Graduate Research Fellowship Program


Shin-Han Shiu


Gustavo de los Campos

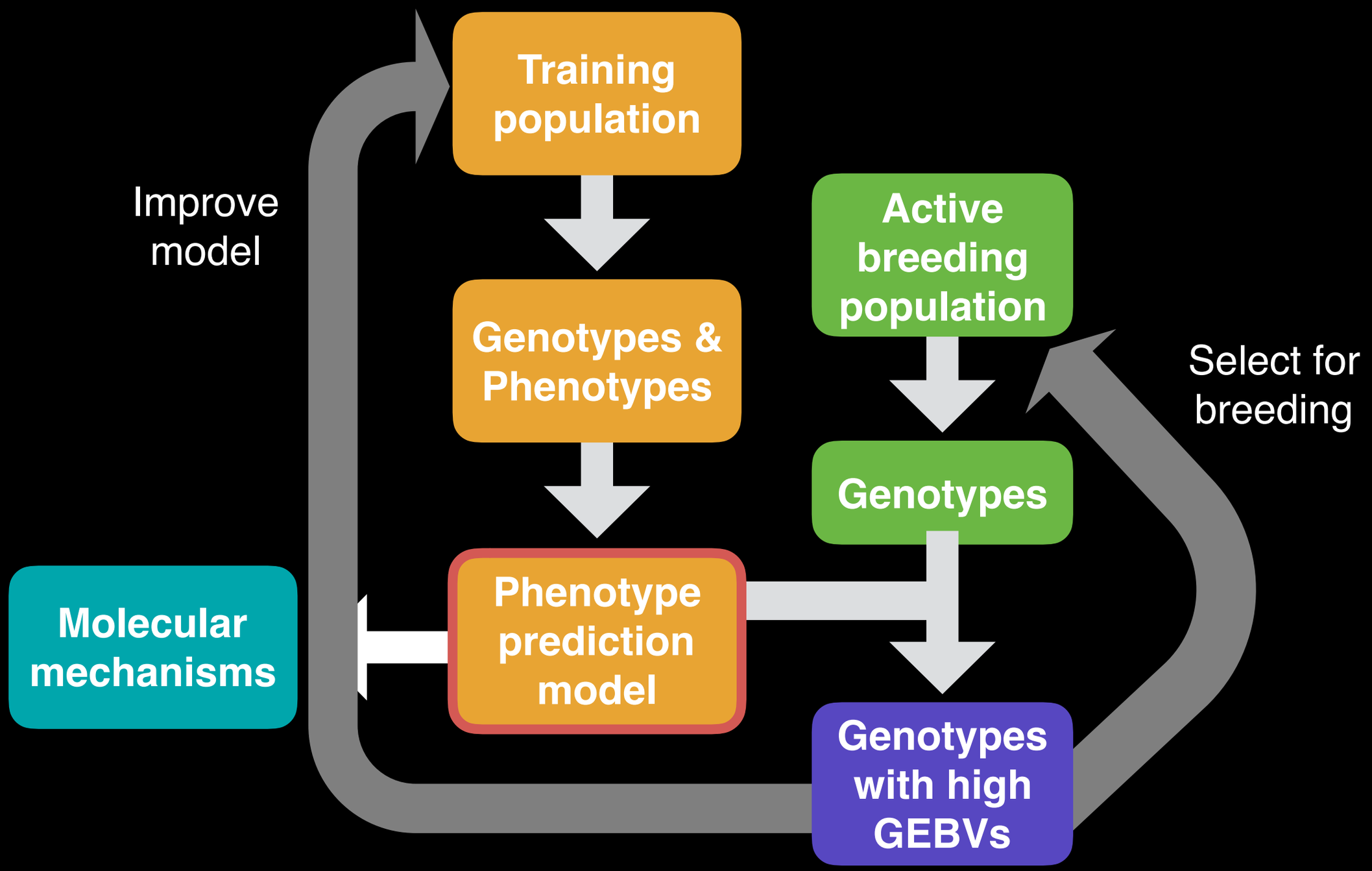

Andrew McCarren


Mark Roantree

Benchmarking algorithms for genomic prediction of complex traits in plants

Christina Azodi
Department of Plant Biology
February 26th, 2019
QuantGen Lab Meeting

Genomic Prediction: Breeding & Basic Science



GEBV: genomic estimated breeding value

Genomic Prediction Models

Quantitative phenotype of individual i

SNP genotype of individual i

$$y_i = g(x_i) + e_i$$

Residual

Function relating genotype to phenotype

Performance Metric (*r*)

$$\mathrm{cor}(y, \hat{y})$$

Outline

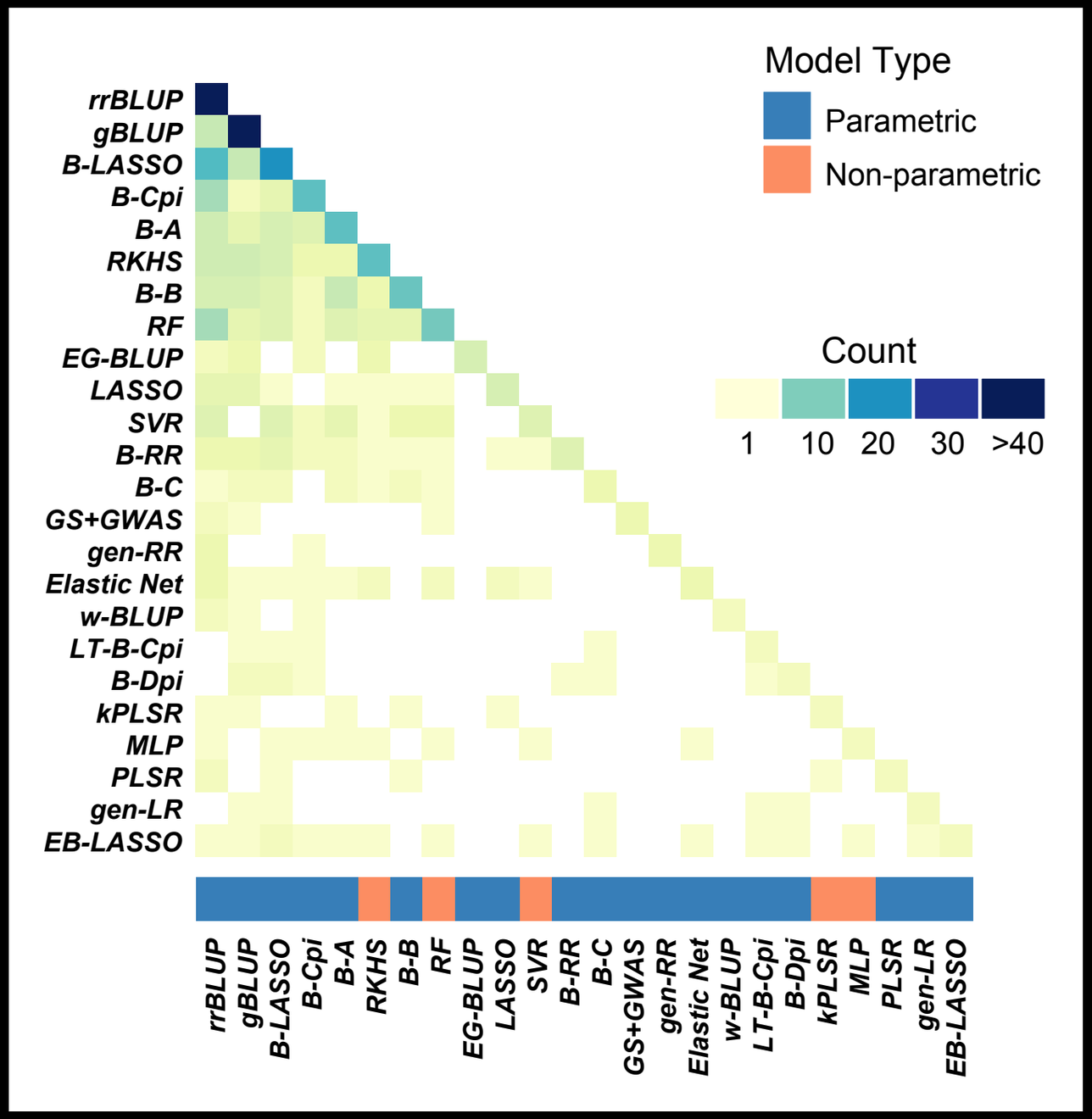
- Data and algorithms included in benchmark
- Intro to Neural Networks
- Performance on predicting height in 6 species
- Improvements to artificial neural networks
    - Feature selection and seeded starting weights
- Final results

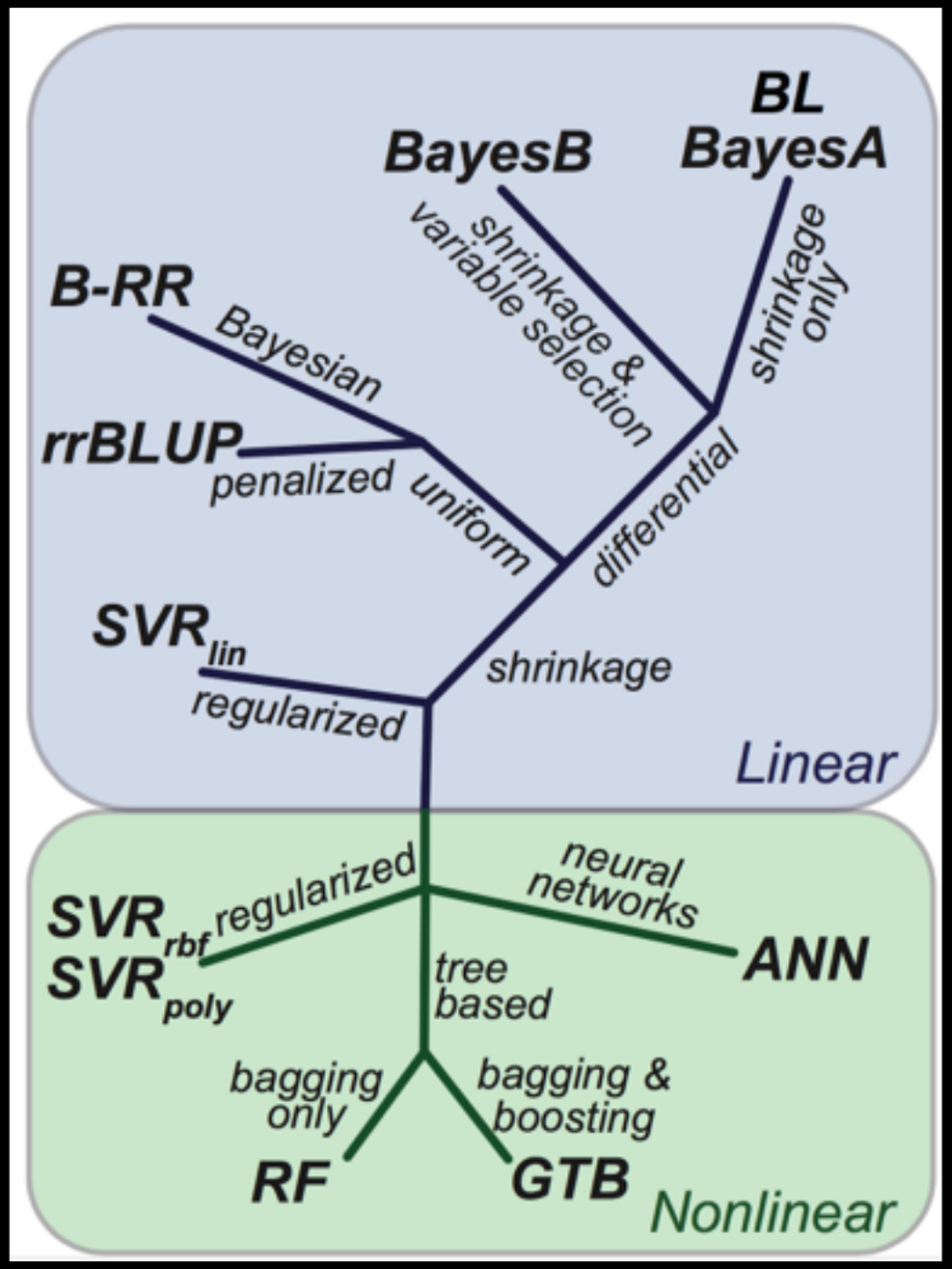
Data Used

Species	Population	Markers	Source
Maize	391	332,178	Hansey et. al 2011 Hirsch et. al 2014
Rice	327	73,147	Spindel et. al 2015
Sorghum	451	58,961	Fernandes et. al 2017
Soy	5,014	4,240	Xavier et. al 2016
Spruce	Partial DM 1,722	6,932	Beaulieu et. al 2014
Switchgrass	514	218,528	Lipka et. al 2014 Evans et. al 2017

Diversity Panel NAM GBS SNP-chip

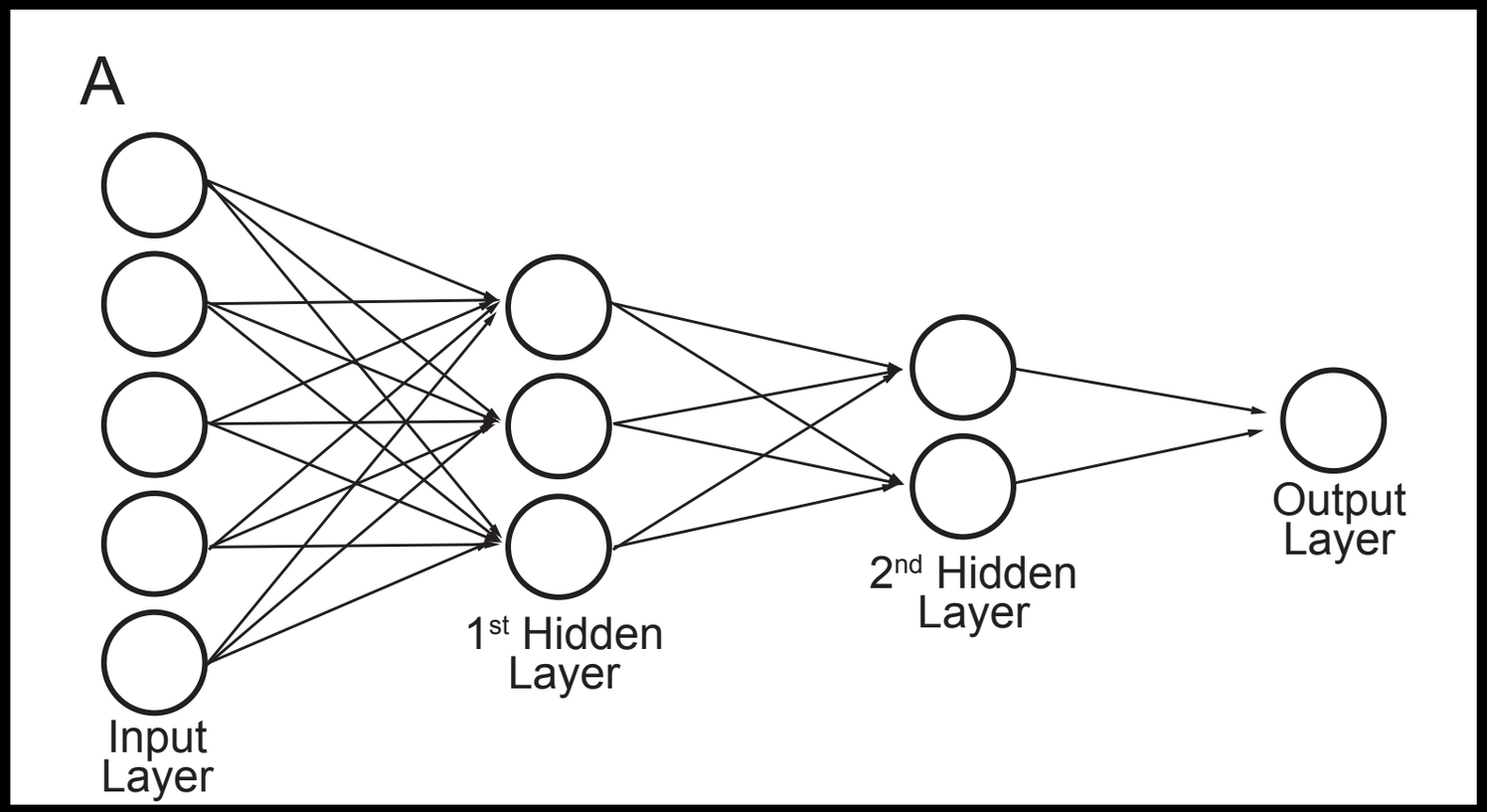
GP Algorithm Comparisons





rrBLUP: ridge regression Best Linear Unbiased Predictor. BL: Bayesian Least Absolute Shrinkage and Selection Operator. SVR: Support Vector Regression. RF: Random Forest. ANN: Artificial Neural Network. GTB: Gradient Tree Boosting.

What is a Neural Network?



Difficulties

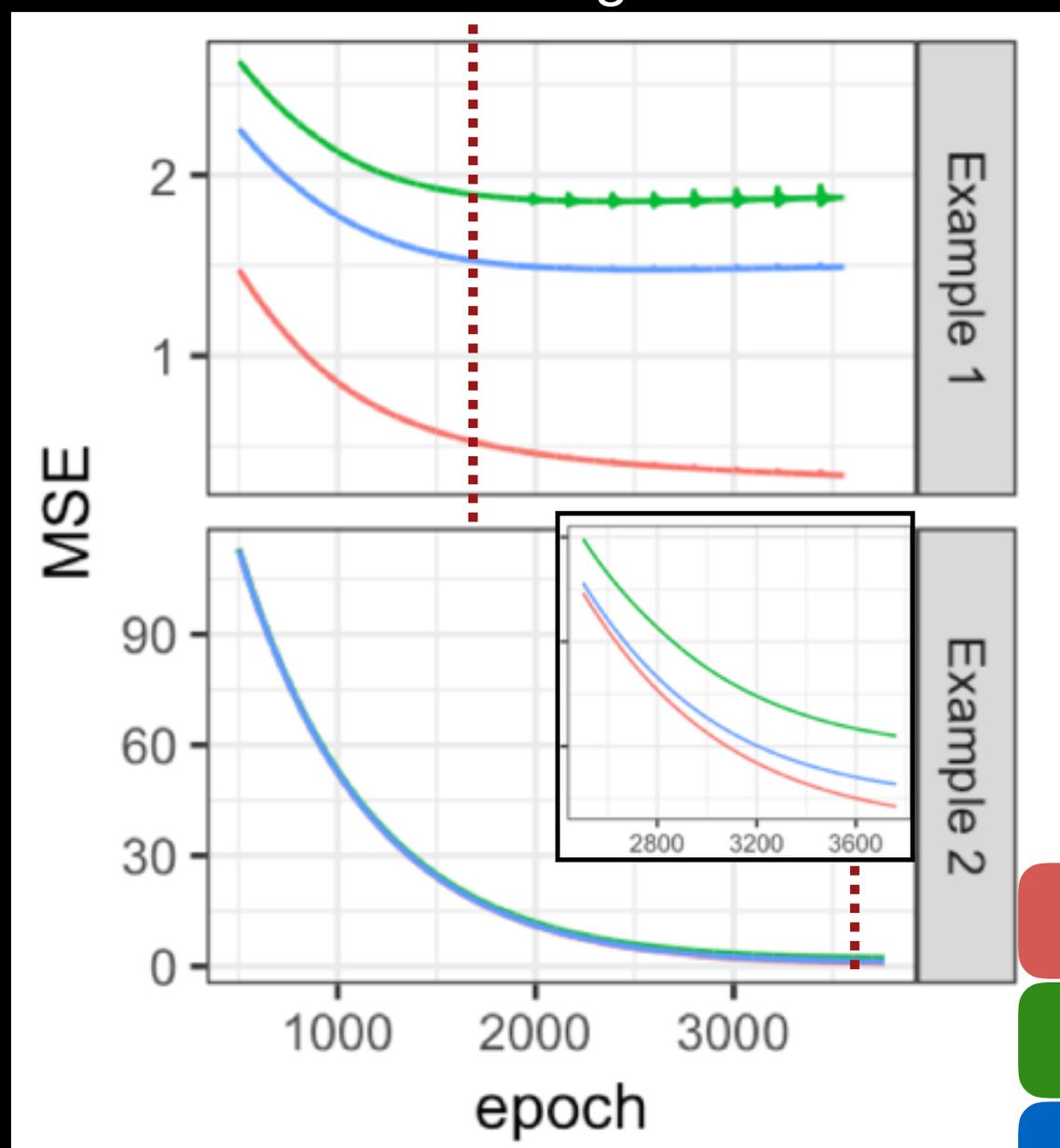
1. Large hyper-parameter space
2. When to stop training?

1. Grid Search

- Network Architecture
- Activation Function
- Learning Rate
- Type/degree of regularization

2. When to stop training?

Models can have very different ideal training times



Minimize validation error

Burn-in

↓

$$\frac{\text{MSE}_{\text{valid}}\ \text{Epoch X}}{\text{MSE}_{\text{valid}}\ \text{Epoch X+1}} < 0.01$$

x 10

↓

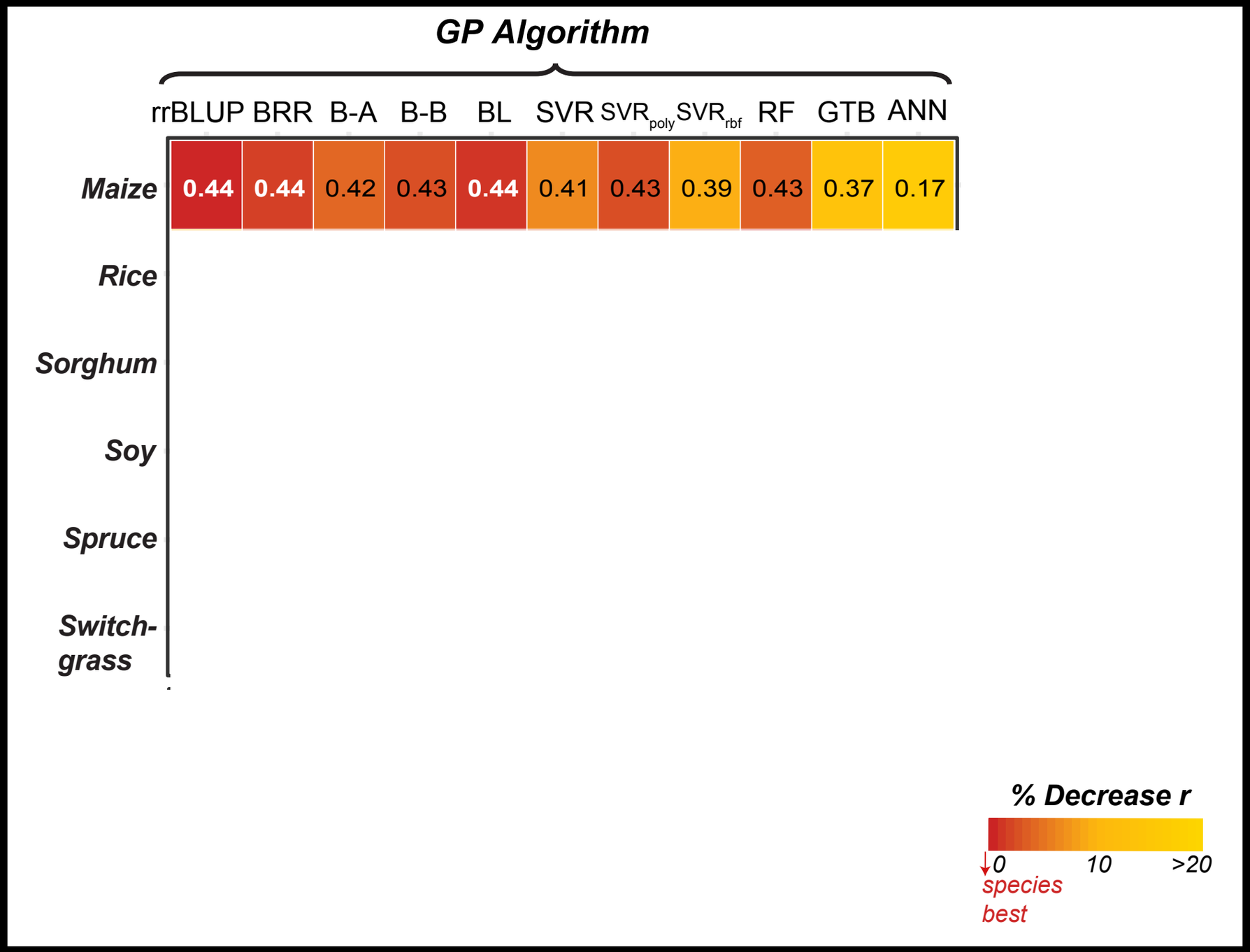
Stop Training

Training

Validation

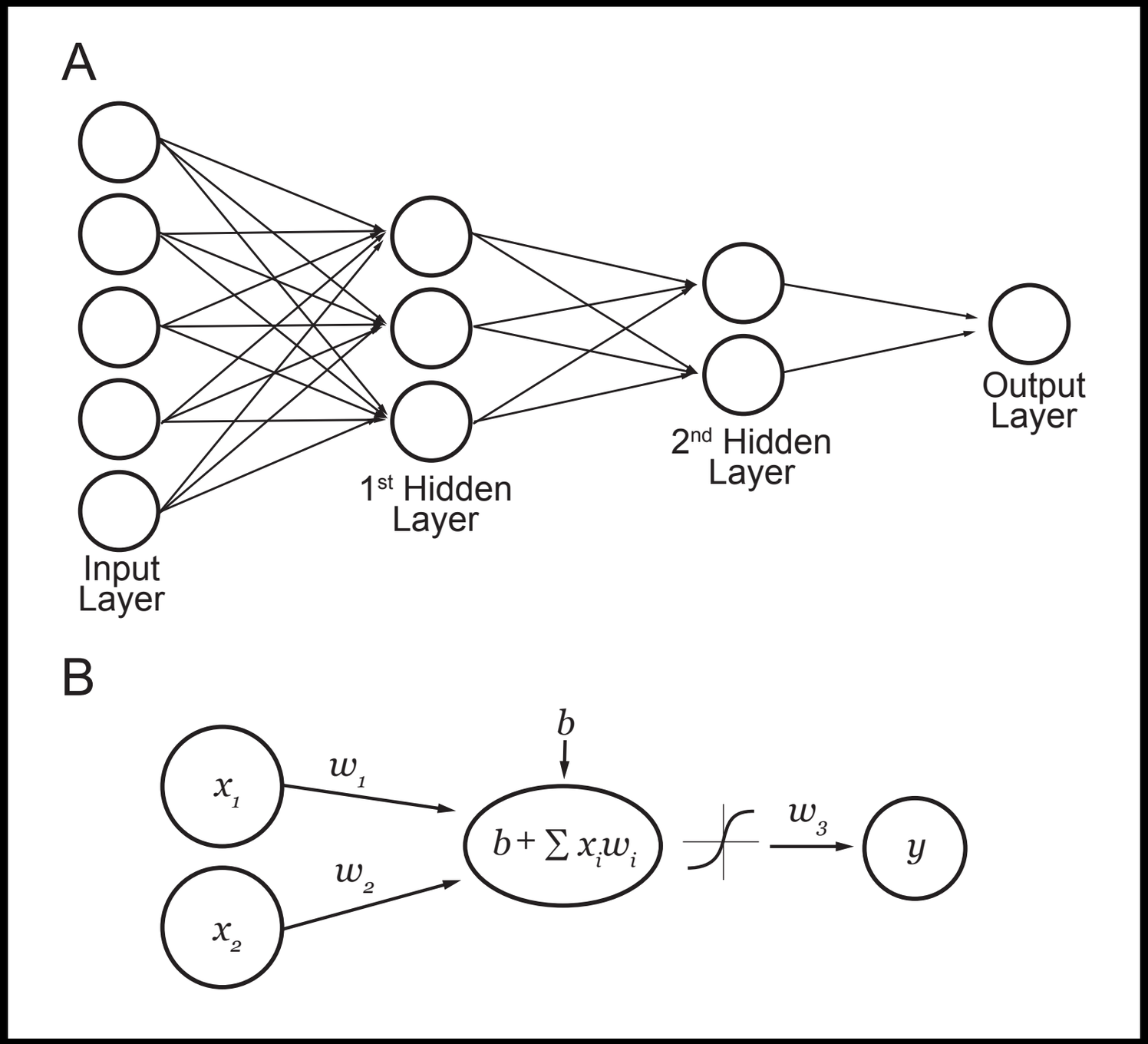
Testing

No one algorithm always performs best



model performance for predicting height in each species

What is a Neural Network?



Difficulties

1. Large hyper-parameter space
2. When to stop training?
3. Not robust when p >> n
4. Large variation in model performance using same data and hyper-parameters.

3. Addressing p >> n with feature selection

1. Feature Selection

Training & Validation Data

Feature Selection

Random Forest
Bayes A
Elastic Net

2. Select hyperparameters & train model

Training Data

Grid Search

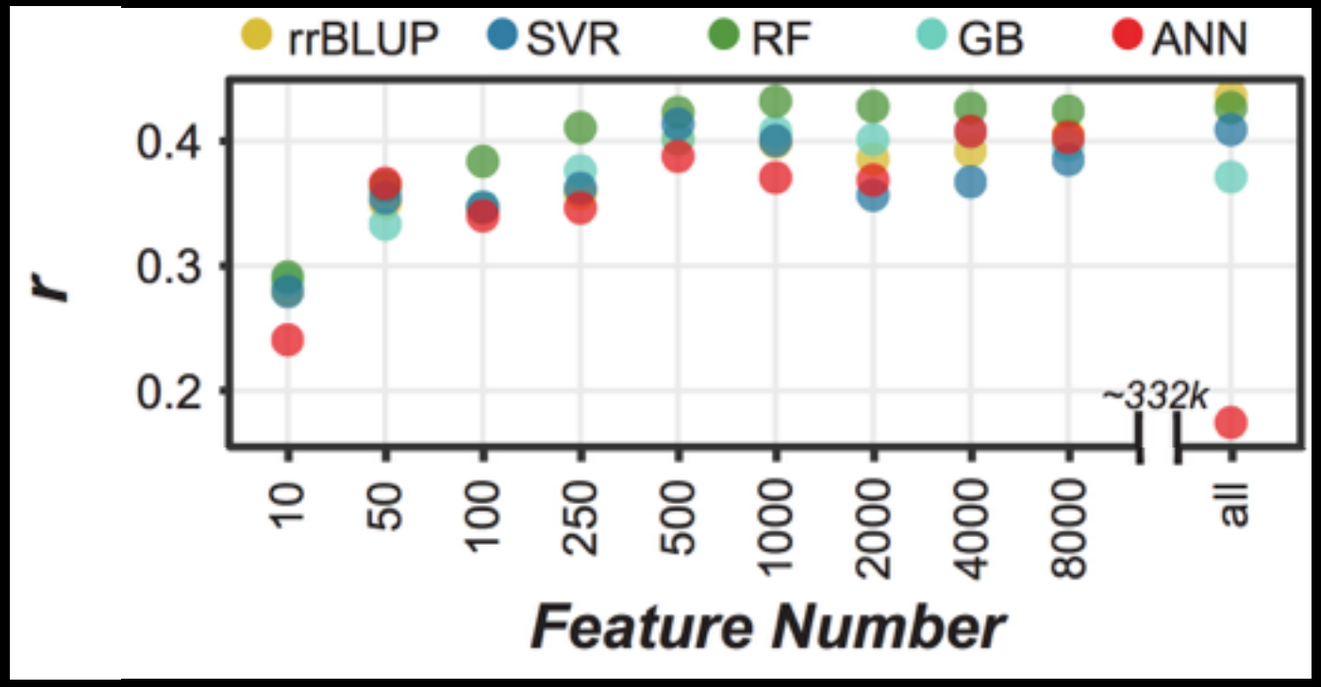
Validation Data

Train model

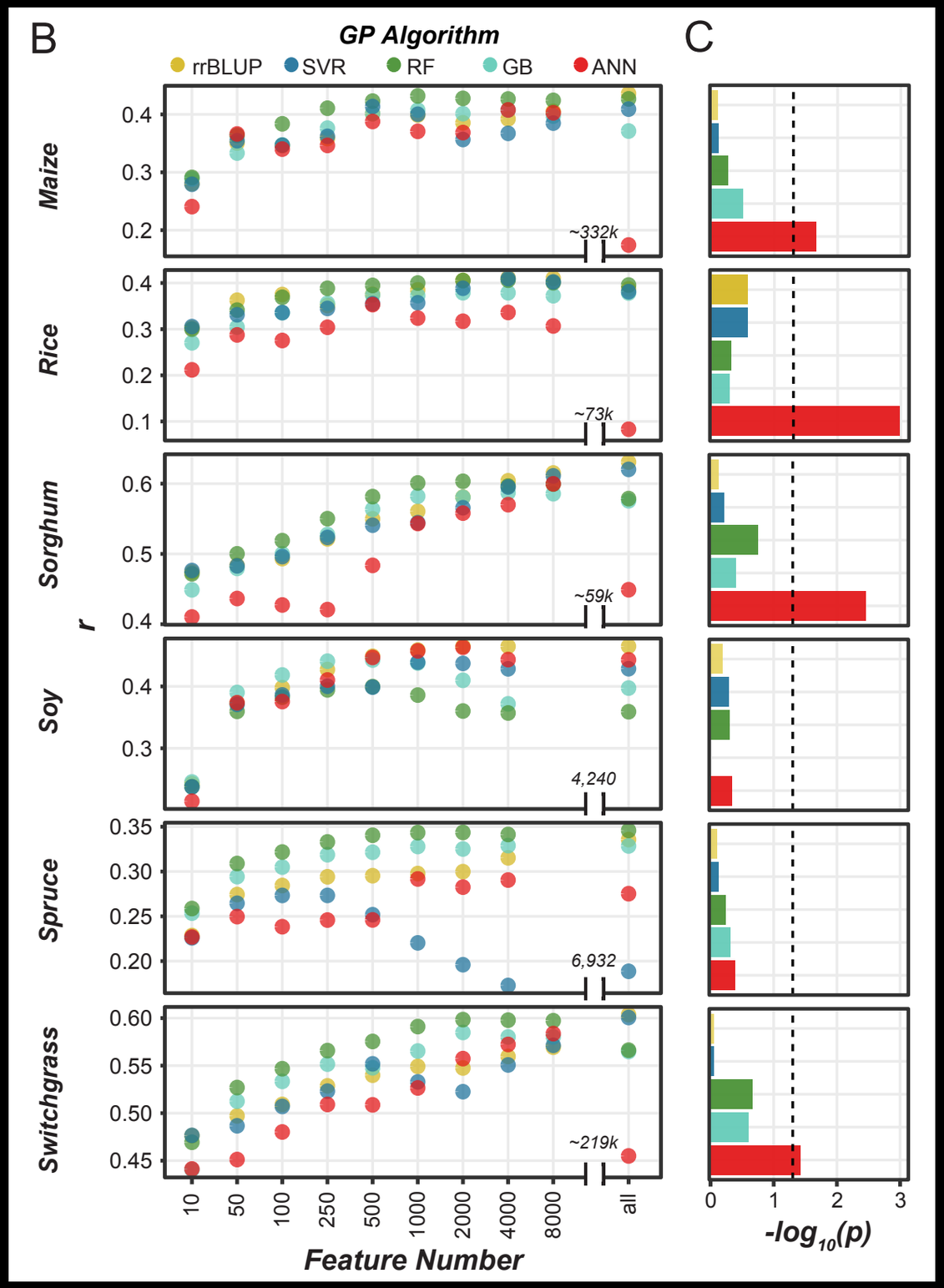
3. Performance

Testing Data

Height in maize



Dimension reduction improves ANN performance



p:n ratio

850

224

131

0.85

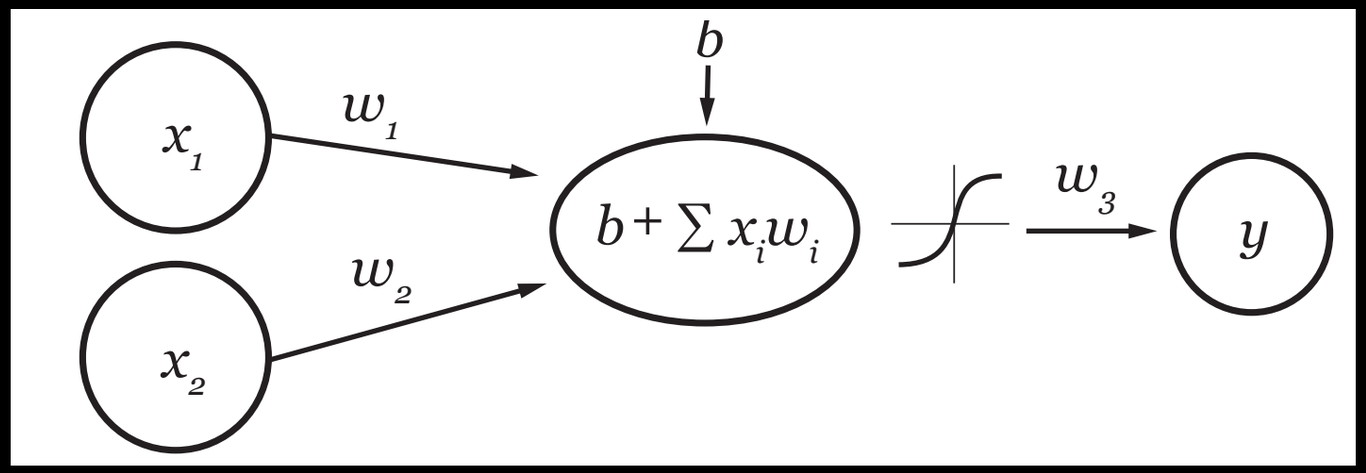
4

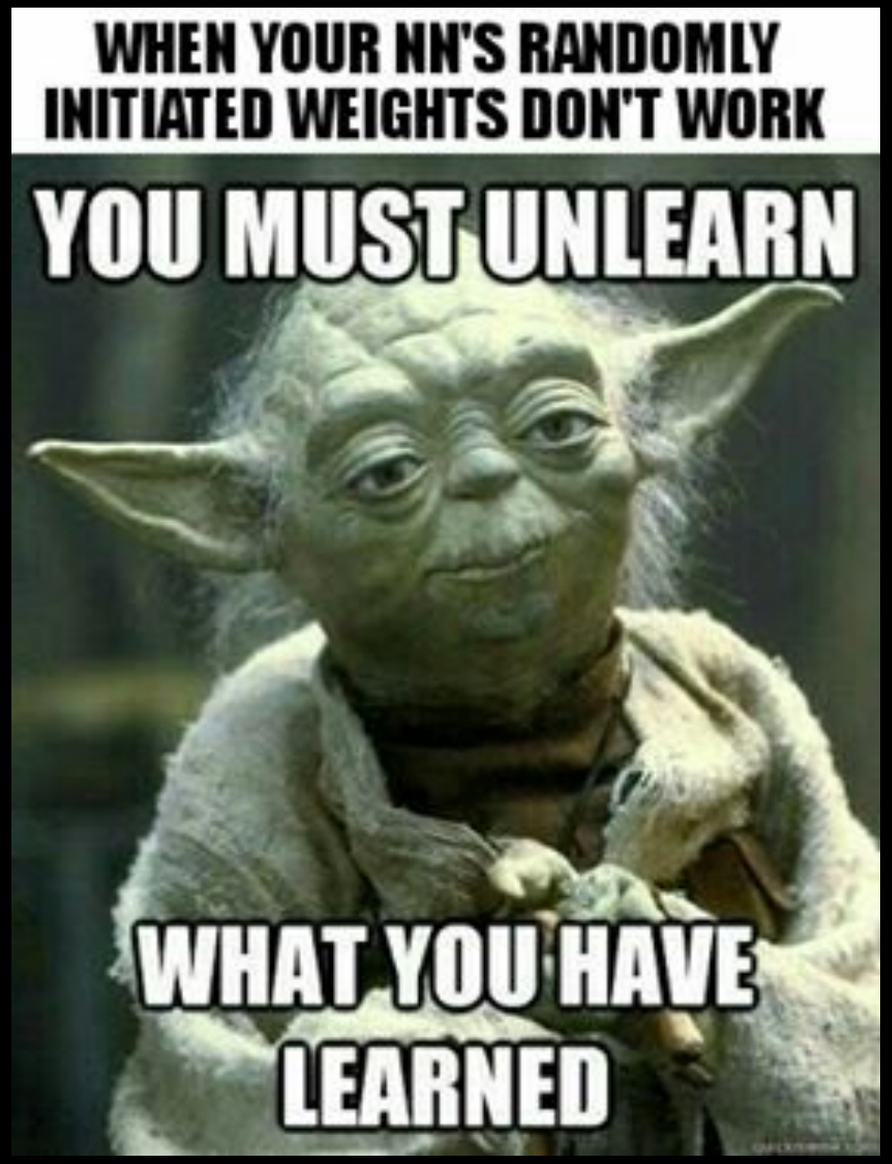
425

Difficulties

1. Large hyper-parameter space
2. When to stop training?
3. Not robust when p >> n
4. Large variation in model performance using same data and hyper-parameters.

Random initialization of starting weights



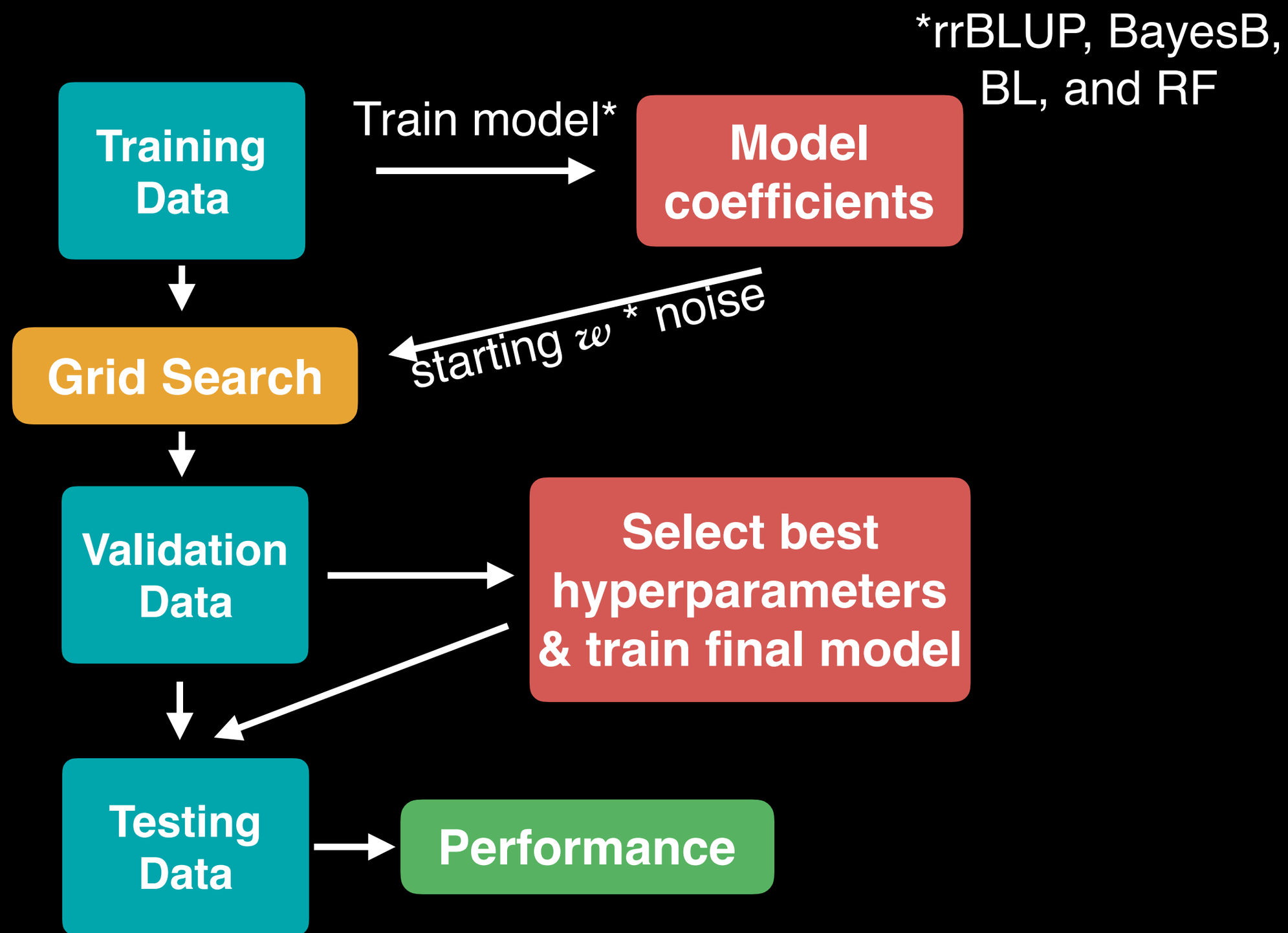


Random Starting Weights

	Node 1	Node 2	Node 3
SNP A	-0.2	0.8	-0.2
SNP B	-0.4	1.2	0.6
SNP C	0.1	0.6	0.8
SNP D	0.4	-0.1	-0.5

****Reduces bias in the model****

Seeded starting weights approach



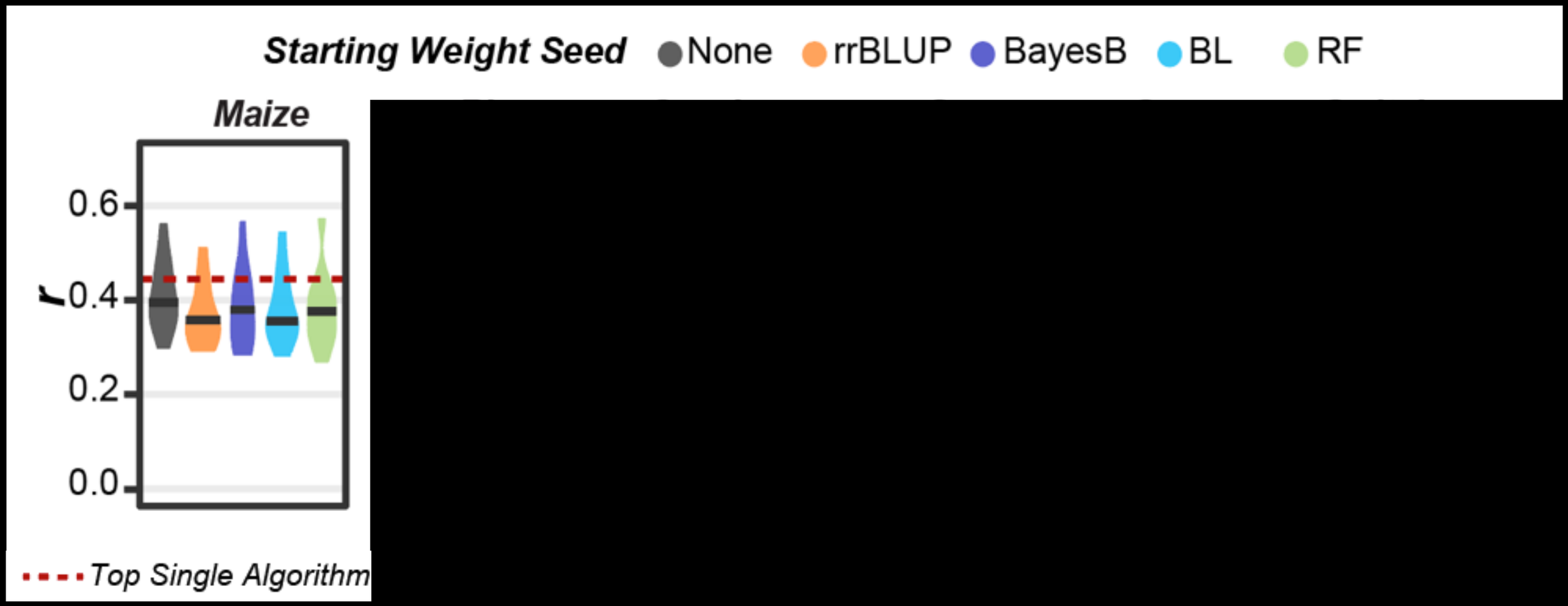
Seeded starting weights approach

25%
Seed Score +
Noise Infusion

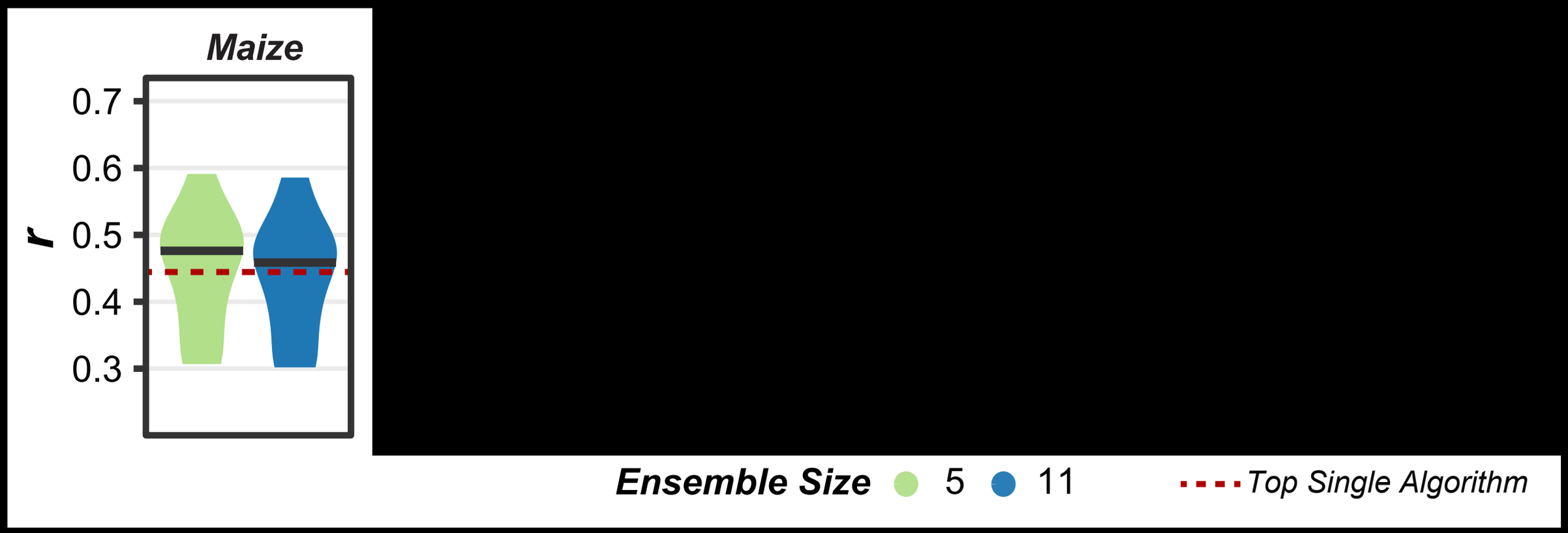
	rrBLUP coef.
SNP A	-0.6
SNP B	-0.4
SNP C	0.1
SNP D	0.4

	Node 1	Node 2
SNP A	-0.7	-0.8
SNP B	-0.5	-0.4
SNP C	0.2	0.1
SNP D	0.4	0.5

Seeded starting weights for top 8k maize markers sometimes improves performance



Ensemble Predictors: mean predicted trait value from 5/11 GP algorithms

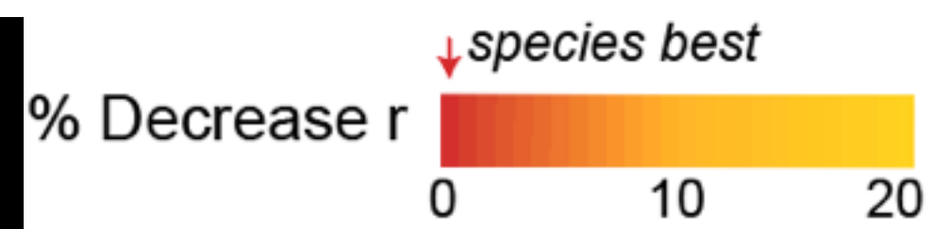


Final Benchmark Analysis Results…

GP Algorithm

	rrBLUP	BRR	B-A	B-B	BL	SVR	SVR_{poly}	SVR_{rbf}	RF	GTB	ANN	ANN_{BA}	EN_5	EN_{11}
Maize	0.44	0.44	0.42	0.43	0.44	0.41	0.43	0.39	0.43	0.37	0.41	0.39	***0.45***	0.44
Rice	0.34	0.39	0.38	0.38	0.39	0.38	0.38	0.38	0.4	0.38	0.25	0.32	0.39	***0.41***
Sorghum	***0.63***	***0.63***	***0.63***	***0.63***	0.62	0.62	***0.63***	***0.63***	0.58	0.58	0.59	0.58	0.62	***0.63***
Soy	0.46	0.46	***0.47***	0.46	0.46	0.43	0.41	0.04	0.36	0.4	0.44		0.46	***0.47***
Spruce	0.32	0.32	0.32	0.32	0.32	0.19	0.33	0.34	***0.35***	0.33	0.28	0.29	0.34	0.34
Switch-grass	***0.61***	***0.61***	***0.61***	***0.61***	***0.61***	0.6	***0.61***	0.6	0.57	0.56	0.53	0.57	0.6	0.6
Ave % decrease	***4%***	***3%***	***3%***	***3%***	***3%***	***8%***	***3%***	***7%***	***7%***	***10%***	***12%***	***13%***	***2%***	***1%***

Height



Final Benchmark Analysis Results…

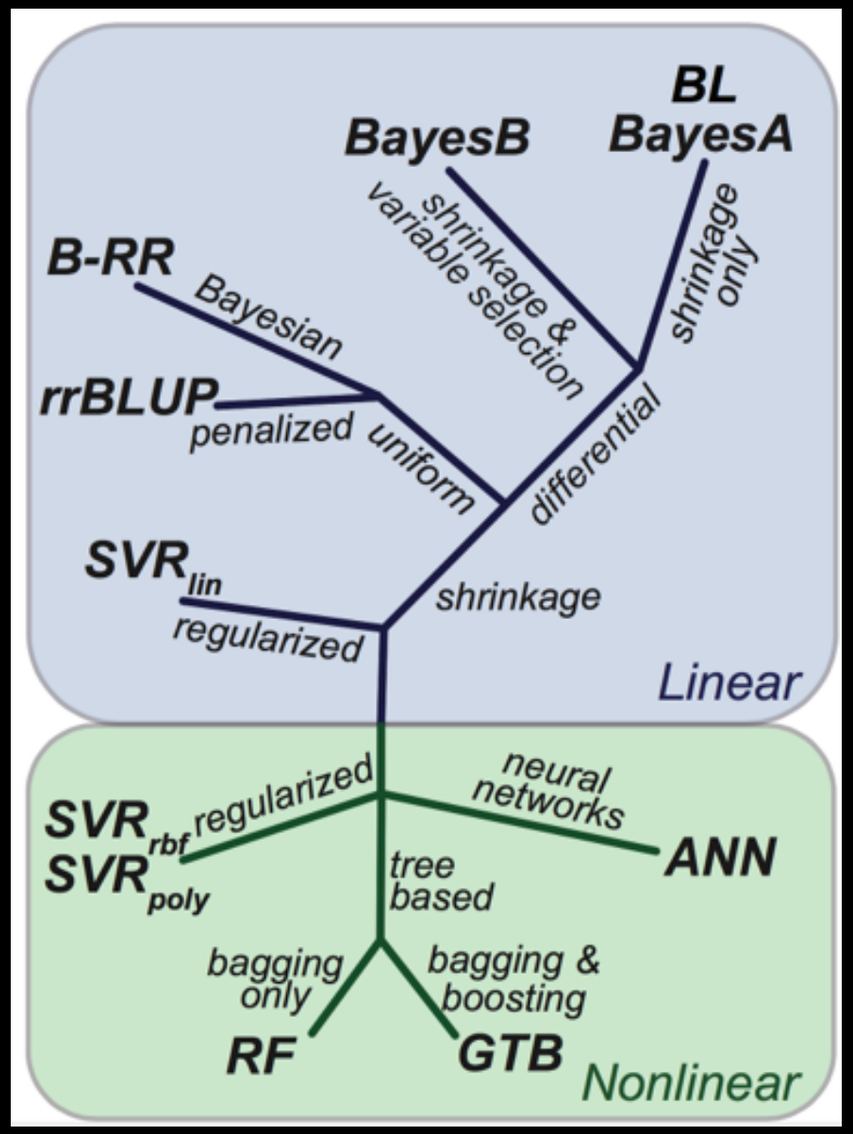
	rrBLUP	BRR	B-A	B-B	BL	SVR	SVR_{poly}	SVR_{rbf}	RF	GTB	ANN	ANN_{BA}	EN_5	EN_{11}	
Maize	0.74	0.77	0.75	0.75	0.74	0.74	0.74	0.66	0.71	0.71	0.67	0.74	0.75	0.75	Developmental Timing
Rice	0.42	0.54	0.64	0.65	0.54	0.5	0.56	0.52	0.54	0.51	0.51	0.57	0.58	0.6	
Soy	0.49	0.49	0.49	0.49	0.49	0.43	0.42	0.29	0.39	0.46	0.46		0.49	0.49	
Switch-grass	0.85	0.85	0.85	0.85	0.84	0.84	0.85	0.85	0.8	0.82	0.66	0.82	0.83	0.84	
	6%	5%	1%	1%	5%	9%	8%	13%	12%	10%	14%	6%	4%	2%	
Maize	0.54	0.54	0.53	0.53	0.54	0.52	0.53	0.5	0.52	0.48	0.51	0.5	0.55	0.54	Yield
Rice	0.27	0.36	0.37	0.37	0.36	0.34	0.33	0.29	0.38	0.37	0.28	0.29	0.37	0.38	
Sorghum	0.39	0.41	0.41	0.4	0.4	0.39	0.41	0.38	0.36	0.4	0.38	0.33	0.41	0.41	
Soy	0.65	0.65	0.65	0.65	0.65	0.61	0.6	0.64	0.6	0.62	0.63		0.65	0.65	
	6%	2%	2%	3%	2%	7%	6%	9%	7%	6%	9%	16%	1%	0%	
Sorghum	0.57	0.57	0.57	0.57	0.57	0.56	0.57	0.57	0.53	0.54	0.51	0.52	0.57	0.57	GM
Spruce	0.24	0.25	0.25	0.25	0.25	0.12	0.27	0.27	0.27	0.25	0.24	0.21	0.26	0.26	DBH
Spruce	0.39	0.38	0.39	0.39	0.39	0.32	0.4	0.39	0.36	0.36	0.31	0.36	0.39	0.4	DE
Switch-grass	0.69	0.69	0.69	0.69	0.69	0.68	0.69	0.69	0.65	0.66	0.55	0.66	0.68	0.69	ST
Ave % decrease	4%	3%	3%	3%	3%	11%	1%	0%	5%	7%	16%	11%	2%	1%	

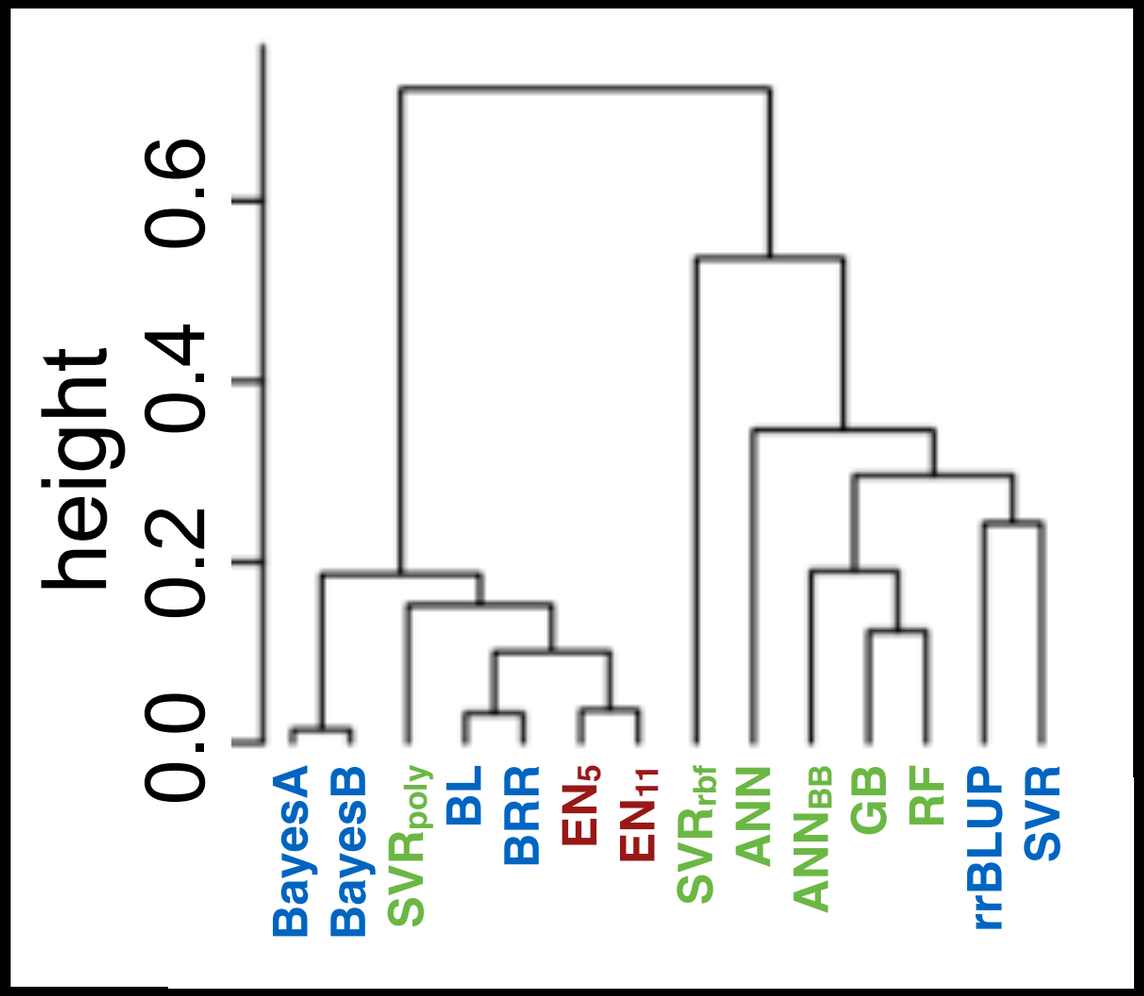
% Decrease r

species best

0 10 20

Hierarchical clustering of GP algorithms





Why do linear algorithms consistently perform well?

- H1: Complex biological systems generate signals that are consistent with linear, additive, genetic models (Hill, Goddard, & Visscher 2008).
- H2: The amount of training data available for most GP problems is insufficient for learning nonlinear interactions when p>>n.

ML and Deep Learning Pipelines

azodichr / **ML-Pipeline**

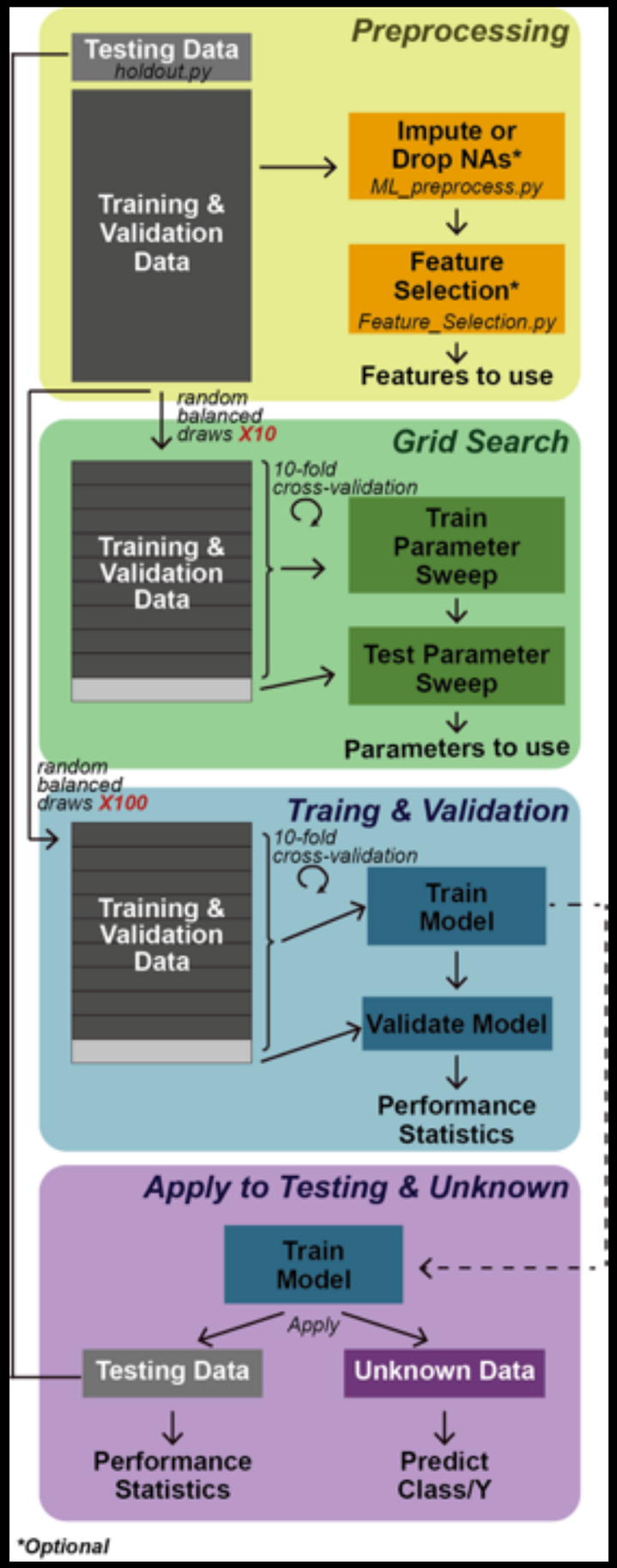
<> Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki

Shiu Lab code base for Machine Learning implemented in SciKit-Learn

azodichr / **ANN_Pipeline**

<> Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki

Shiu Lab code base for Artificial Neural Networks implemented in Tensorflow

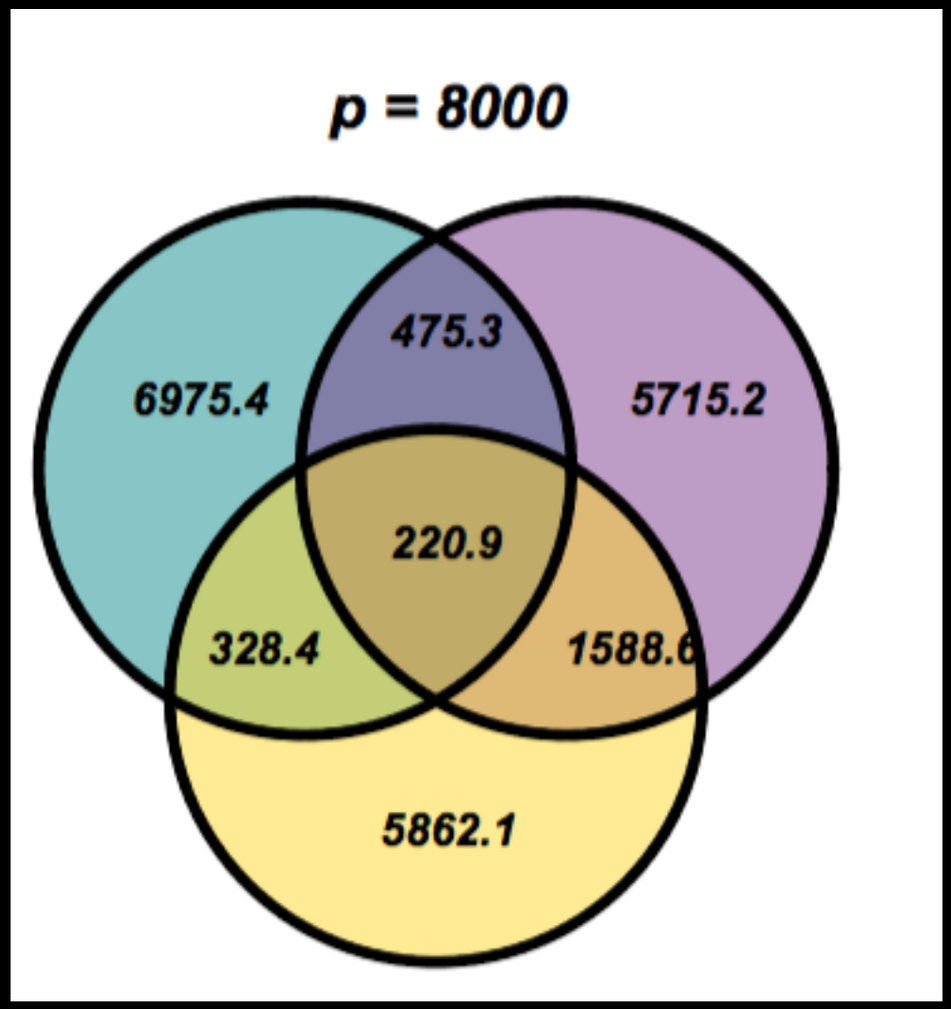


azodichr/ML-Pipeline/README.md

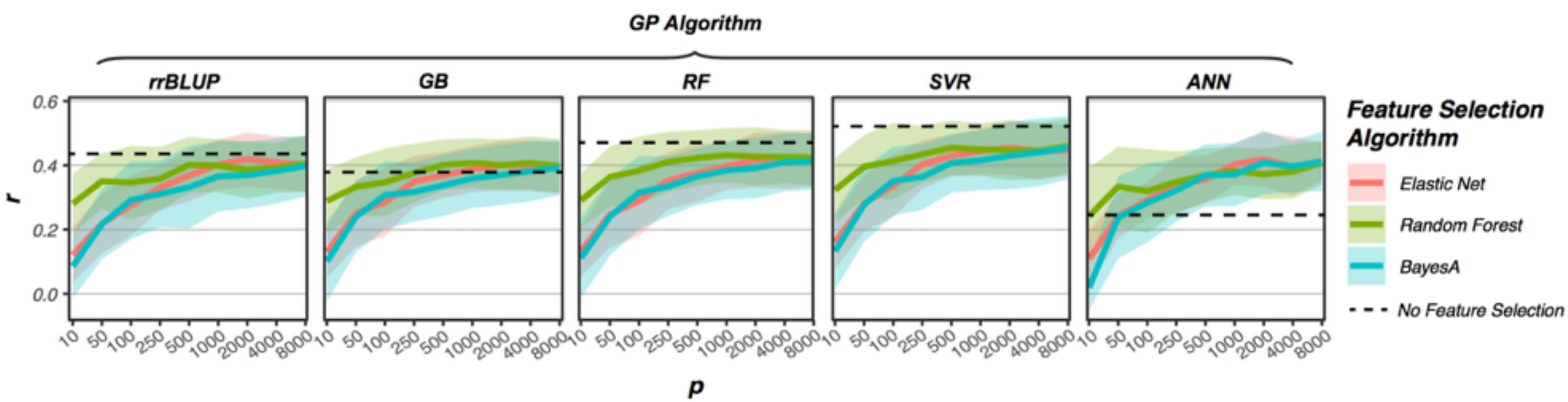
Future Directions

- Wrap up and submit GP algorithm comparison manuscript.
- Use gene expression levels as input into GP algorithms.
  - Are transcriptomes more predictive than genotype data?
  - Can we better predict traits by combining genotype & transcriptome data?
  - Is transcript data useful when dissecting GP models to learn about the genetic basis of a trait?
- Use “deep Connection Weight” (dCW) approach to interpret deep learning models to identify markers and interactions between markers important for a trait

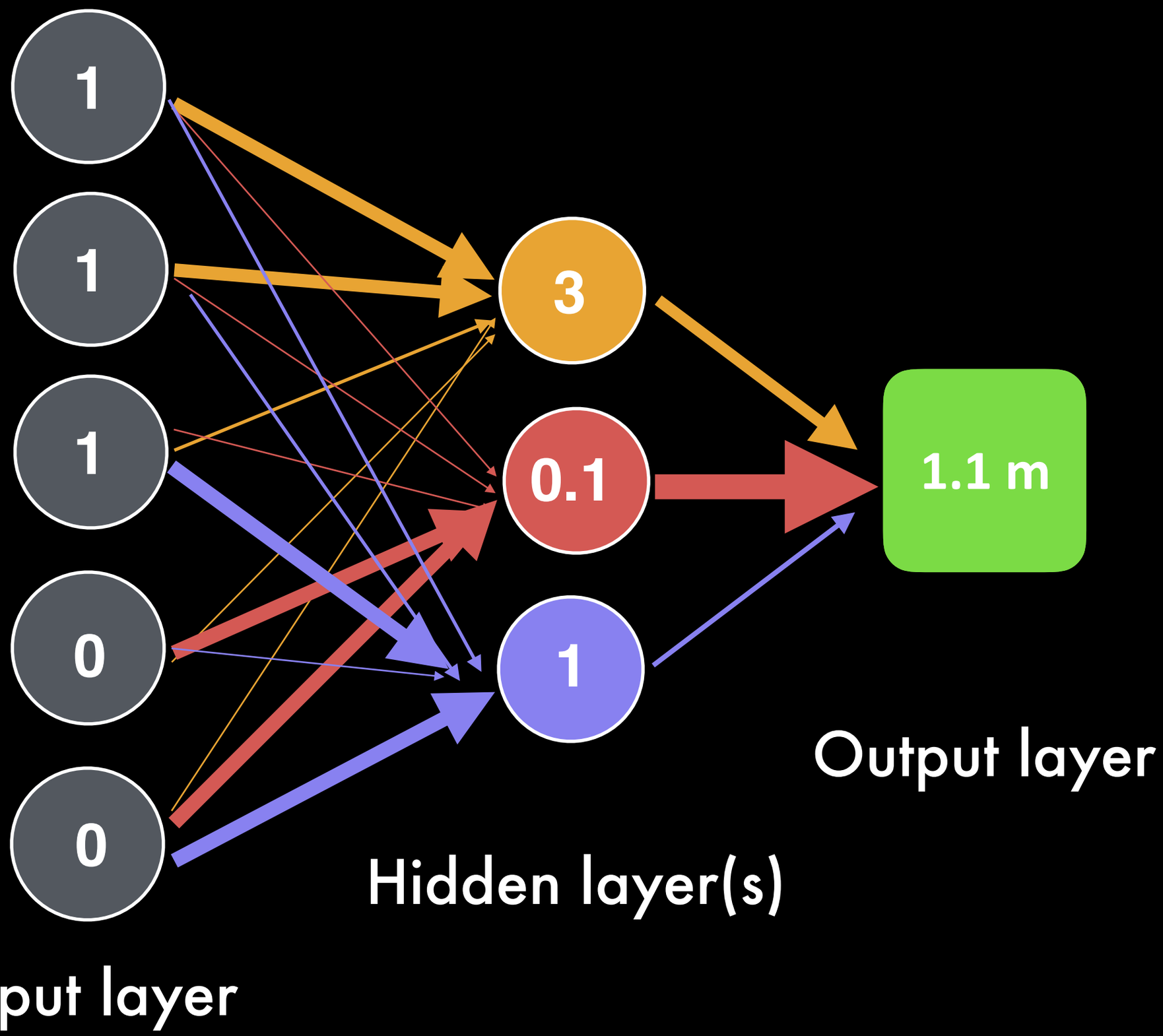
Thanks!



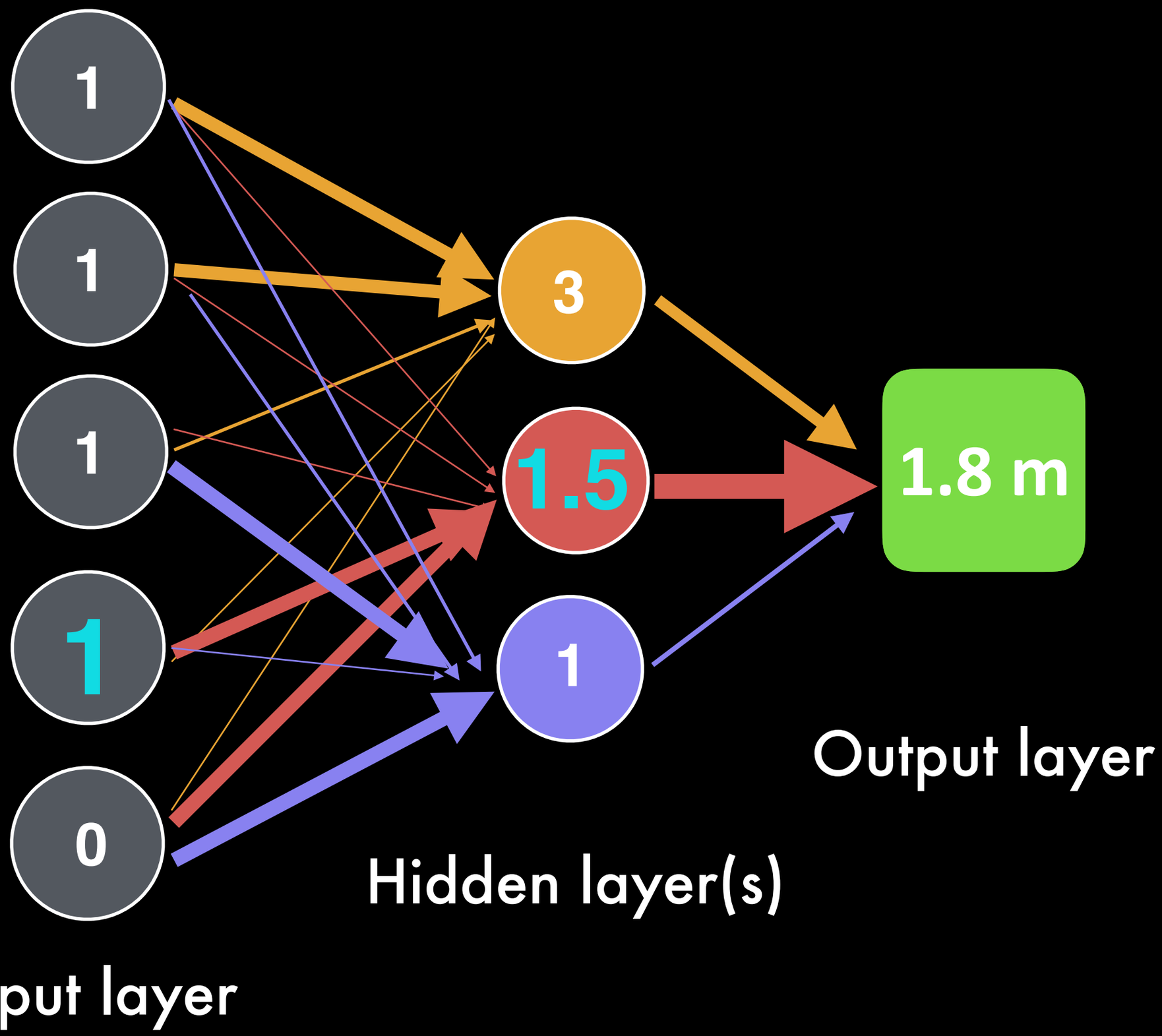
Random expectation for triple overlap: ~10



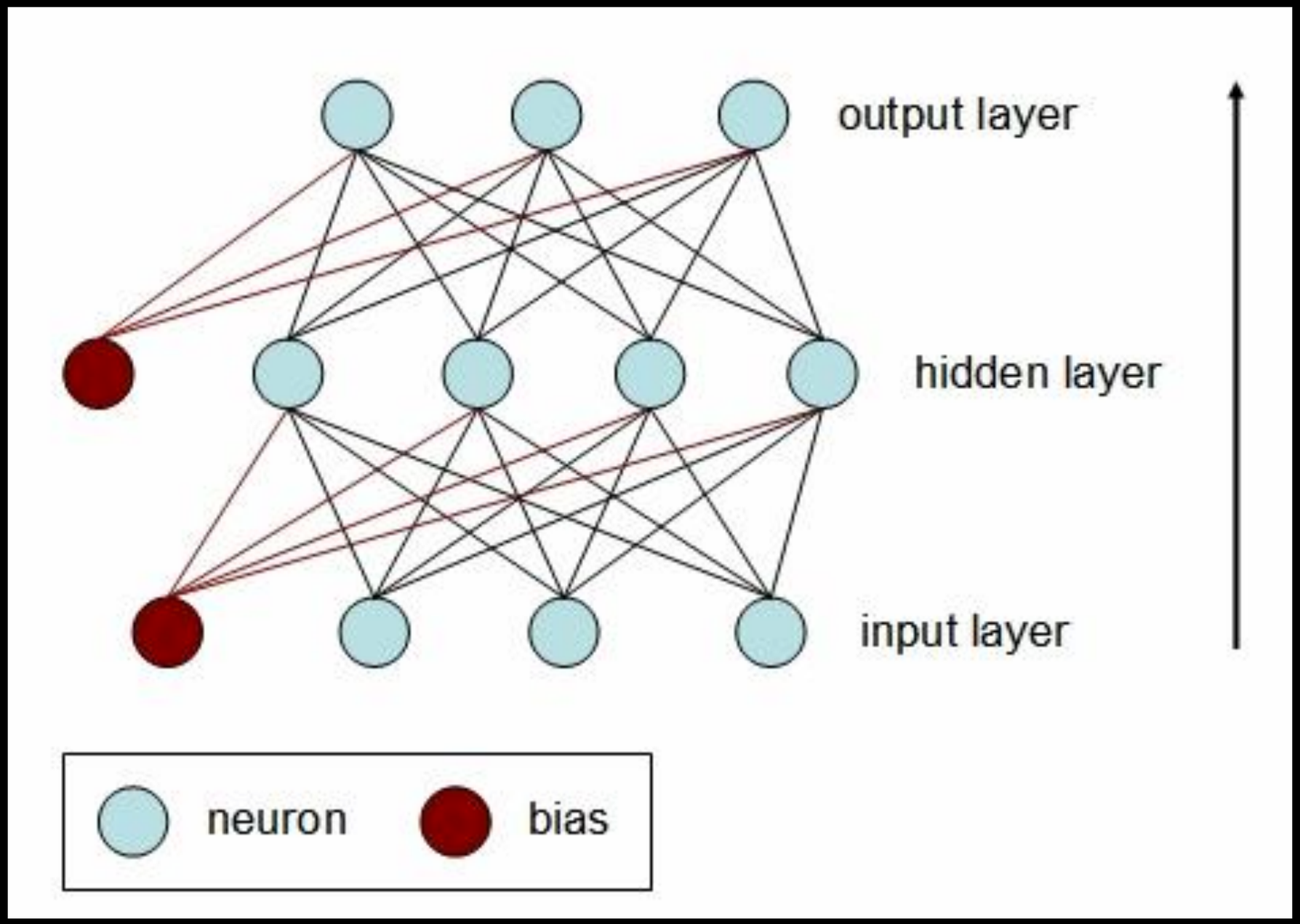
Genotype X



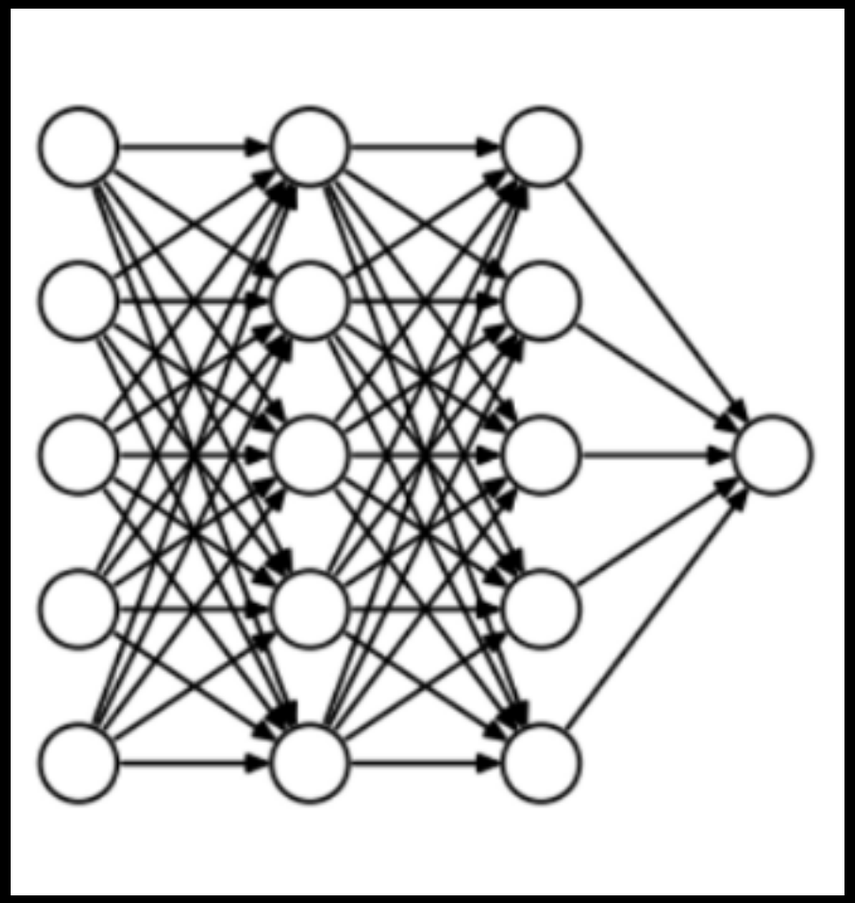
Genotype Y

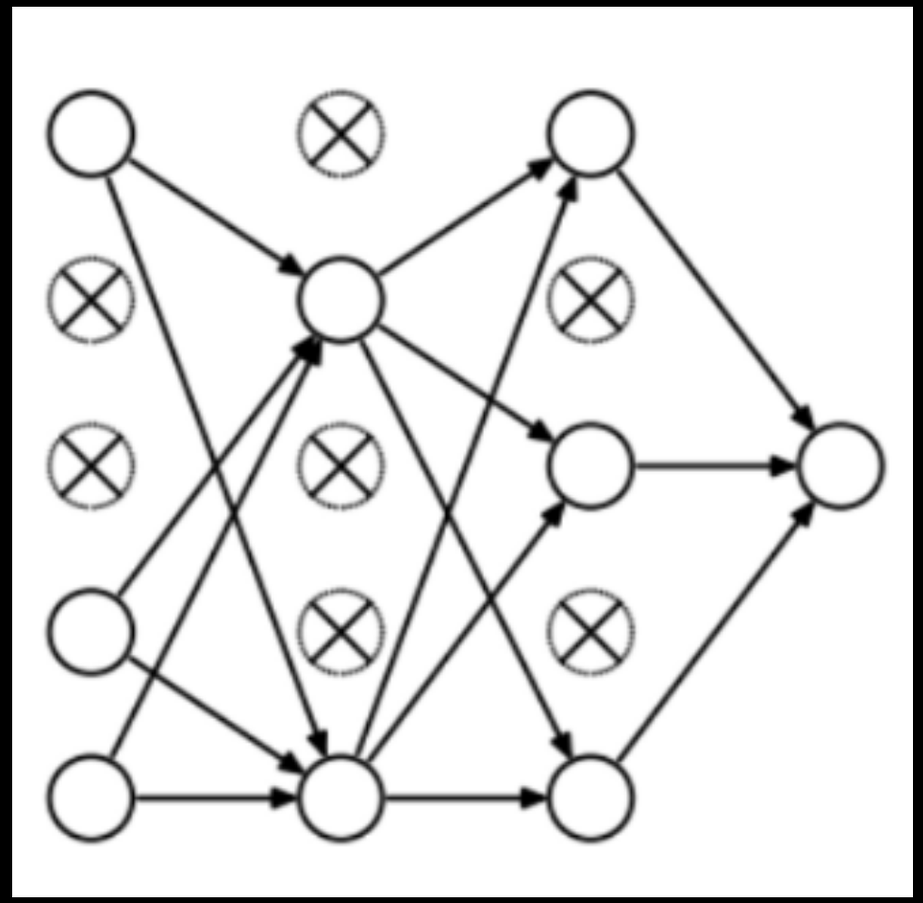


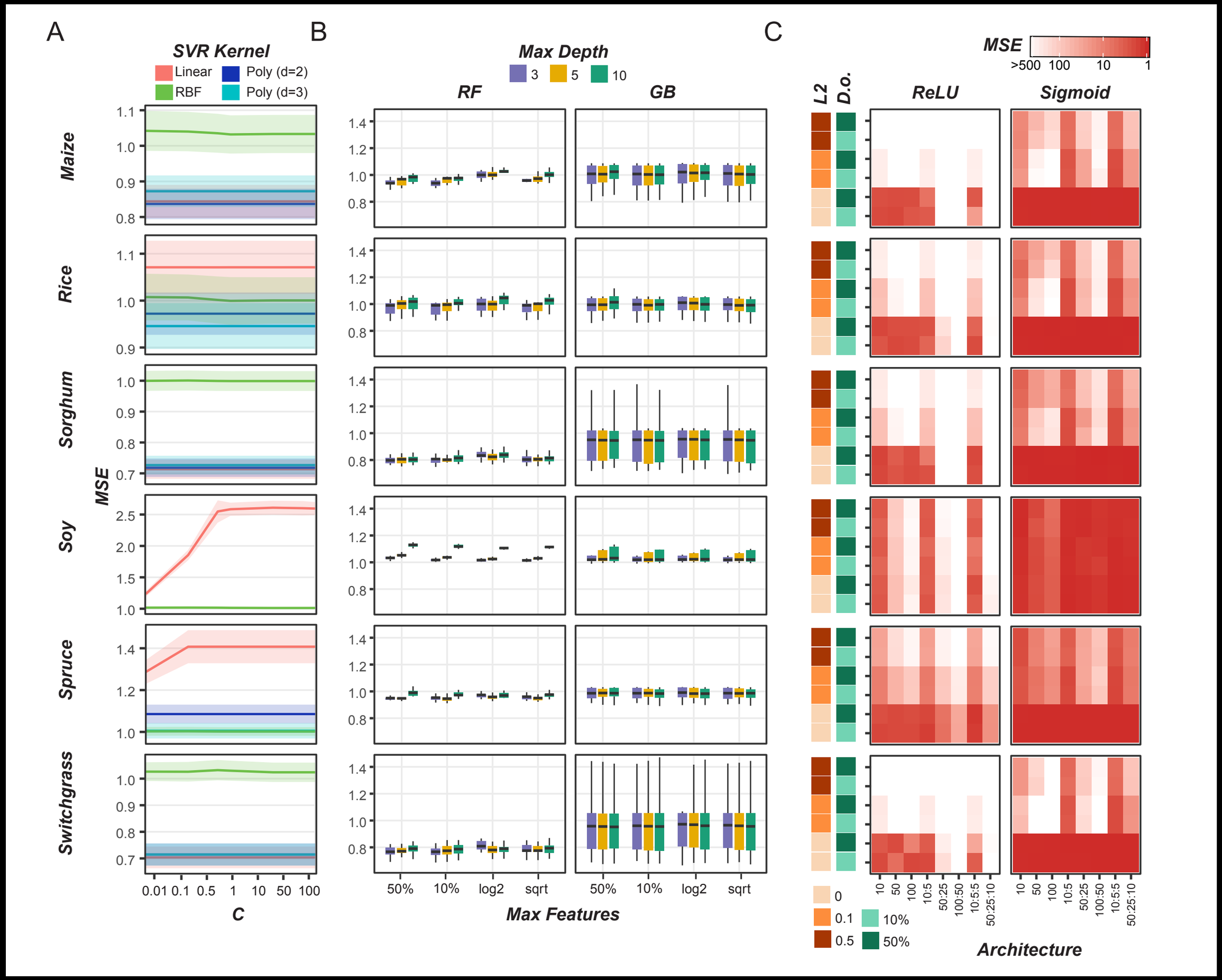
Role of the bias term....

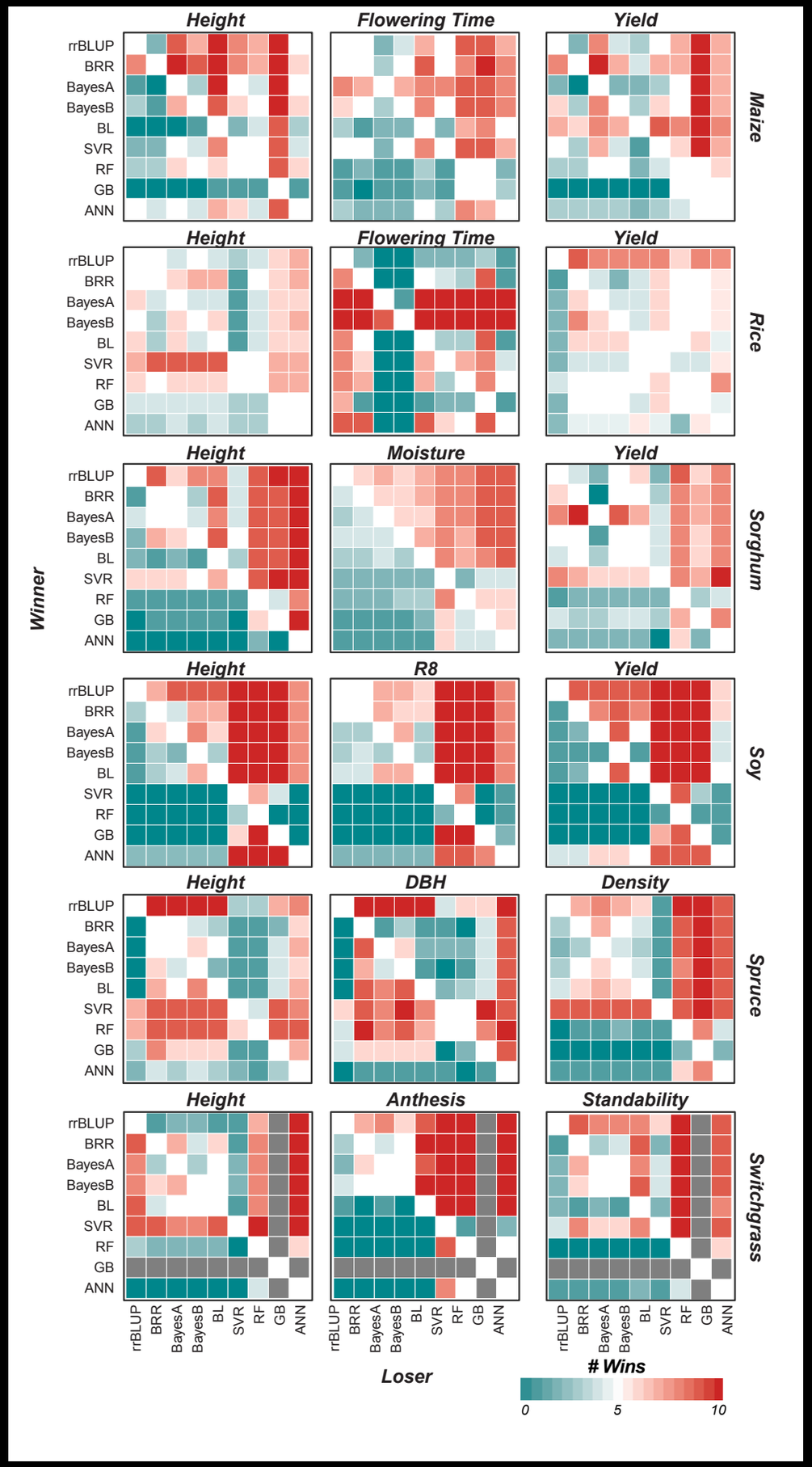


Drop out Explanation






A
B
C
SVR Kernel
Linear
Poly (d=2)
RBF
Poly (d=3)
Max Depth
3
5
10
RF
GB
MSE
>500
100
10
1
L2
D.o.
ReLU
Sigmoid
Maize
Rice
Sorghum
Soy
Spruce
Switchgrass
MSE
C
0.01 0.1 0.5 1 10 50 100
50% 10% log2 sqrt
Max Features
0
0.1
0.5
10%
50%
10 50 100 10:5 50:25 100:50 10:5:5 50:25:10
Architecture


Height
Flowering Time
Yield
Maize
Height
Flowering Time
Yield
Rice
Height
Moisture
Yield
Sorghum
Height
R8
Yield
Soy
Height
DBH
Density
Spruce
Height
Anthesis
Standability
Switchgrass
Winner
rrBLUP
BRR
BayesA
BayesB
BL
SVR
RF
GB
ANN
Loser
Wins
0
5
10

1st Hidden Layer

	Node A1	Node A2	…	Node An
SNP A	-0.66	-0.8	…	-0.21
SNP B	-0.44	0.66	…	0.64
SNP C	0.18	0.66	…	0.84
SNP D	0.45	0.78	…	0.52

2nd Hidden Layer

	Node B1	Node B2	…	Node Bn
Node A1	-0.44	0.18	…	-0.38
Node A2	0.15	0.15	…	0.24
…	…	…	…	…
Node An	1.59	0.65	…	0.37

Output Layer

	Output Layer
Node B1	0.00001
Node B2	1.44
…	…
Node Bn	2.14