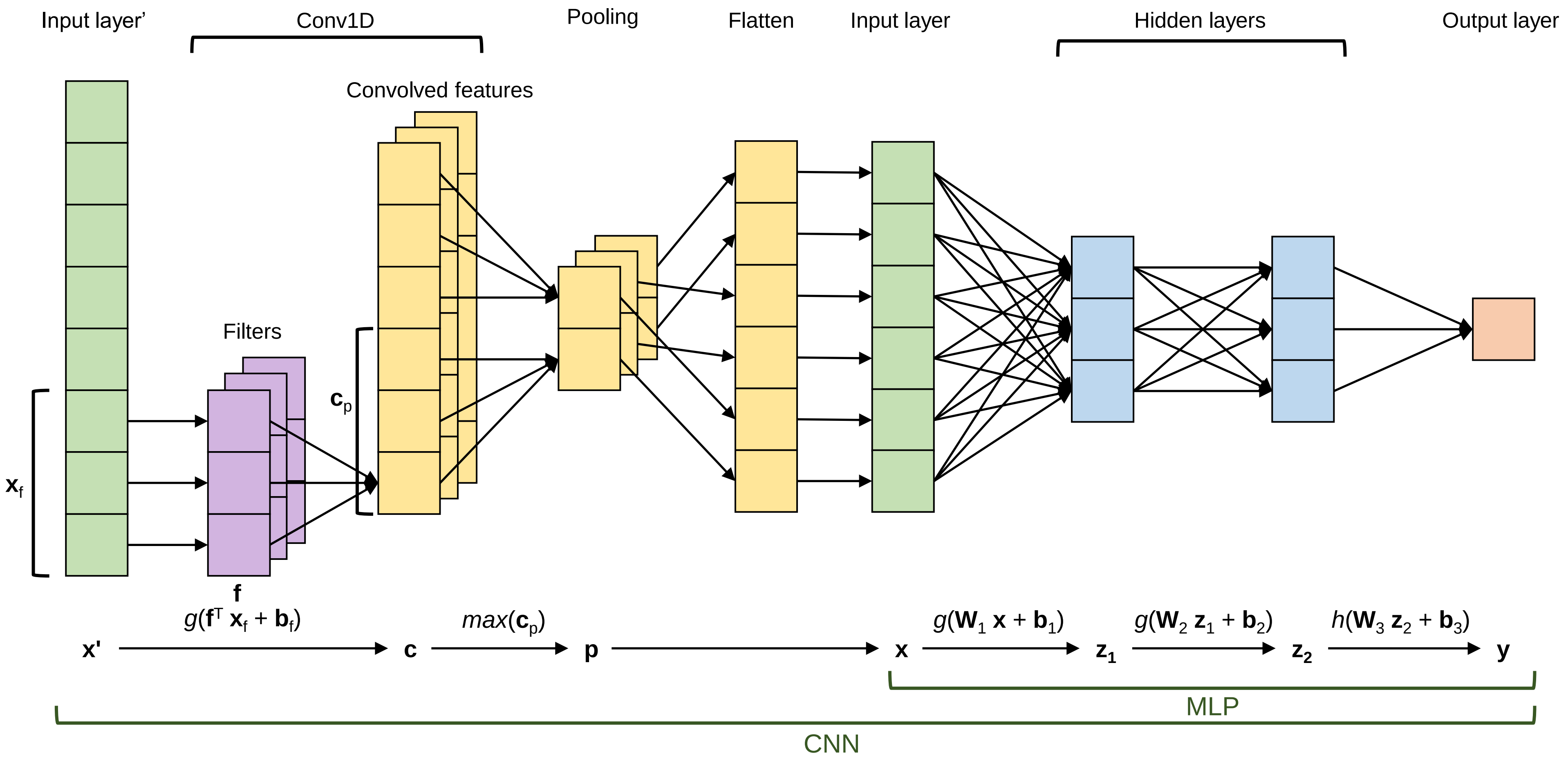
[1]

[1] Modified from de los Campos et al. (2013)

3. ANNs

ANN: multilayer stacks of neurons. Neurons compute non-linear transformations of the weighted sum of an input

Architectures: MLP and CNN



3. ANNs

Input options: top-m and kerPC

top-m

	m_1	m_2		$m_{10,000}$
a_1	0	2		1
a_2	1	0		0
a_n	0	1		0

kerPC **G** matrix $\xrightarrow{\text{eigen()}}$

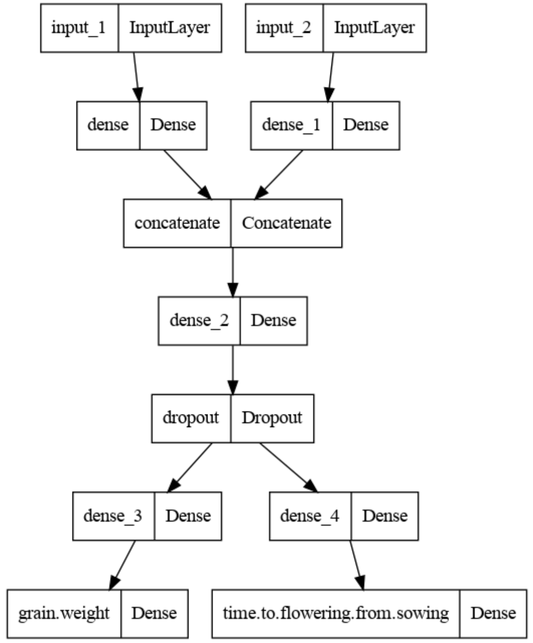
	v_1	v_2		v_{738}
a_1	0.021	-0.019		0.023
a_2	0.027	0.012		0.026
a_n	0.029	-0.005		0.055

SNP — single.in
all — single.in (concatenated) / multi.in

Output number: single.out and multi.out

3 multiple output models:

- 2.cont.traits
- 2.bin.traits
- 4.traits

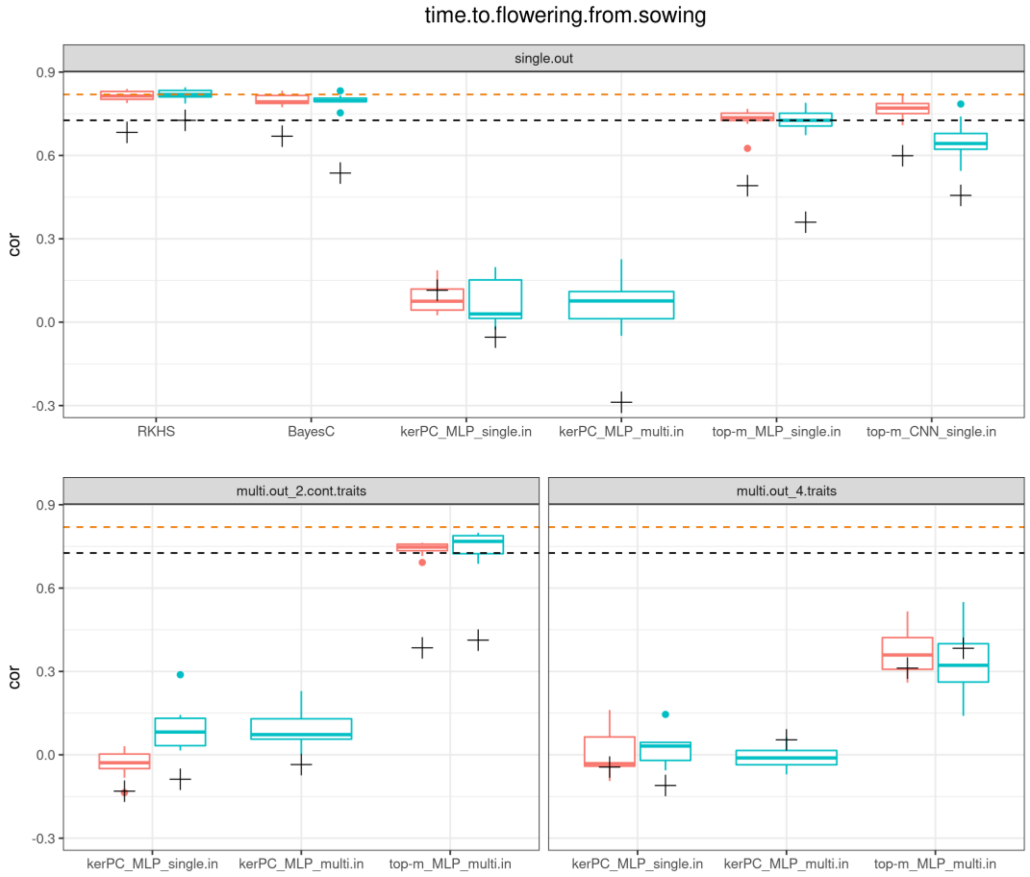


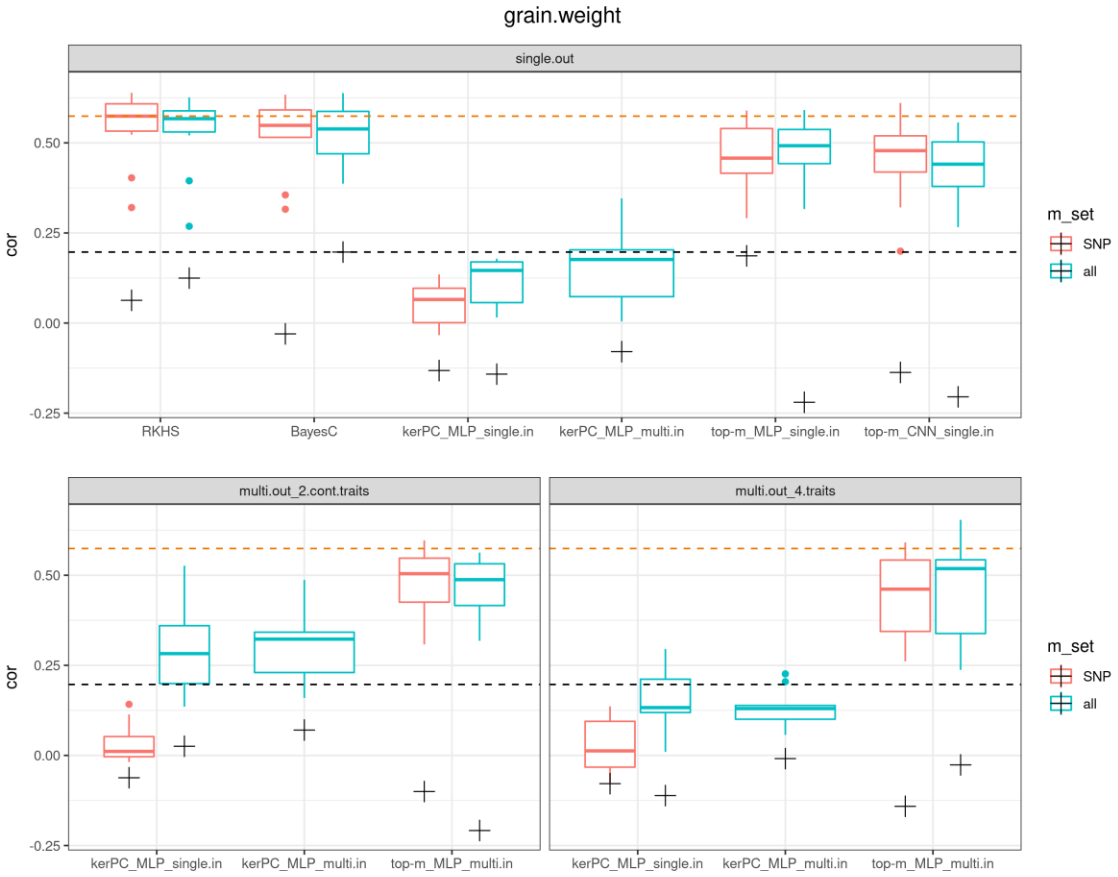
Input option	Architecture	Input number	Output number	Marker sets
kerPC	MLP	single.in	single.out / multi.out	SNP / all
kerPC	MLP	multi.in	single.out / multi.out	all
top-m	MLP	single.in	single.out	SNP / all
top-m	MLP	multi.in	multi.out	SNP / all
top-m	CNN	single.in	single.out	SNP / all

4. Results

Regression: cor (from 0 to 1)

Linear > top-m ANNs > kerPC ANNs





Prediction ✓

Generalization across pops. ✓

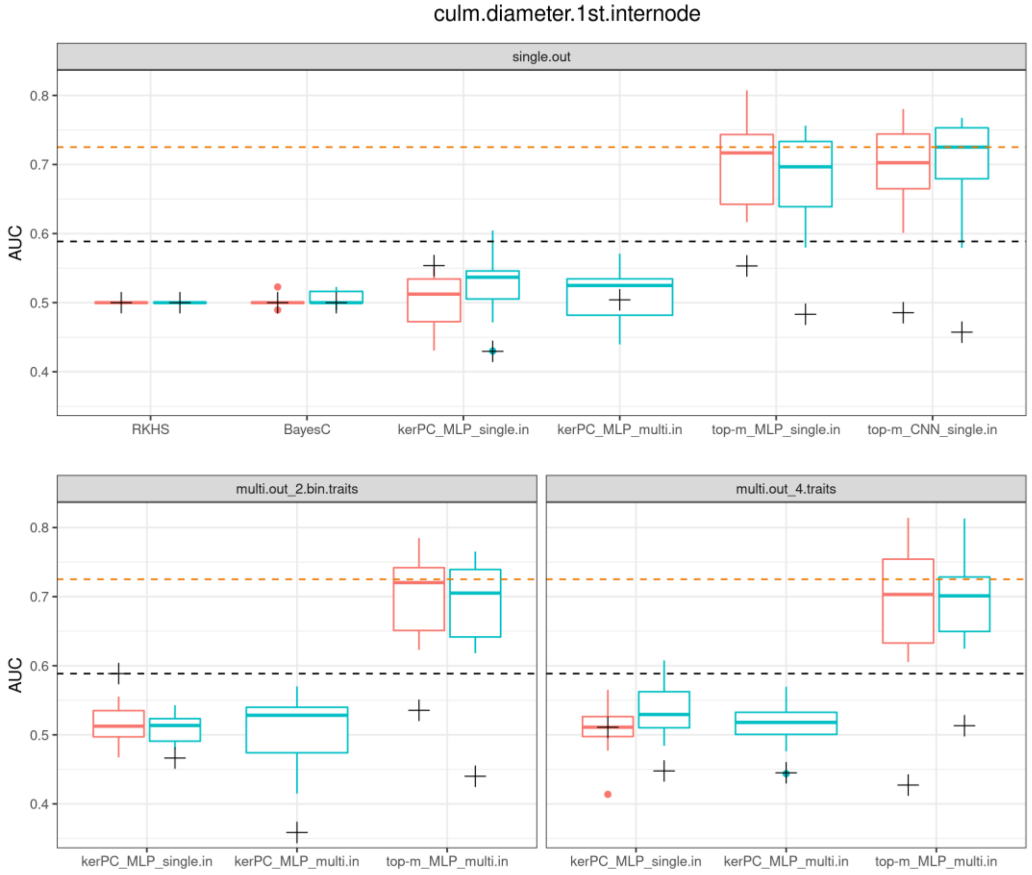
Prediction ~

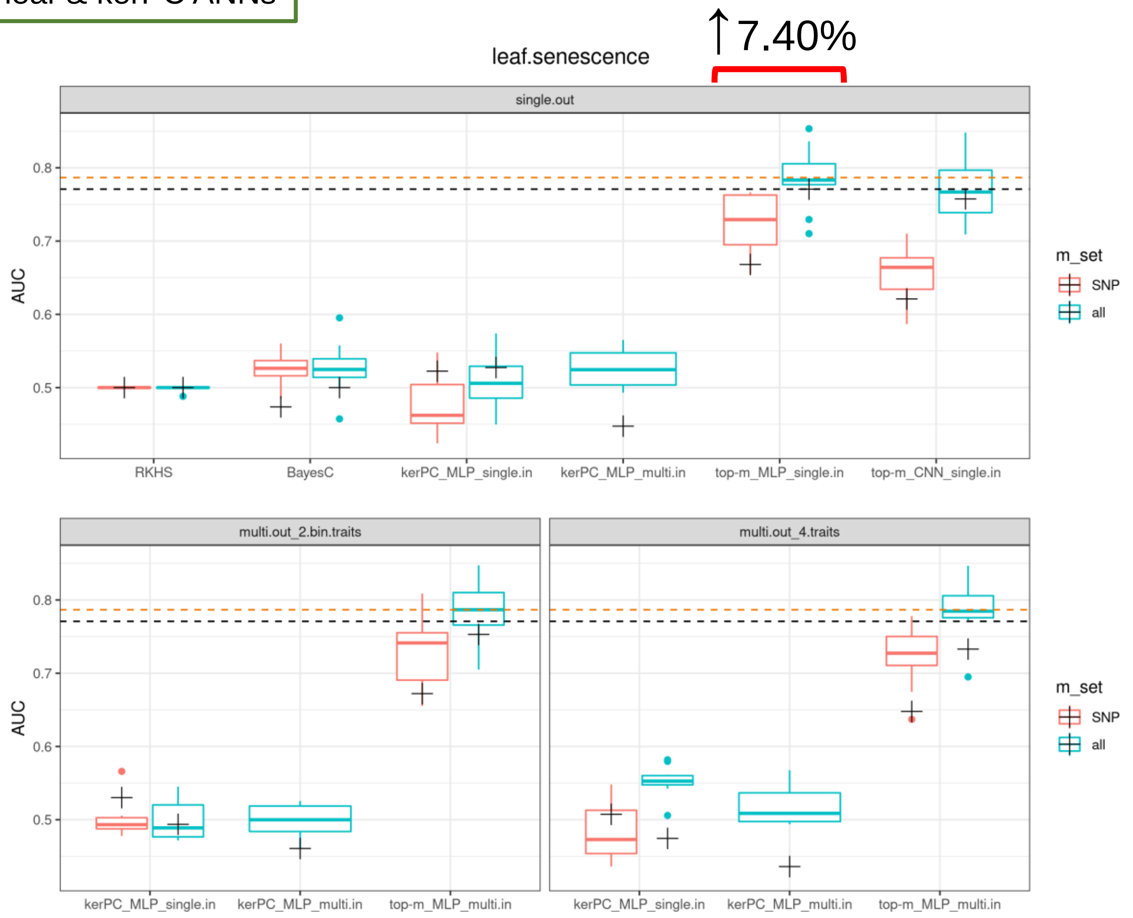
Generalization across pops. ✗

4. Results

Classification: AUC (from 0.5 to 1)

top-m ANNs > Linear & kerPC ANNs





Prediction ~

Generalization across pops. ✗

Prediction ✓

Generalization across pops. ✓

- **SVs can improve GP** for traits in which SVs are a significant source of genetic variance.

SVs
▼
QTL?

- **Linear models** perform very well for **regression**, but less so with classification.

- **ANN models** are competitive with linear models for regression, and they could be a good option for traits with important non-additive genetic effects. They can also deal with **classification** tasks.

Linear
vs.
DL

- The **eigenvectors** of genomic relationship matrices are **not good input options** for ANNs.

- For the traits analyzed here, **CNNs** and **multi-trait** models **do not seem to pose an advantage** versus MLPs and single-trait models, respectively.

DL
⚙

- Prediction across populations is more difficult for grain weight and culm diameter than for time to flowering and leaf senescence.

*

MSc in Bioinformatics - UAB

Genomic Prediction of Complex Traits in Rice:

Leveraging Structural Variation and Applying Deep Learning Methods

Alejandro Navarro Martínez

Tutor: Miguel Pérez Enciso

14th September 2022

crag
CENTRE FOR RESEARCH
IN AGRICULTURAL GENOMICS

UAB
Universitat Autònoma
de Barcelona

Citations and assets

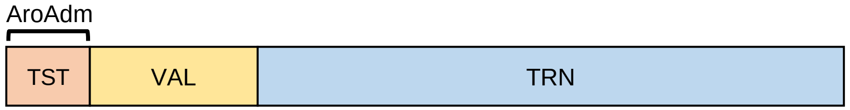
Qu P, Shi J, Chen T, Chen K, Shen C, Wang J, Zhao X, Ye G, Xu J, Zhang L. 2020. Construction and integration of genetic linkage maps from three multi-parent advanced generation inter-cross populations in rice. Rice. 13(1):1–16. doi:10.1186/s12284-020-0373-z.

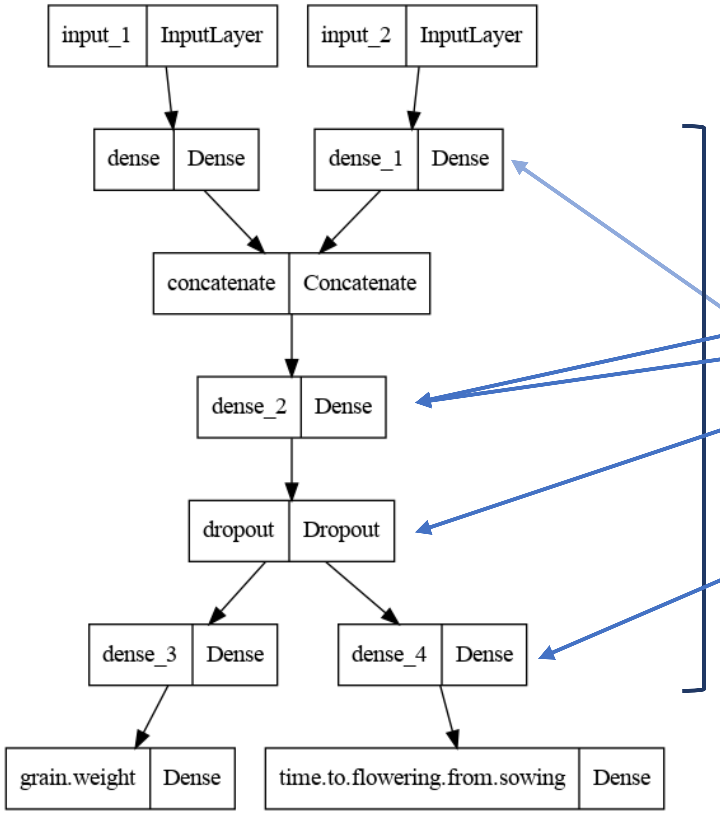
de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics. 193(2):327–345. doi:10.1534/genetics.112.143313.

Presentation designed using resources from Flaticon.com.

3. ANNs

Hyperparameter tuning





Hyperparameter	Values	Models
activation	['relu', 'tanh', 'linear', 'softplus']	All
regularization_type	['L2', 'L1']	All
regularization_rate	[0.0, 0.1, 0.01, 0.001]	All
n_layers	'min_value': 1, 'max_value': 5, 'step': 1	All
units_x	[10, 20, 40, 80, 160]	All
dropout_rate	'min_value': 0.0, 'max_value': 0.3, 'step': 0.05	All
first_units	[10, 20, 40, 80, 160]	kerPC_MLP_multi.in, top-m_MLP_multi.in
last_units	[10, 20, 40, 80, 160]	top-m_MLP_multi.in, (kerPC_MLP_single.in), (top-m_MLP_single.in)
n_filters	[16, 32, 64, 128]	top-m_CNN_single.in