Leveraging Machine Learning and Environmental Data to Enhance Genomic Prediction in Canola

S. LIBAN[1,3], L. EHRIG[2], L. GABUR[2], E. GILLIS[3], S.FOX[3], L.RACH[3], J.DUGUID[3], J.MUTCHESON[3], M.DOMARATZKI[4], R.SNOWDON[2], R.DUNCAN[1]

SEPT 27, 2023

1. UNIVERSITY OF MANITOBA, 50 SIFTON RD, WINNIPEG MB, CANADA
2. UNIVERSITY OF GIESSEN, LUDWIG STREET 23, GIESSEN HESSE, GERMANY
3. DL SEEDS INC., 25028 RD. 17N STANLEY, MB, CANADA
4. WESTERN UNIVERSITY, 1151 RICHMOND STREET, LONDON ON, CANADA

Canola: A Canadian Innovation

- Canola is the #1 revenue crop in Canada
  - 12 billion in 2021 to farmers
- 30 Billion to Canadian economy

Canola Council of Canada, 2019

Breeding Canola: Phenotype

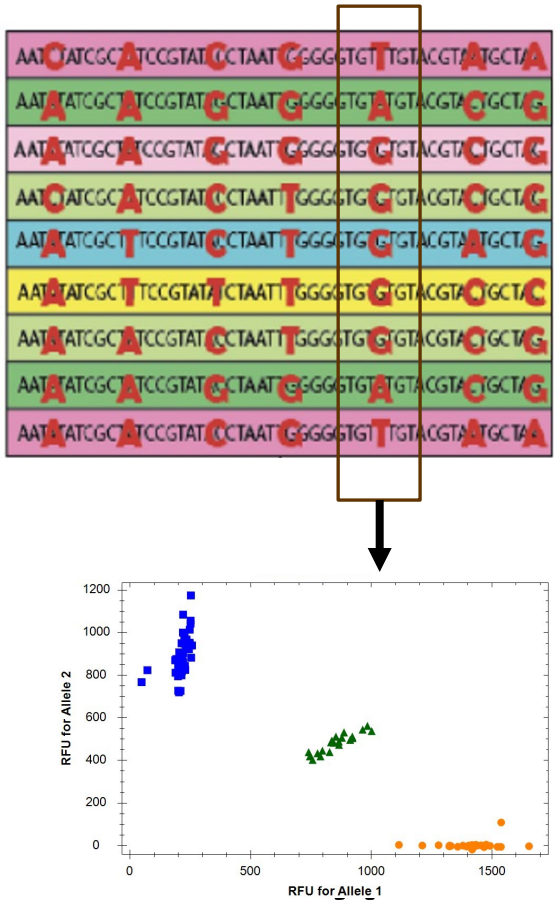
Phenotype

Breeding Canola: Genotype

Phenotype

Marker Assisted Selection

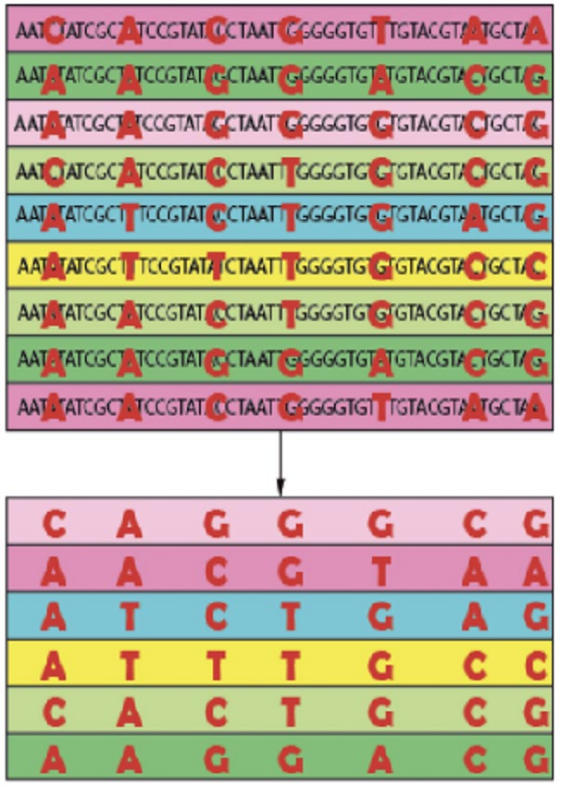
Trait linked SNP



Breeding Canola: Genotype

Phenotype

Genomic Selection Models



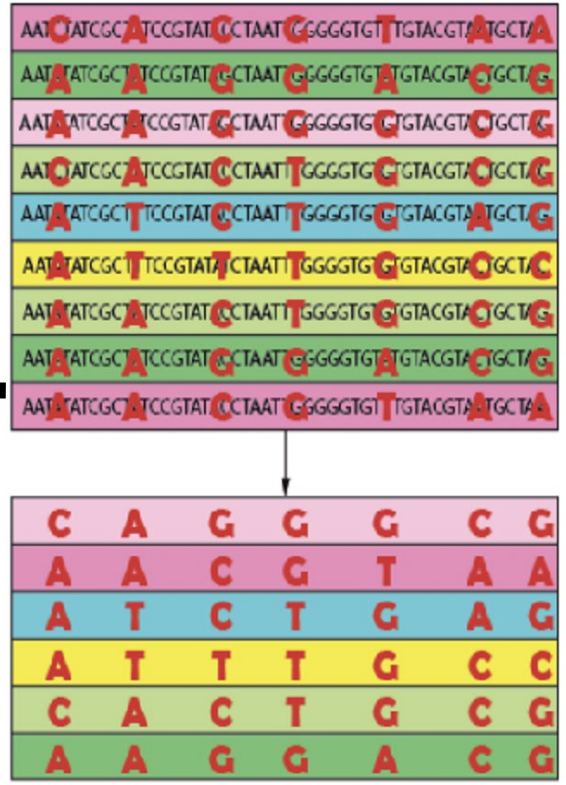
Genotype

Breeding Canola: Environment

Environment

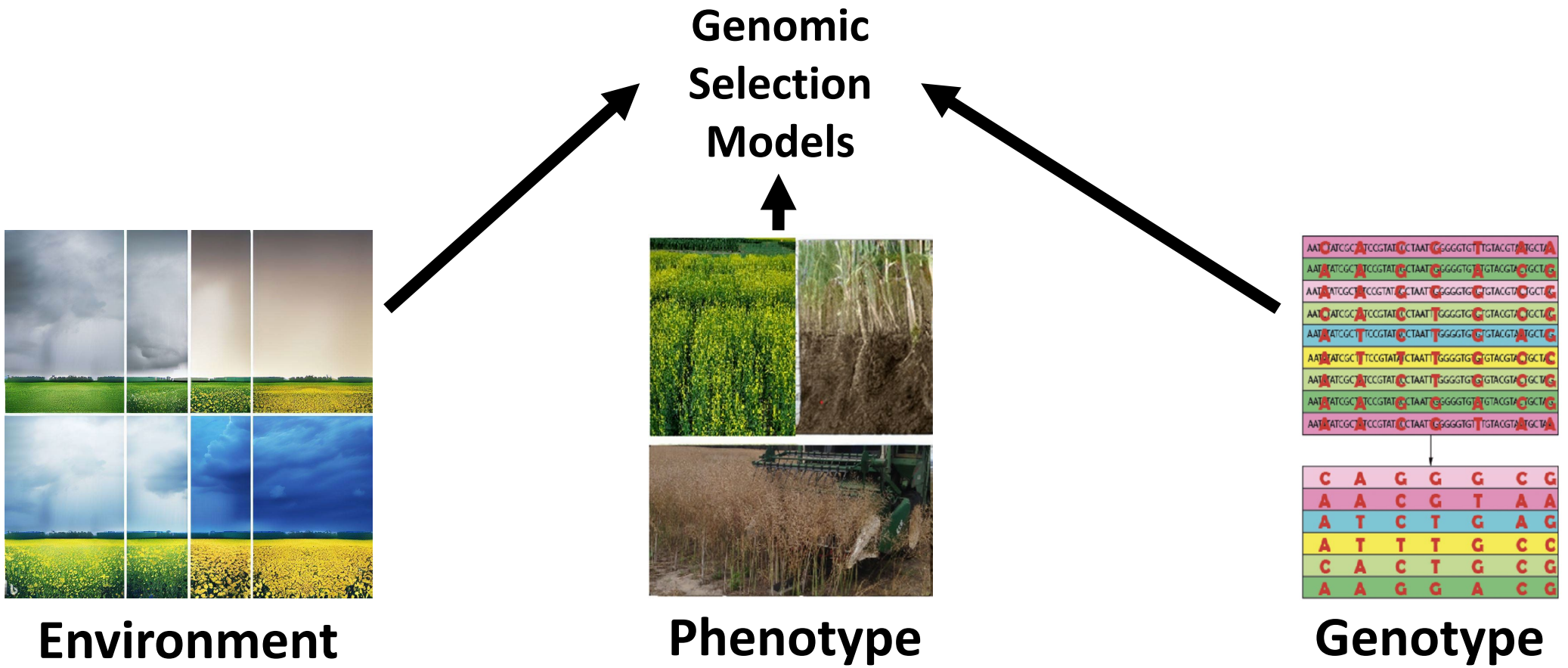
Phenotype

Genomic Selection Models



Genotype

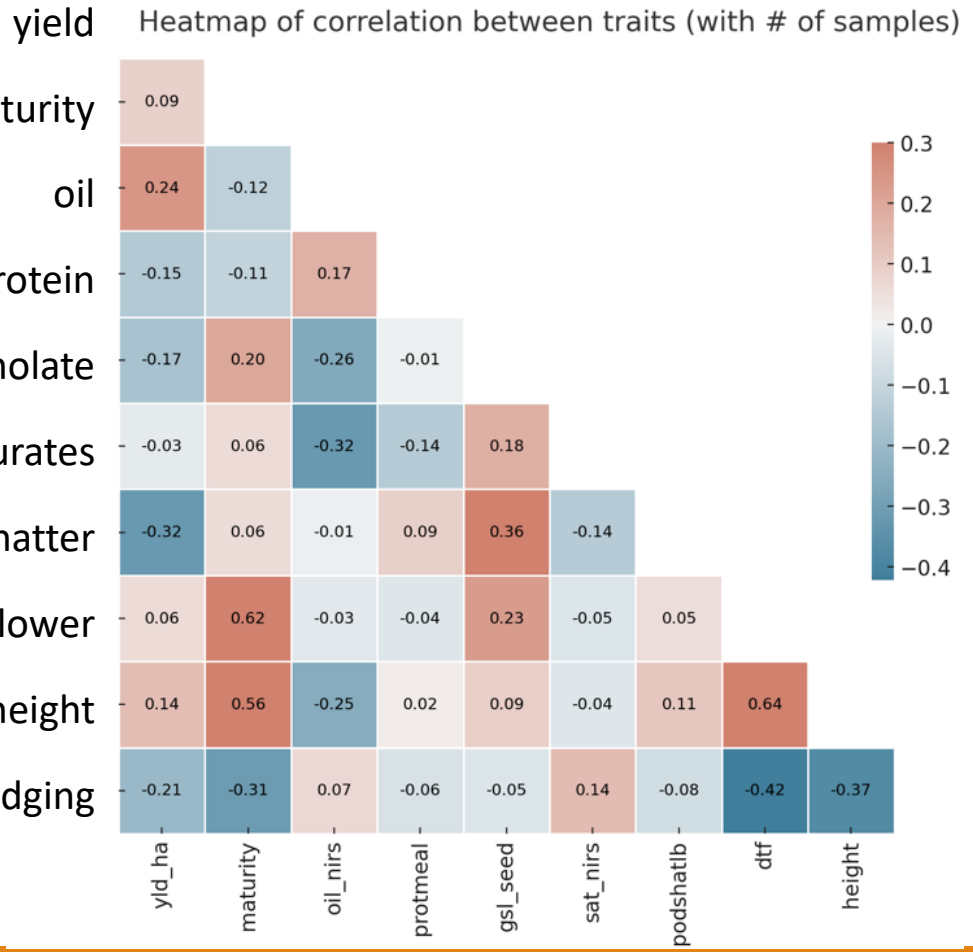
Breeding Canola: Adding the Environment



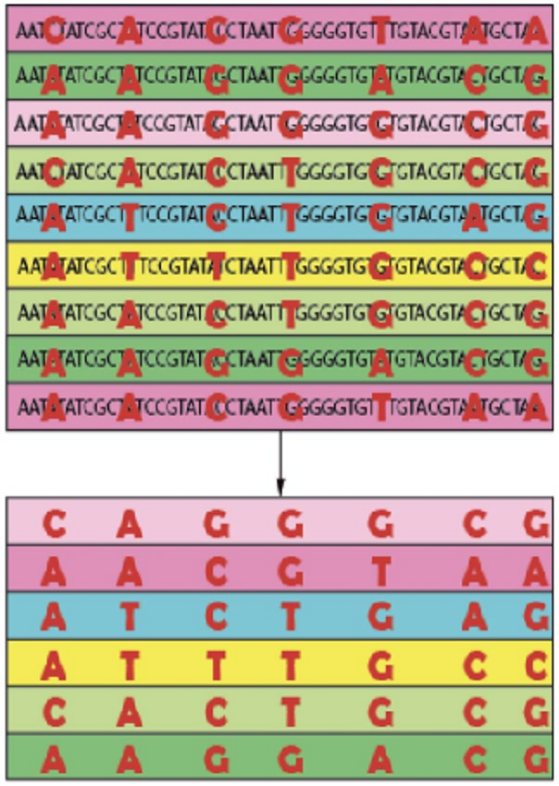
Phenotypic Data

2016-2020

- 5146 yield
- 4938 maturity
- 4957 oil
- 4957 protein
- 4958 glucosinolate
- 770 saturates
- 870 pod shatter
- 2112 days to flower
- 801 height
- 444 lodging



Genotypic Data



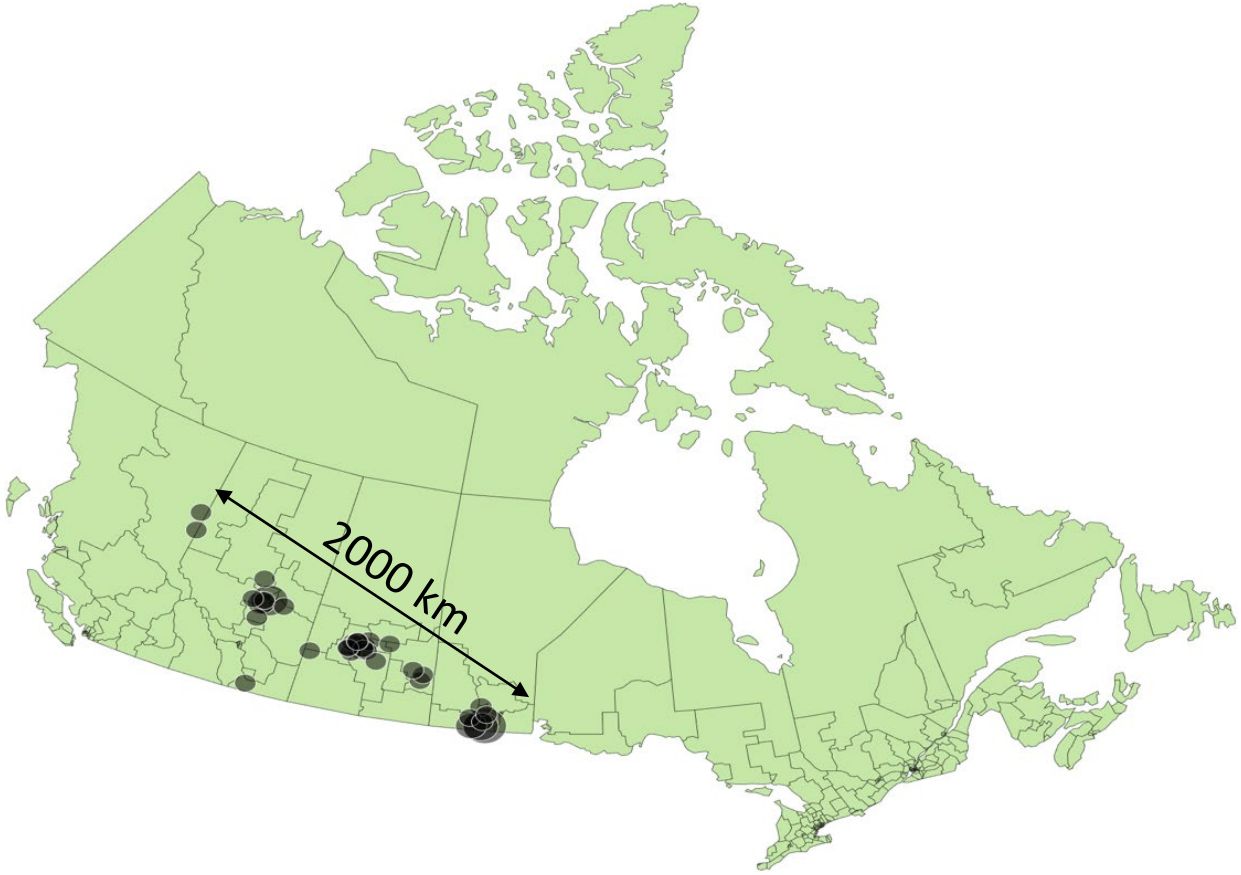
Genotype

- 19K Array (subset of Illumina 60k array)
  - 1500 Parents genotyped
    - 5000 F1 Hybrids generated

P1	Hybrid	P1
0	0.5	1
2	1	0
0	0	0
1	1.5	2
2	2	2
2	1	0
0	0	0
1	1	1

0	0.5	1	1.5	2

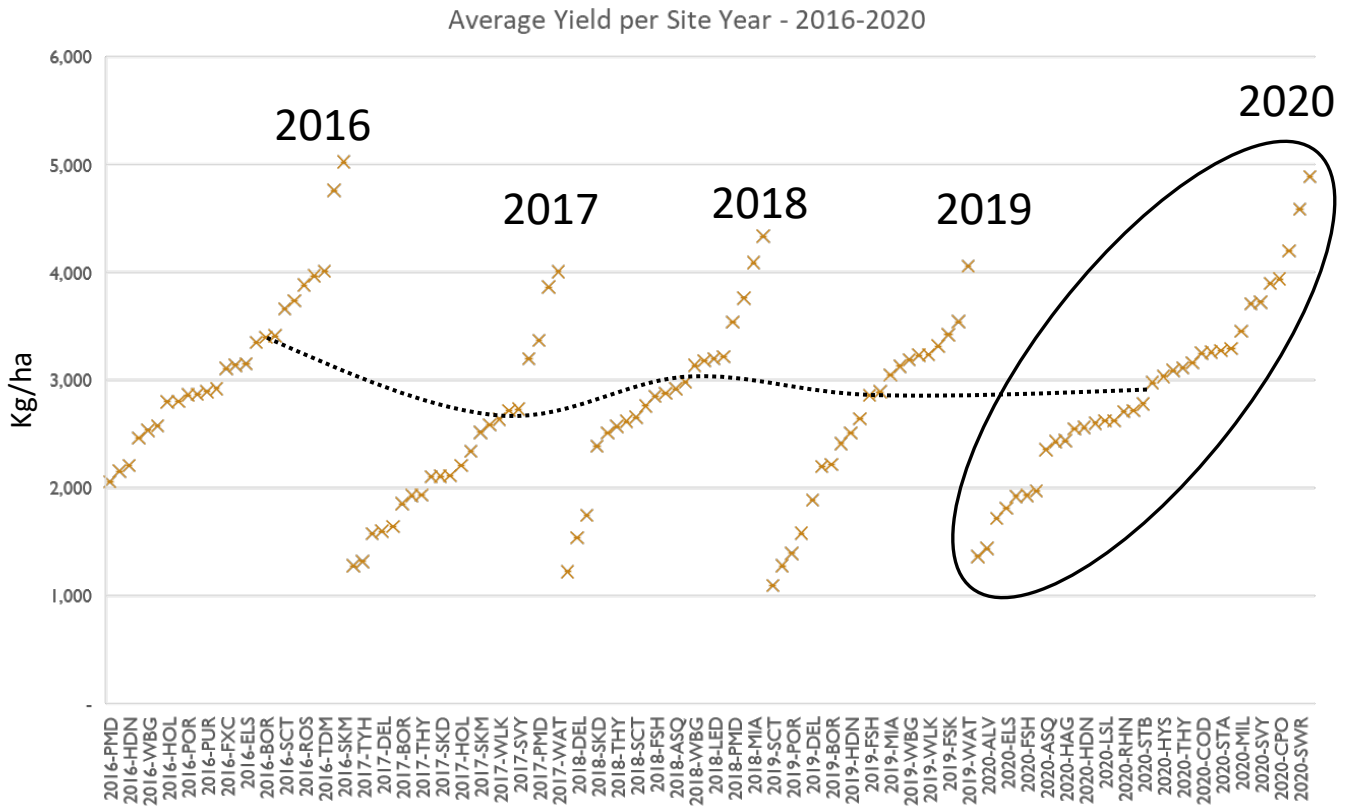
Environmental Data



Environment

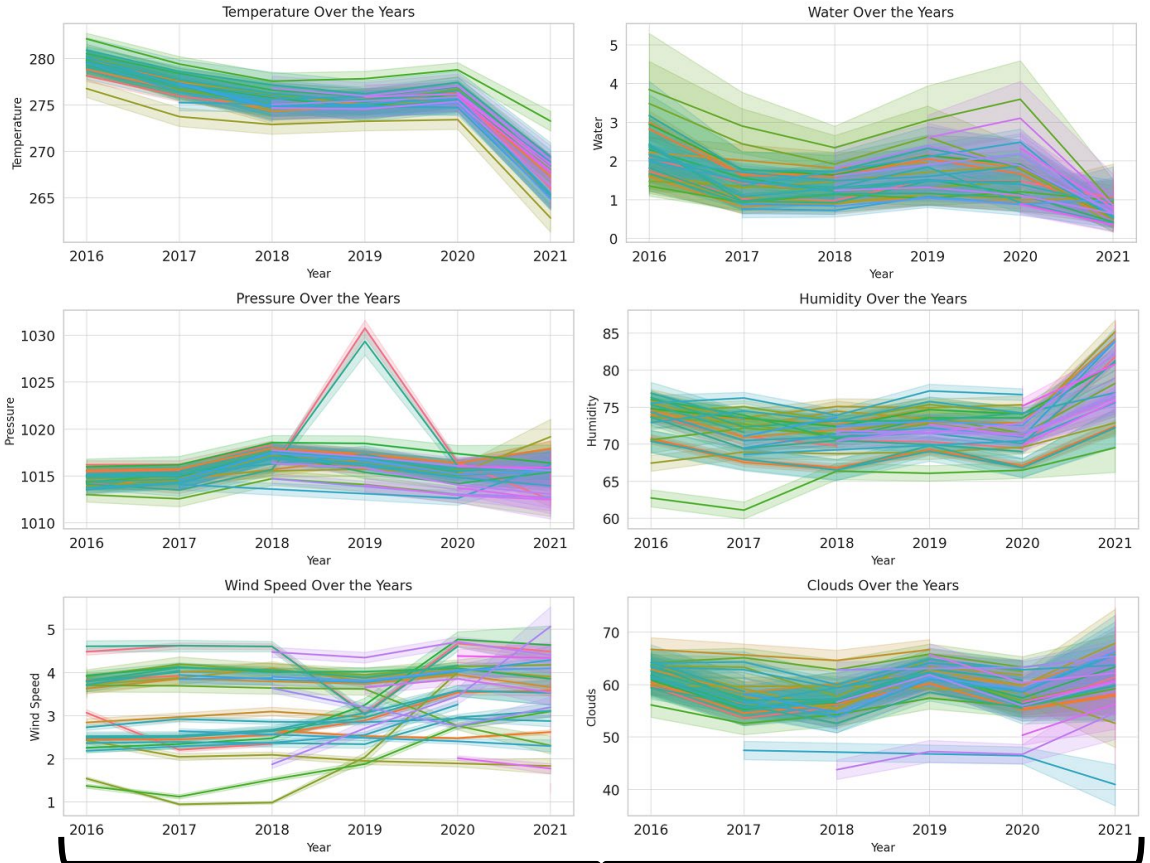
Breeding Canola: The Data

Environment



Breeding Canola: The Data

Environment

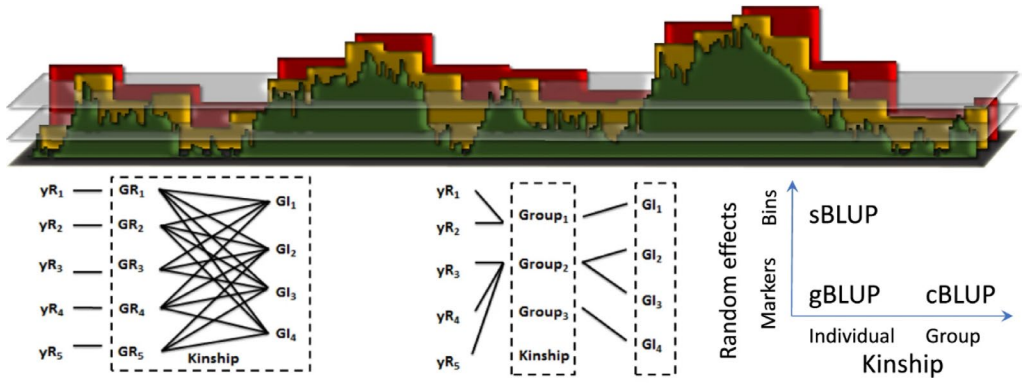


792 Environmental Variables

Genomic Selection: The Functions

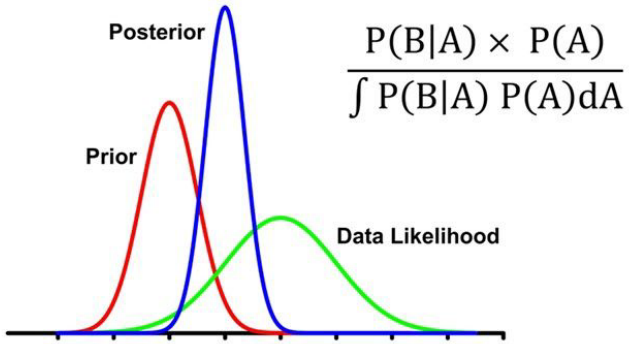
Genomic BLUP (Best Linear Unbiased Predictor)

$y=\mu+Zu+e$, $u\sim N(0,G)$, $e\sim N(0,R)$



Wang et al., 2018

Bayesian Functions



$$\frac{P(B|A) \times P(A)}{\int P(B|A)\, P(A)dA}$$

Model (prior density)	Hyper-parameters
Flat (FIXED)	Mean (μ_β) Variance (σ^2_β)
Gaussian (BRR)	Mean (μ_β) Variance (σ^2_β)
Scaled-t (BayesA)	Degrees of freedom (df_β) Scale (S_β)
Double-Exponential (BL)	λ^2
Gaussian Mixture (BayesB)	π (prop. of non-null effects) df_β S_β
Scaled-t Mixture (BayesC)	π (prop. of non-null effects) df_β S_β

Compas & Rodriguez., 2022

Genomic Selection: Linear Approach

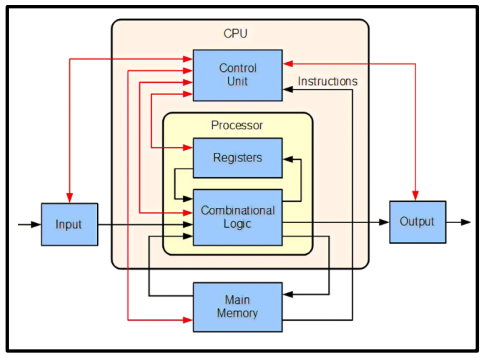
$y=X\beta+\epsilon, \epsilon \sim N(0,\sigma 2I), \beta \sim N(0,\lambda 2I)$

$y=\mu+Zu+e,\ u \sim N(0,G),\ e \sim N(0,R)$

Mathematical Functions

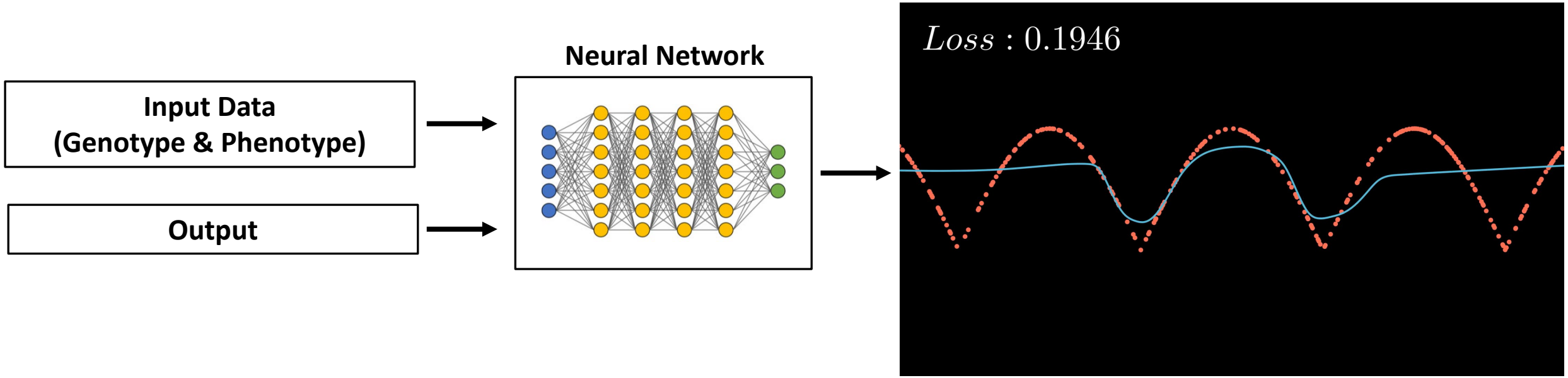
Input Data
(Genotype & Phenotype)

CPU



Output

Genomic Selection: Artificial Intelligence



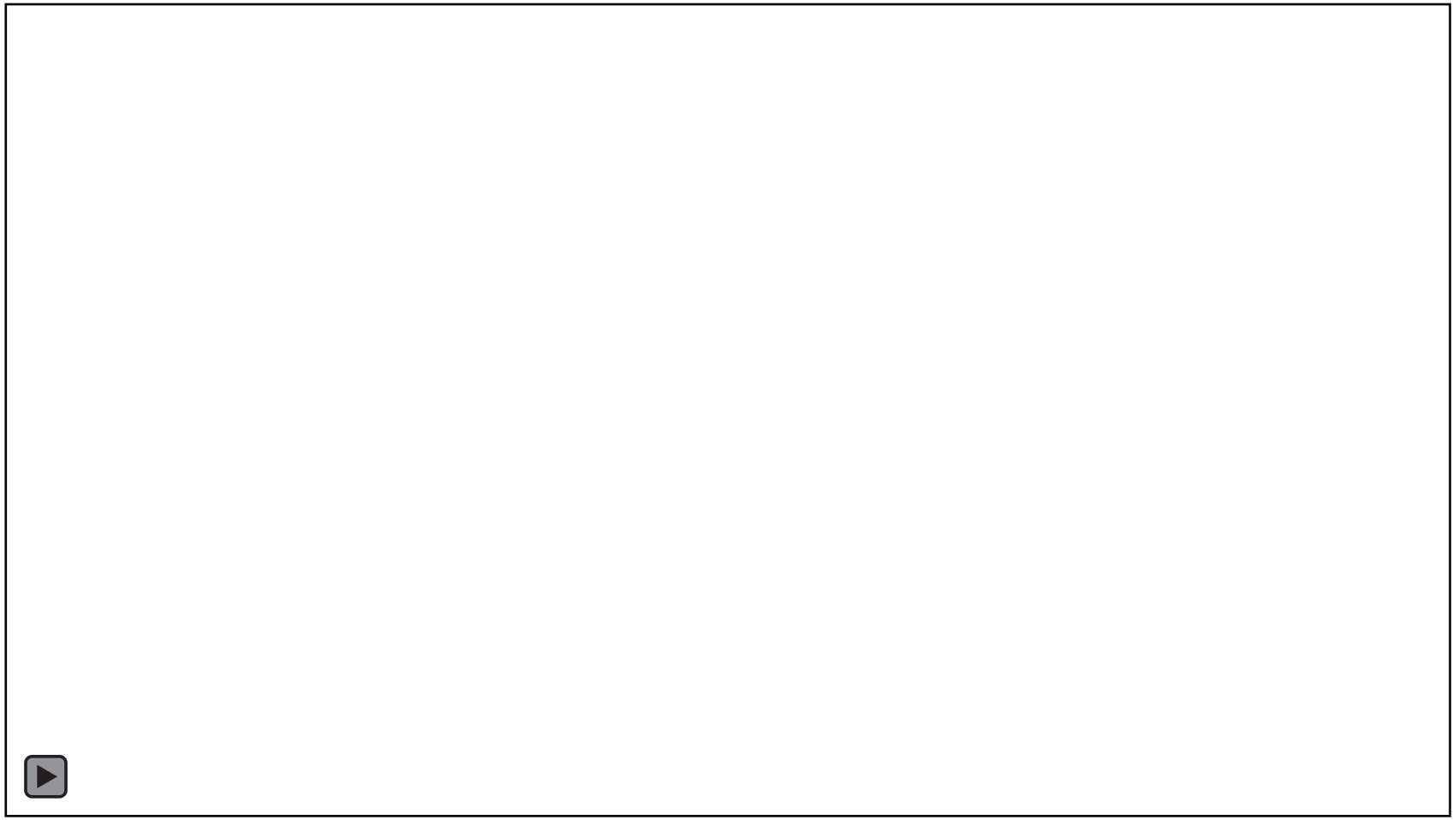
Neural Networks are Universal Function Approximators !

Genomic Selection: AI Optimization

Neural Network Factors

- **Data Size and Quality**
  - Phenotype and Genotype
- **Compute**
  - Optimization and training
- **Architecture**
  - 14 Hyper-parameters

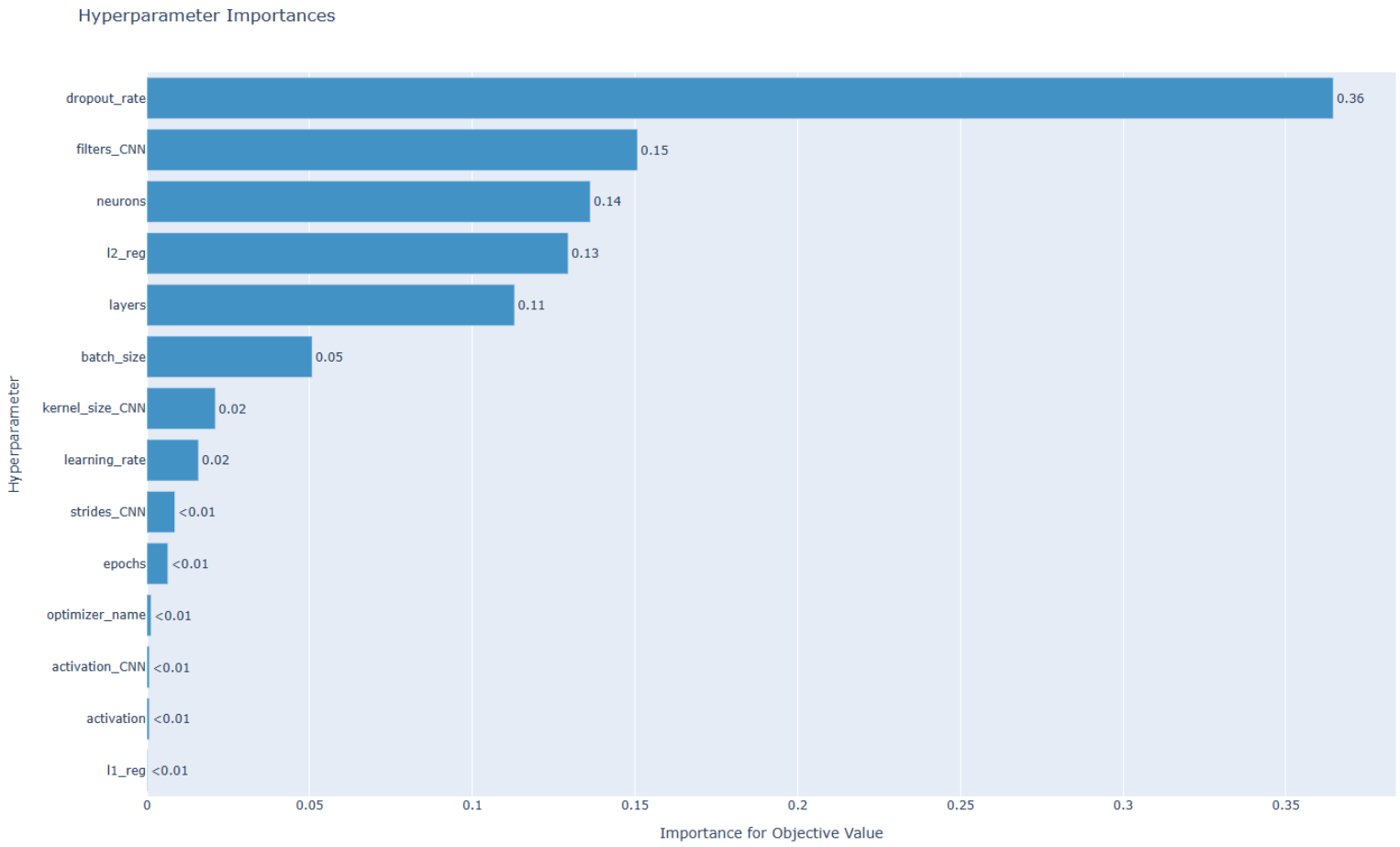
- Parallel cloud instances
  - 1000 compute hours
    - 4090 24GB
    - A100 80GB



Genomic Selection: AI

Neural Network Optimization Outcomes

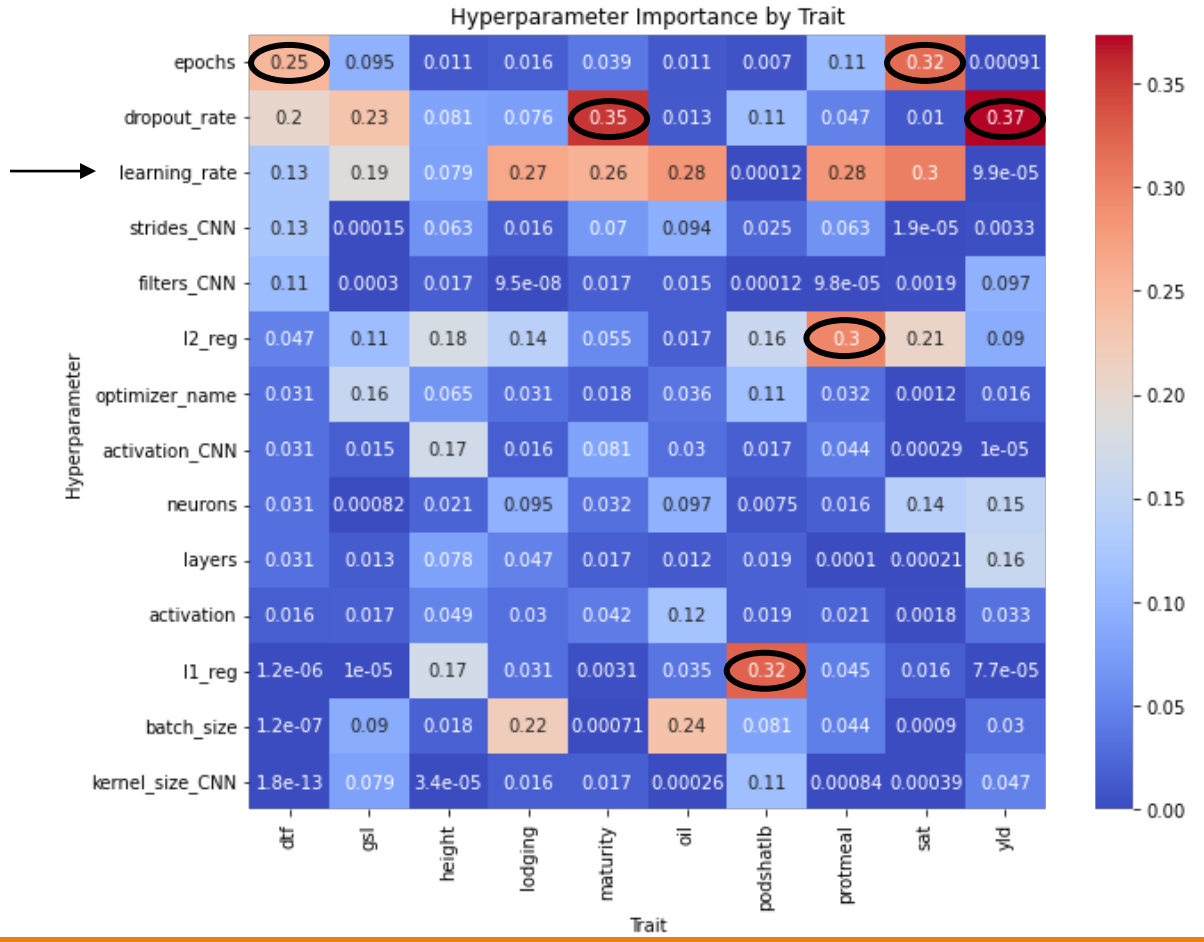
- Hyperparameters have different optimal values
  - trait x parameter combination
- Relative importance varies by trait



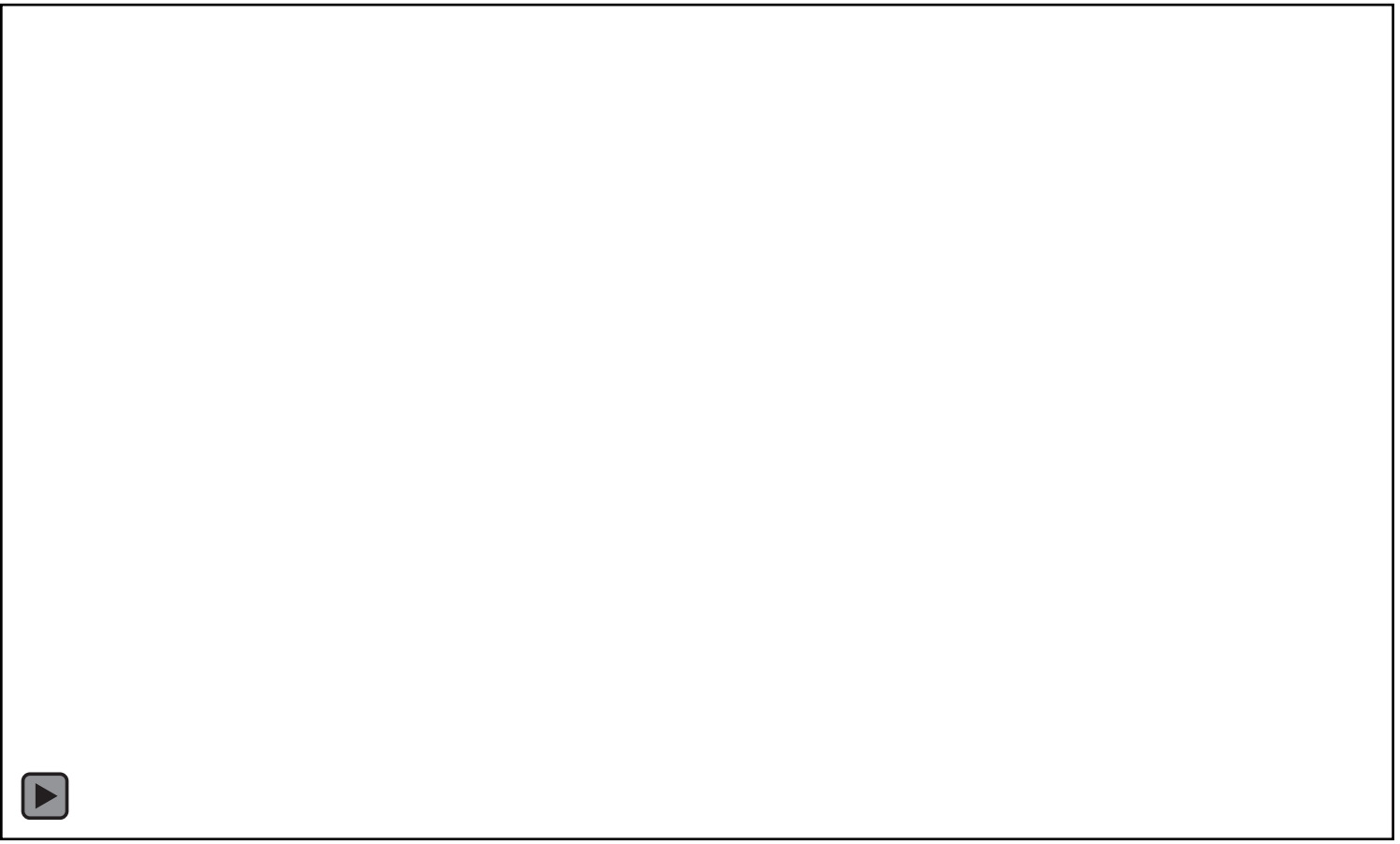
Genomic Selection: AI

Neural Network Optimization Outcomes

- Hyperparameters have different optimal values
  - trait x parameter combination
- Relative importance varies by trait

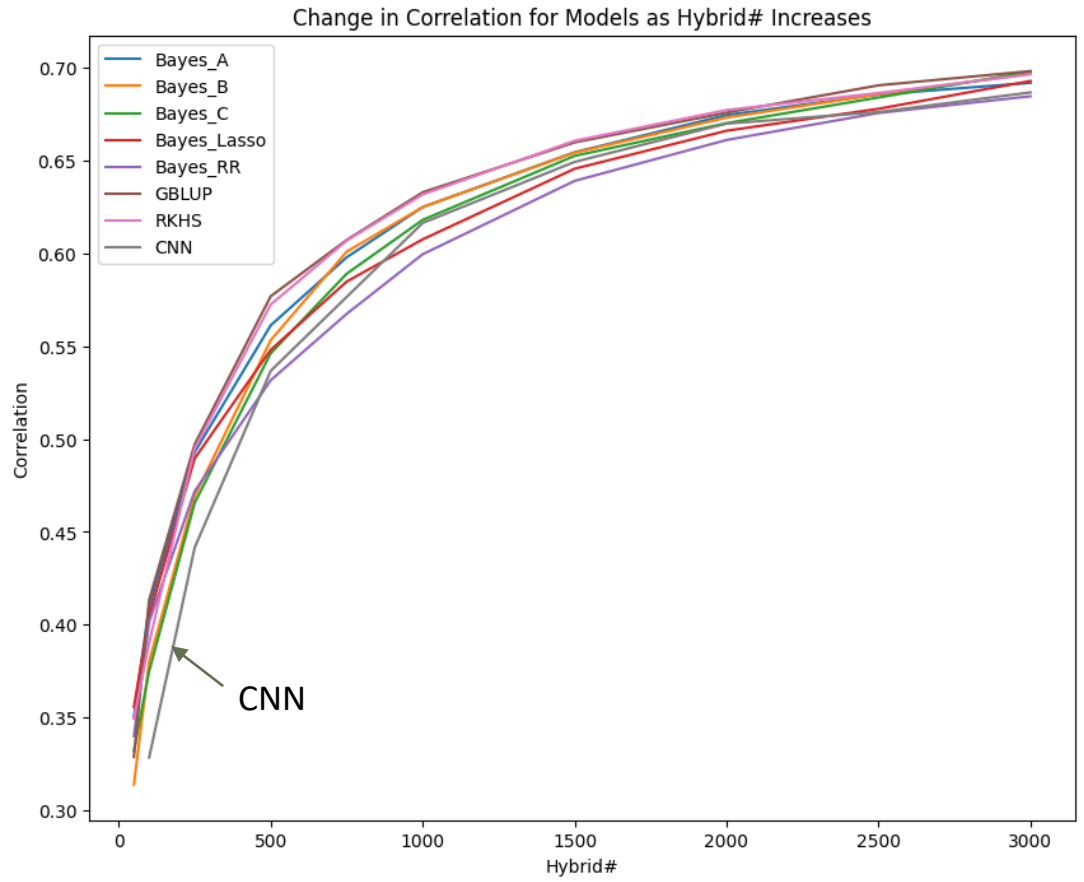


Genomic Selection:
Data Quantity vs Trait (Avg All models)



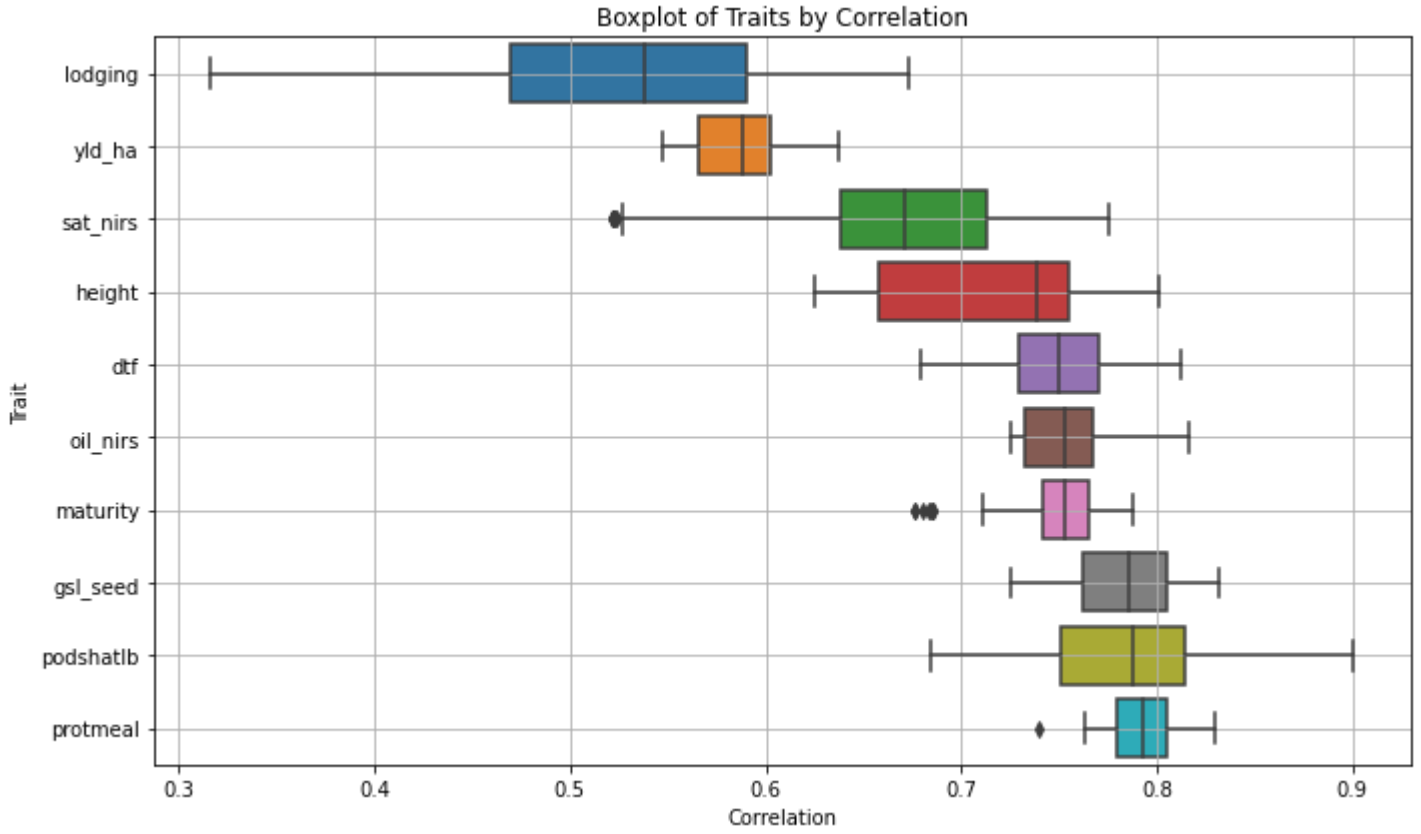
- Models scale positively with sample size
- Diminishing returns with a logarithmic curve
- Approaching limits of model improvement by data quantity!

Genomic Selection: Data Quantity vs Model (All traits)



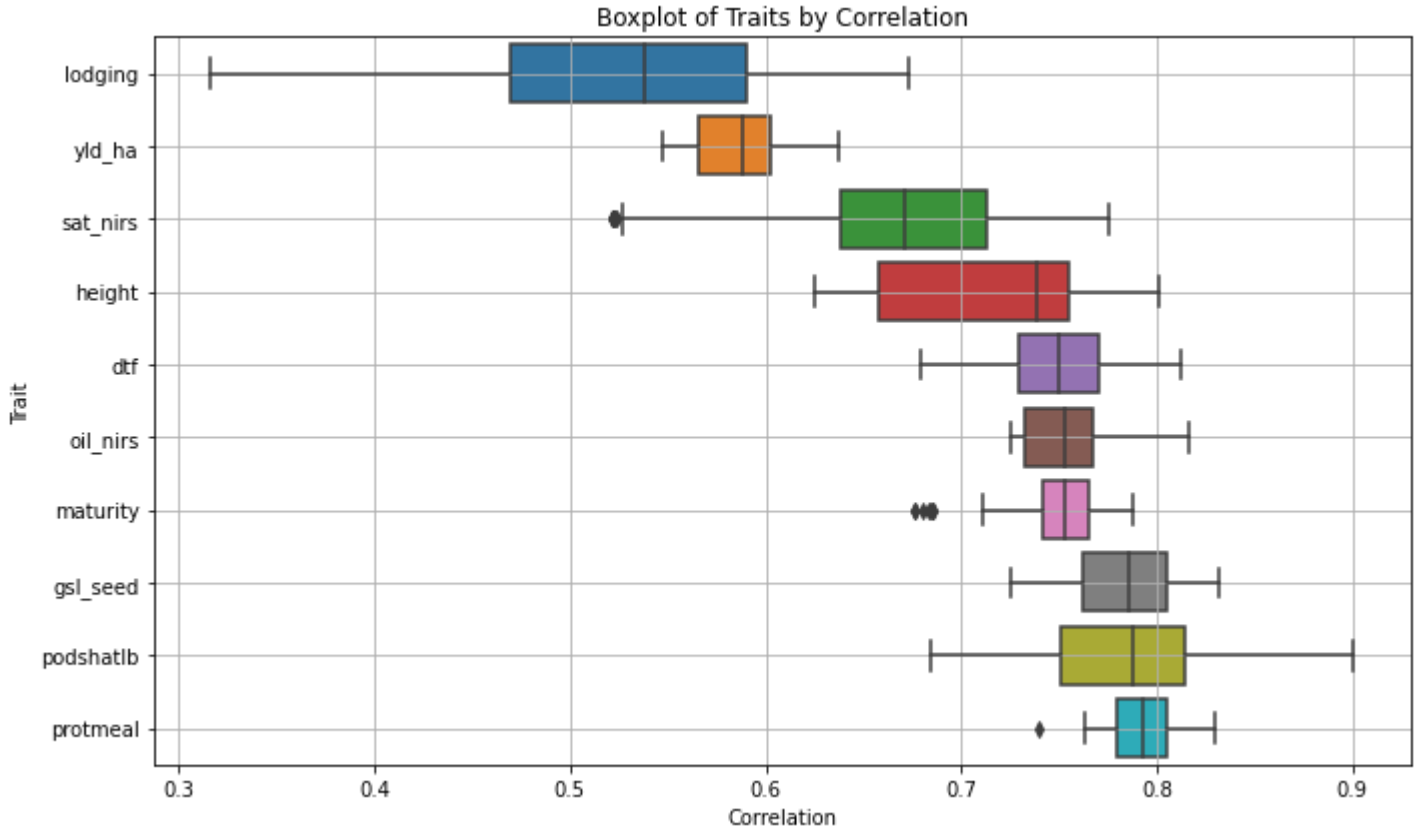
- Models respond similarly to sample size overall
- Less variance between models than between traits
- Neural network worse at lower hybrid numbers

Genomic Selection:
Trait Comparison across all Models



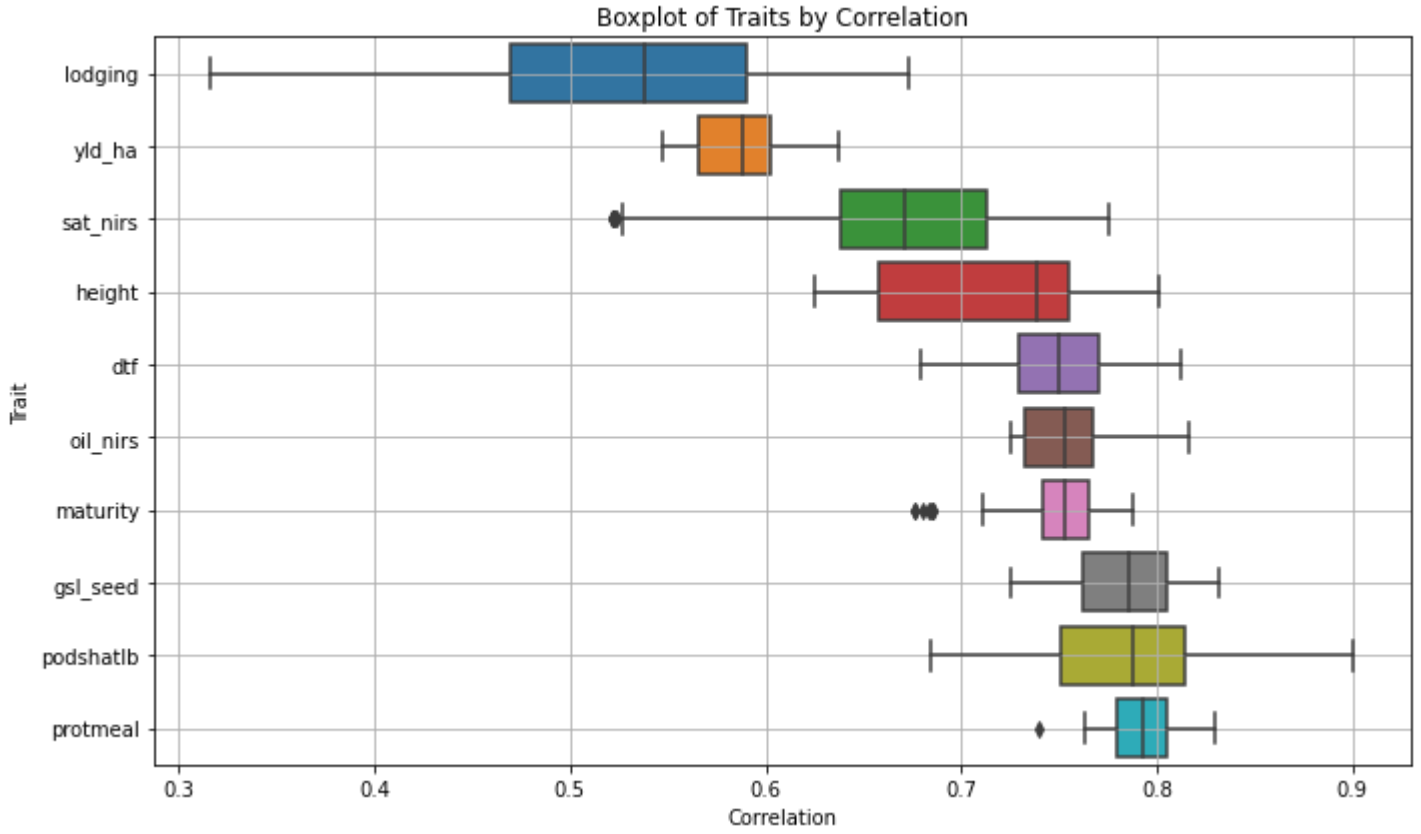
# Hybrids	Trait	Avg Correlation
2112	days to flower	74.9%
4958	glucosinolate	78.2%
801	height	71.5%
444	lodging	53.2%
4938	maturity	75.0%
4957	oil	75.3%
870	pod shatter	78.2%
4957	protein	79.4%
770	saturates	66.7%
5146	yield	58.8%

Genomic Selection:
Trait Comparison across all Models



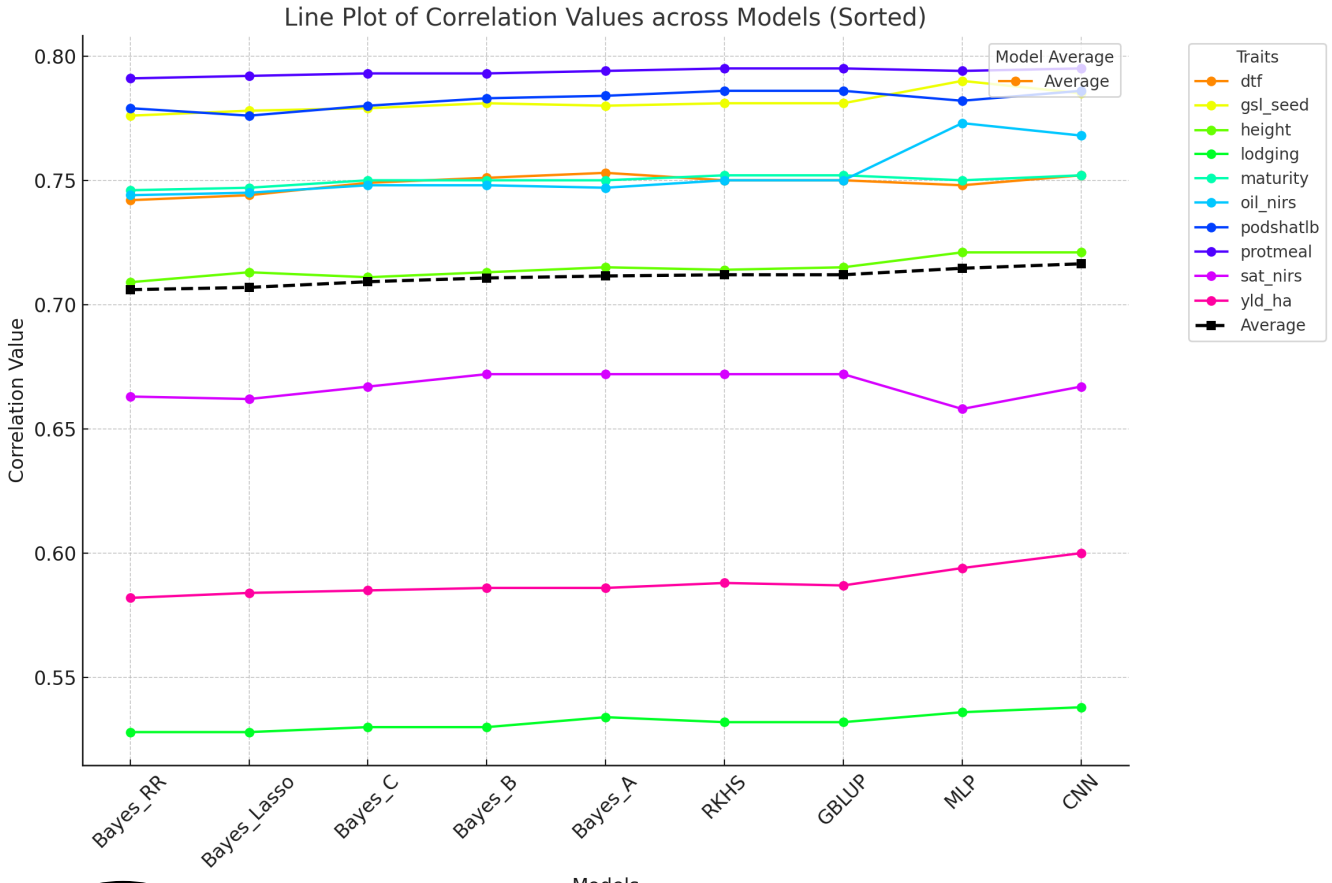
# Hybrids	Trait	Avg Correlation
2112	days to flower	74.9%
4958	glucosinolate	78.2%
801	height	71.5%
444	lodging	53.2%
4938	maturity	75.0%
4957	oil	75.3%
870	pod shatter	78.2%
4957	protein	79.4%
770	saturates	66.7%
5146	yield	58.8%

Genomic Selection:
Trait Comparison across all Models



# Hybrids	Trait	Avg Correlation
2112	days to flower	74.9%
4958	glucosinolate	78.2%
801	height	71.5%
444	lodging	53.2%
4938	maturity	75.0%
4957	oil	75.3%
870	pod shatter	78.2%
4957	protein	79.4%
770	saturates	66.7%
5146	yield	58.8%

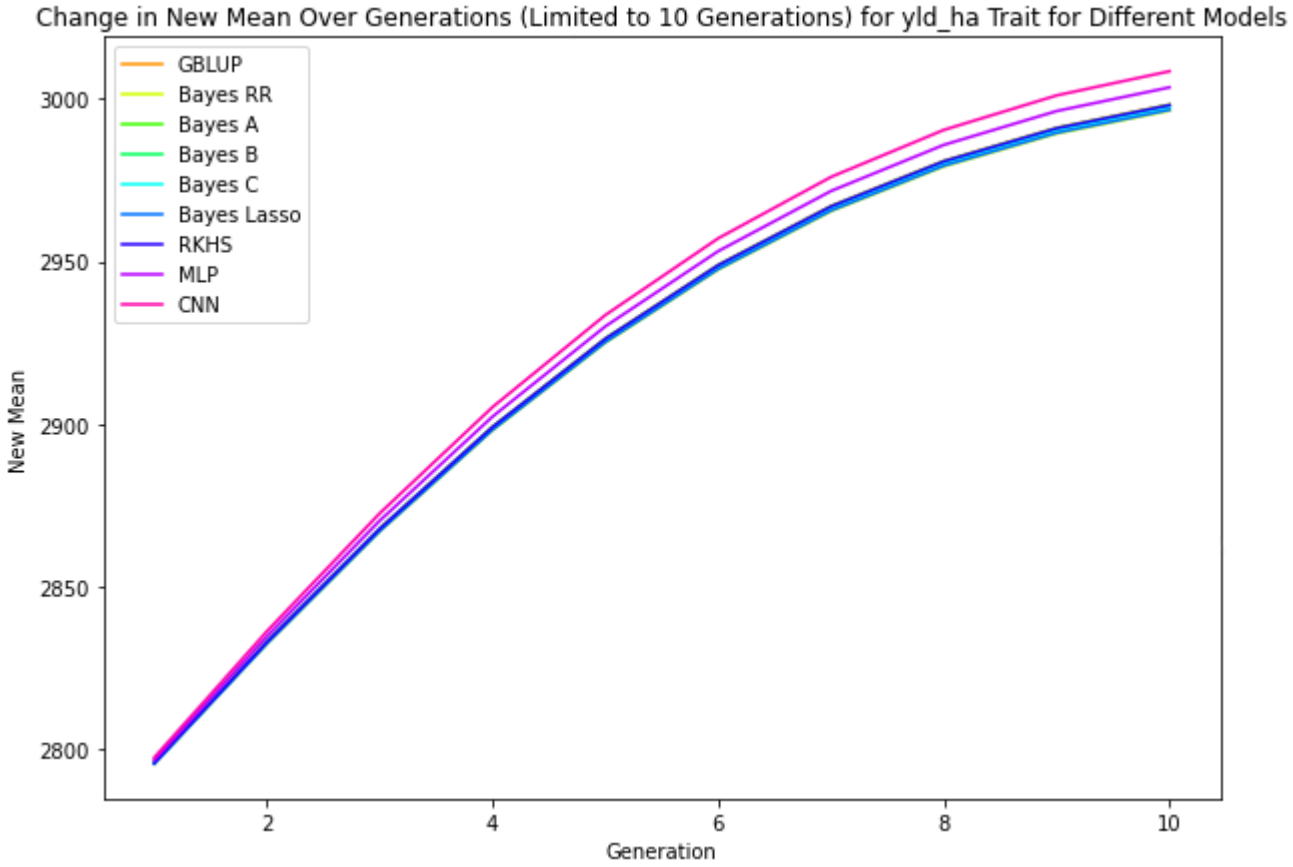
Genomic Selection: Model x Trait Comparison



70.6% 70.7% 70.9% 71.1% 71.1% 71.2% 71.2% 71.5% **71.6%**

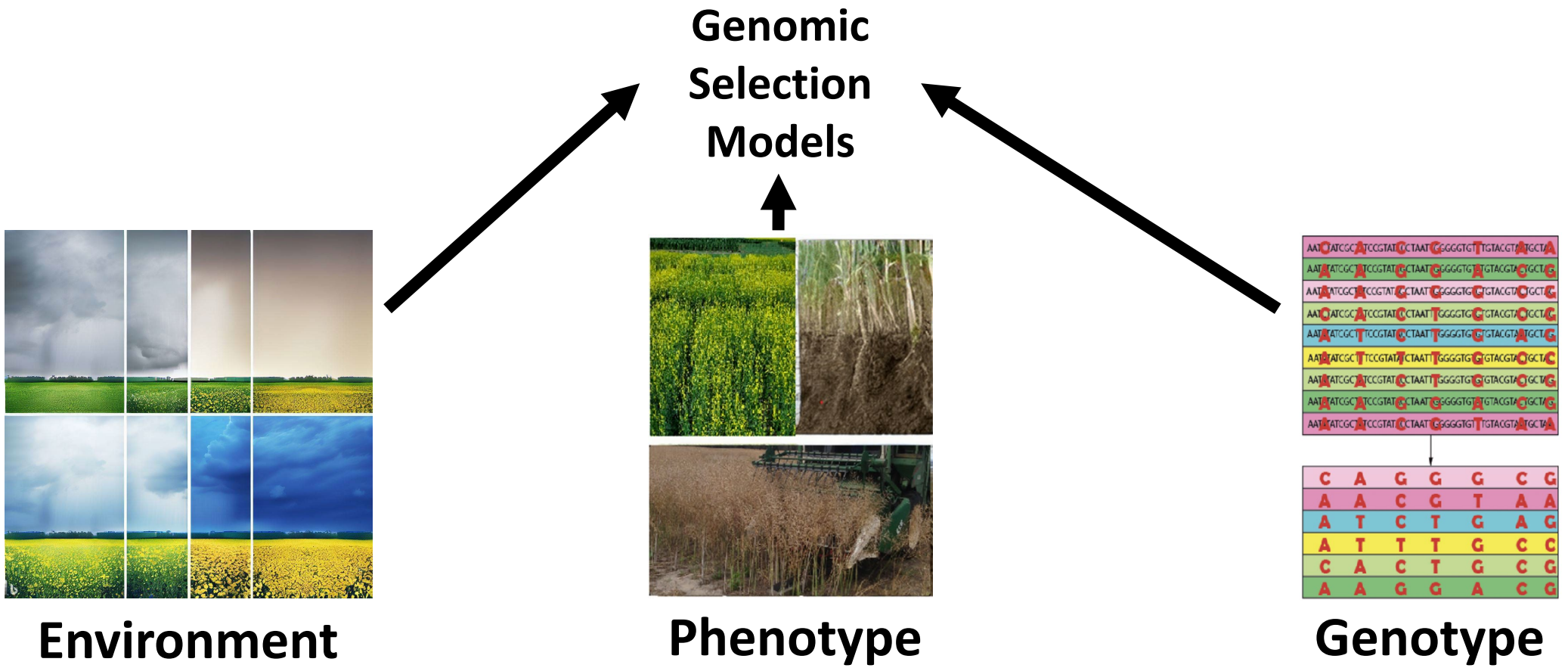
# Hybrids	Trait	Min	Max	Avg	Variance
2112	days to flower	74.2%	75.3%	74.9%	1.1%
4958	glucosinolate	77.6%	79.0%	78.2%	1.4%
801	height	70.9%	72.1%	71.5%	1.2%
444	lodging	52.8%	53.8%	53.2%	0.9%
4938	maturity	74.6%	75.2%	75.0%	0.6%
4957	oil	74.5%	77.3%	75.3%	2.8%
870	pod shatter	77.6%	78.6%	78.2%	1.0%
4957	protein	79.1%	79.5%	79.4%	0.4%
770	saturates	65.8%	67.2%	66.7%	1.4%
5146	yield	58.2%	60.0%	58.8%	1.9%

Genomic Selection: Model x Trait Comparison

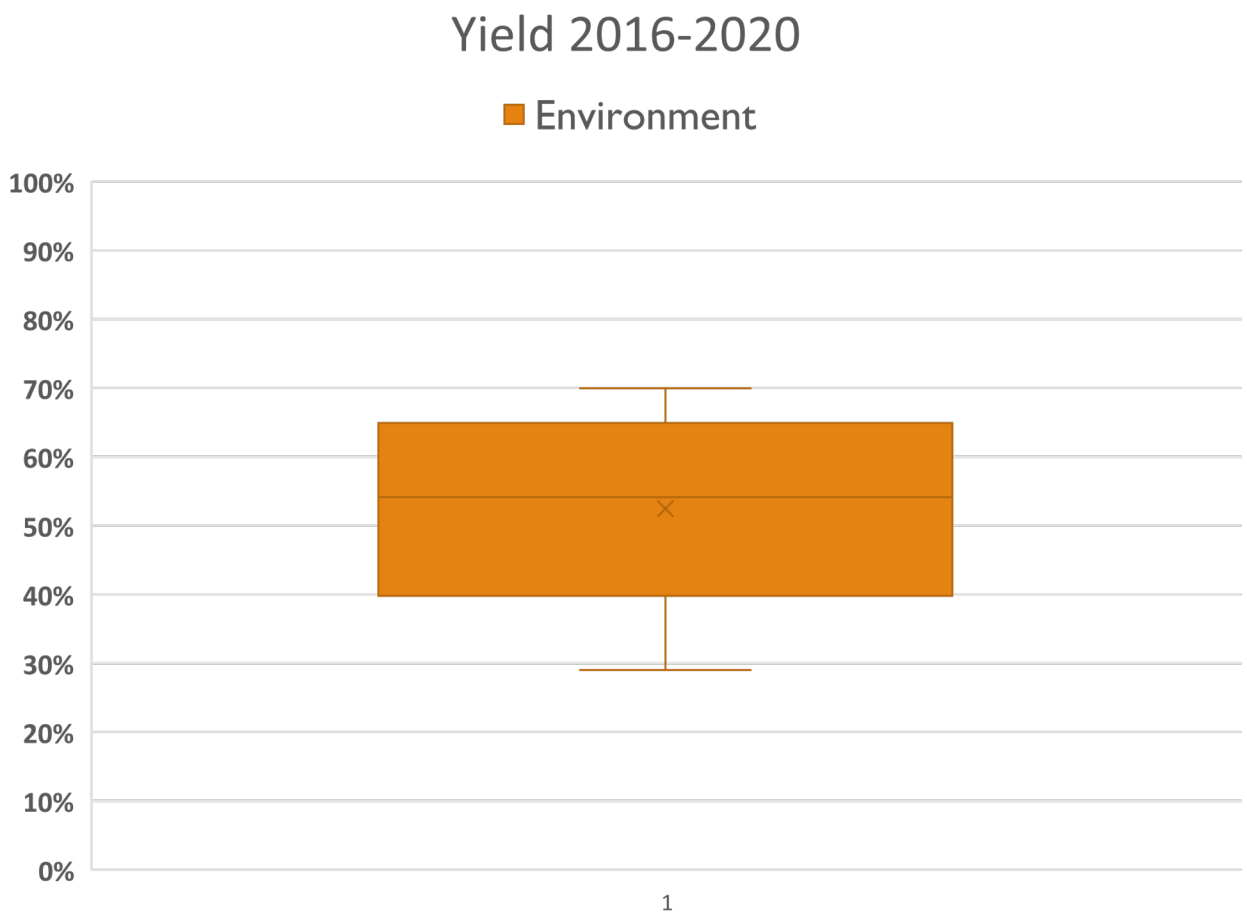


- Impact of data quality and quantity is more impactful than model choice
- Small model variance is measurable over generations
- Optimized AI models have small advantage overall
  - Further improvements and innovations in this field may increase gains

Breeding Canola: Adding the Environment



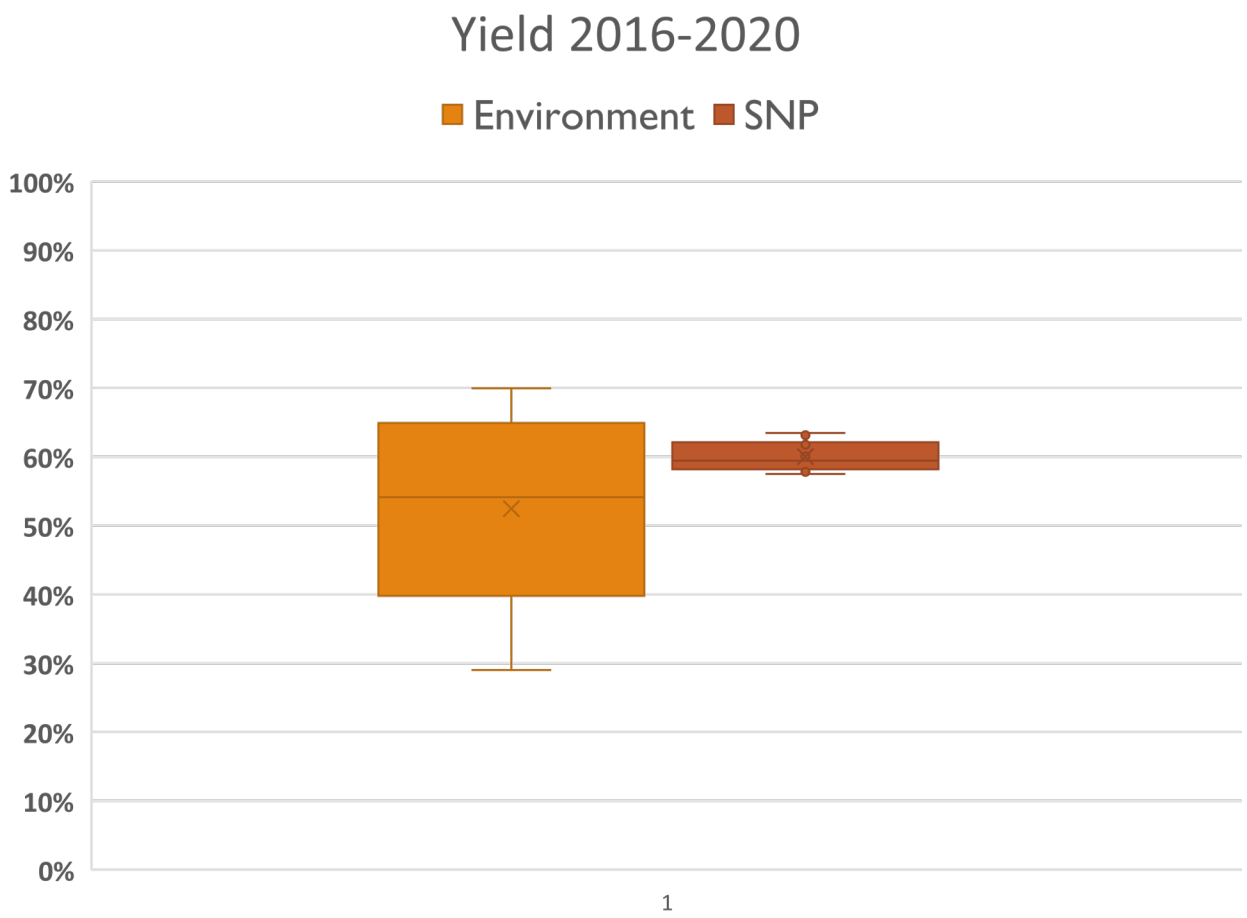
Genomic Selection: Environmental Data



- Environment Only
  - 108 Site years (GPS specific)
  - 792 environmental parameters
  - Prediction of average yield of site

Model	Pearson	StdDev	# Variables
Environment	52%	14%	1K

Genomic Selection: Environmental Data

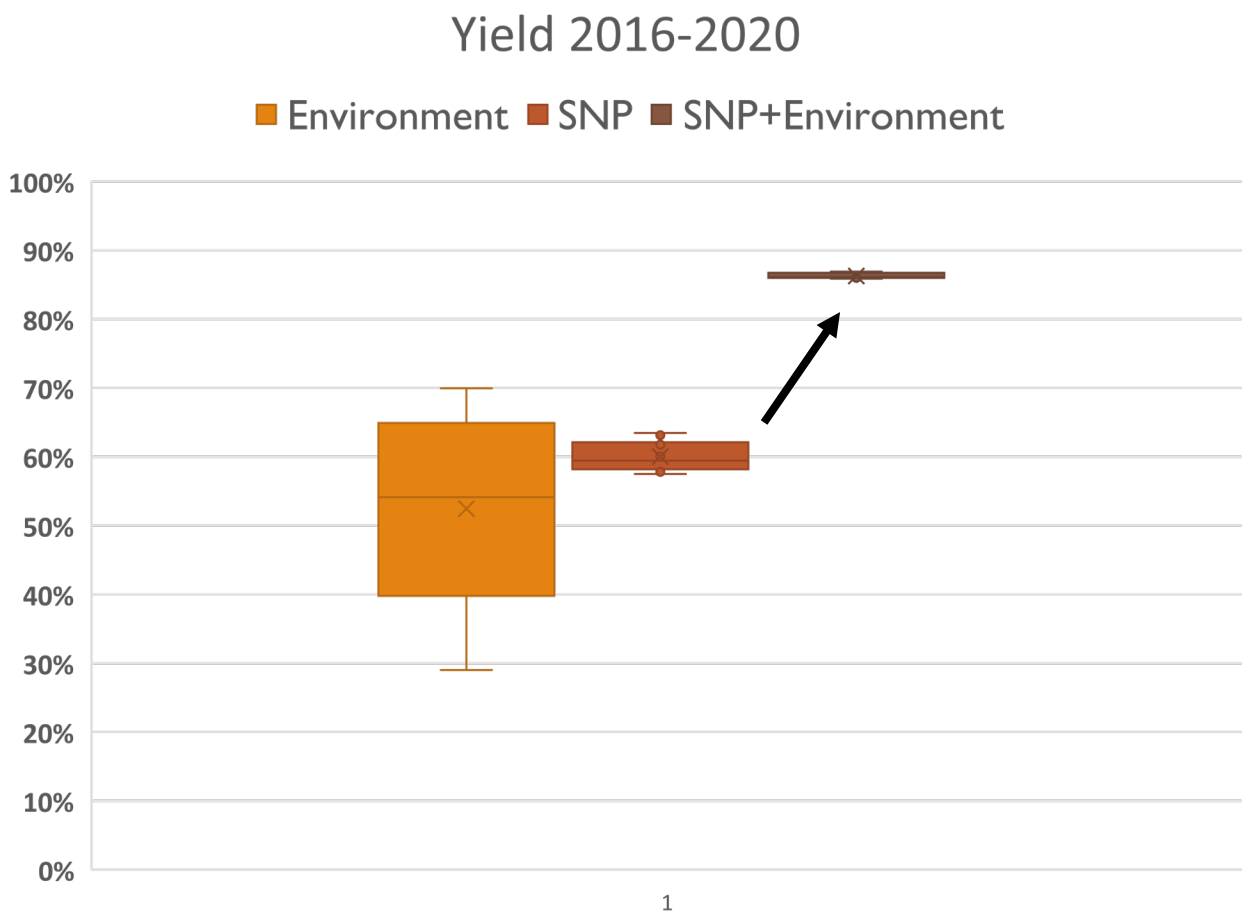


- Genotype
  - 18,945 SNP Markers
  - 5146 Hybrids (Mixed model average)

5146 x 18,945 Matrix
97 Million Cells

Model	Pearson	StdDev	# Variables
Environment	52%	14%	1K
Genotype	60%	2.4%	19K

Genomic Selection: Environmental Data

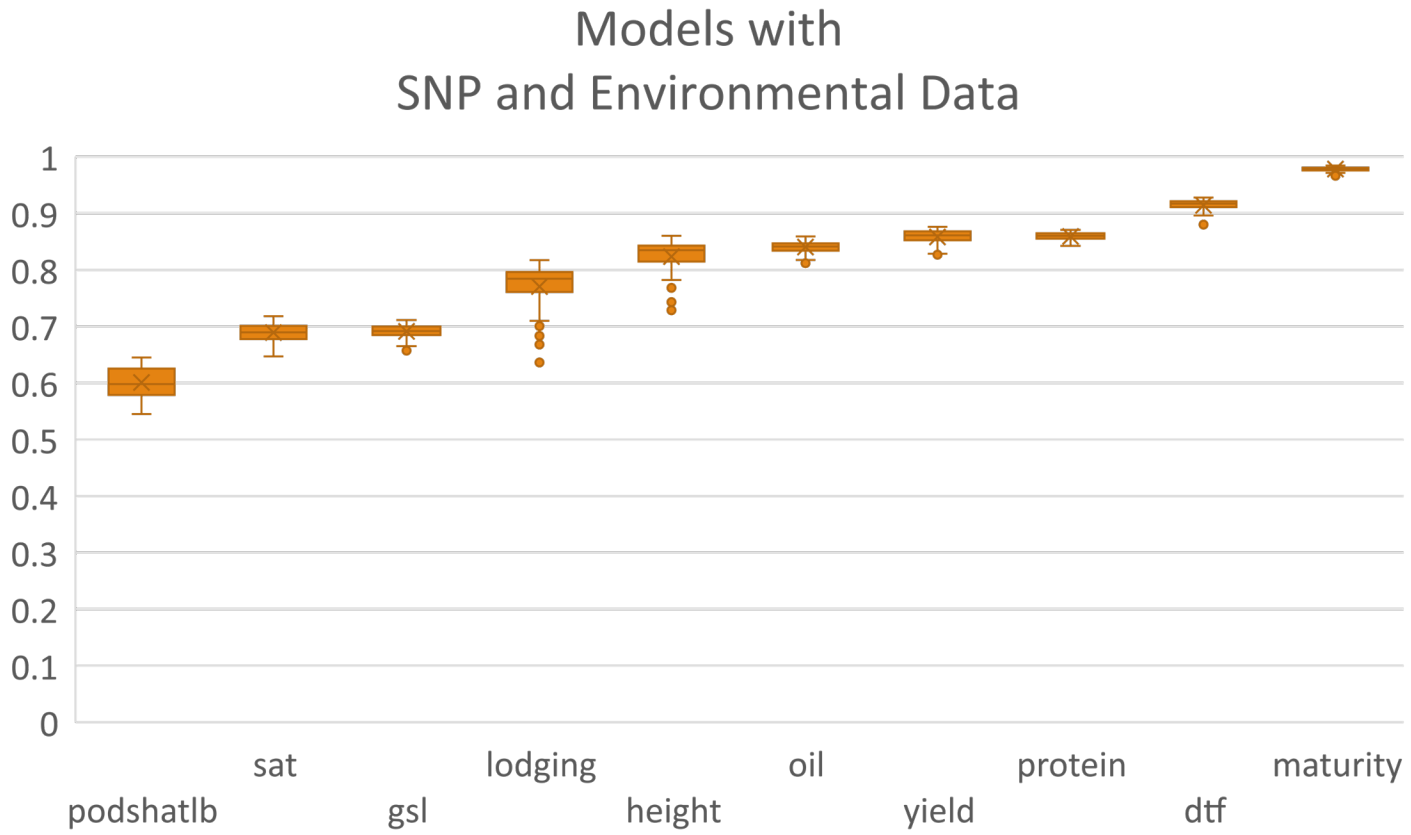


- Genotype x Environment
  - 108 Site years
  - 33,234 hybrid x Site Year Phenotype
  - 18,945 SNP + 792 Environment parameters

33,234 x 19,737 Matrix
655 Million Cells

Model	Pearson	StdDev	# Variables
Environment	52%	14%	1K
Genotype	60%	2.4%	19K
Genotype x Environment	85%	1.4%	20K

Genomic Selection: Environmental Data



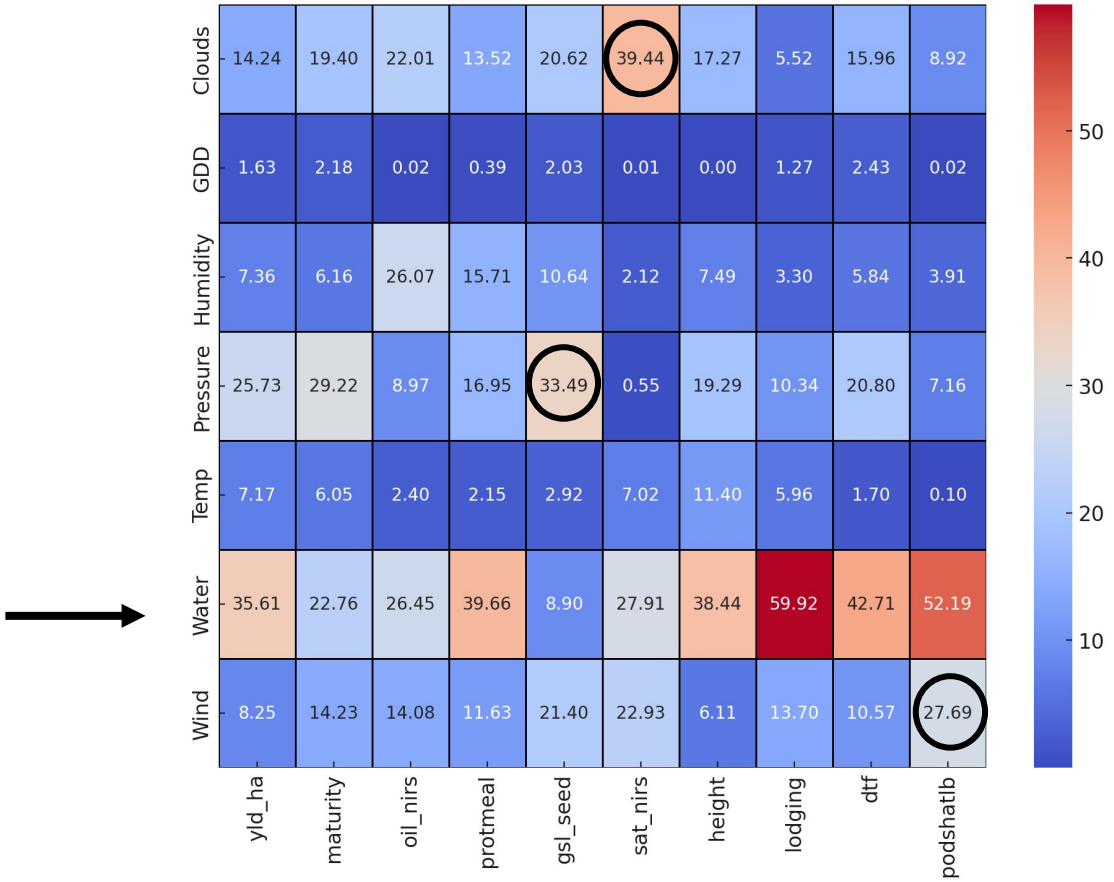
- Across all traits we observed an average 10% gain with environmental data
  - 792 parameters added

- **0.53 to 0.79** with Genotype only

↓

- **0.60 to 0.97** with Genotype + Environment

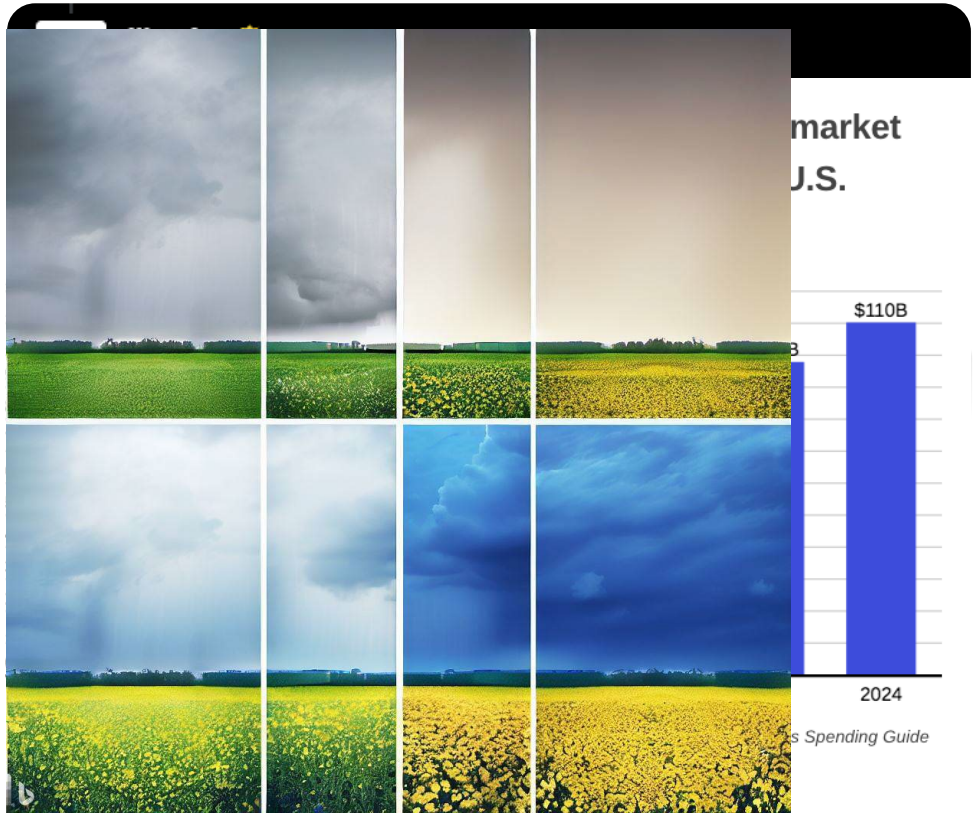
Genomic Selection: Environmental Variables



- Water (rain, snow) is the most important environmental variable overall
- Expected
  - Wind ~ pod shattering
    - Peaks 100 days post seeding
- Unexpected
  - Cloud cover ~ saturate levels
    - Peaks 70 days post seeding
  - Air pressure ~ Glucosinolate content
    - Cumulative mean 0 to 39 days post seeding

Genomic Selection: Opportunities

- Genotyping costs now less than phenotyping
  - Illumina $2/GB
  - Cost of SNP arrays and GBS < Plot
- Rapid growth in AI and machine learning
  - Spillover of gains from other fields
  - Human algorithms vs Neural Networks
- Genomic selection
  - Untested hybrids
  - Untested environments
    - Balance datasets (trials x locations x year)
    - Climate change and climate models



ACKNOWLEDGEMENTS

- DL Seeds
  - Valuable data set
  - Many colleagues Stephen, Lon, Evan, Janice, and the team
  - Technical support

- University of Manitoba
  - Supervisor Rob Duncan
  - Committee members Curt McCartney, Mike Domaratzki, and Dilantha Fernando

- University of Giessen
  - Collaborative support Iulian Gabur, Lennard Ehrig
  - Technical assistance

QUESTIONS