Exploring Non-Linear Genetic Relationships Between Correlated Traits Using Deep Learning

F. SHOKOR[1,2*], P. CROISEAU[2], R. SAINTILAN[1,2], T. MARY-HUARD[3], H. GANGLOFF[3], BCD.CUYABANO[2]

[1]Eliance, 149 Rue de Bercy, 75012 Paris, France

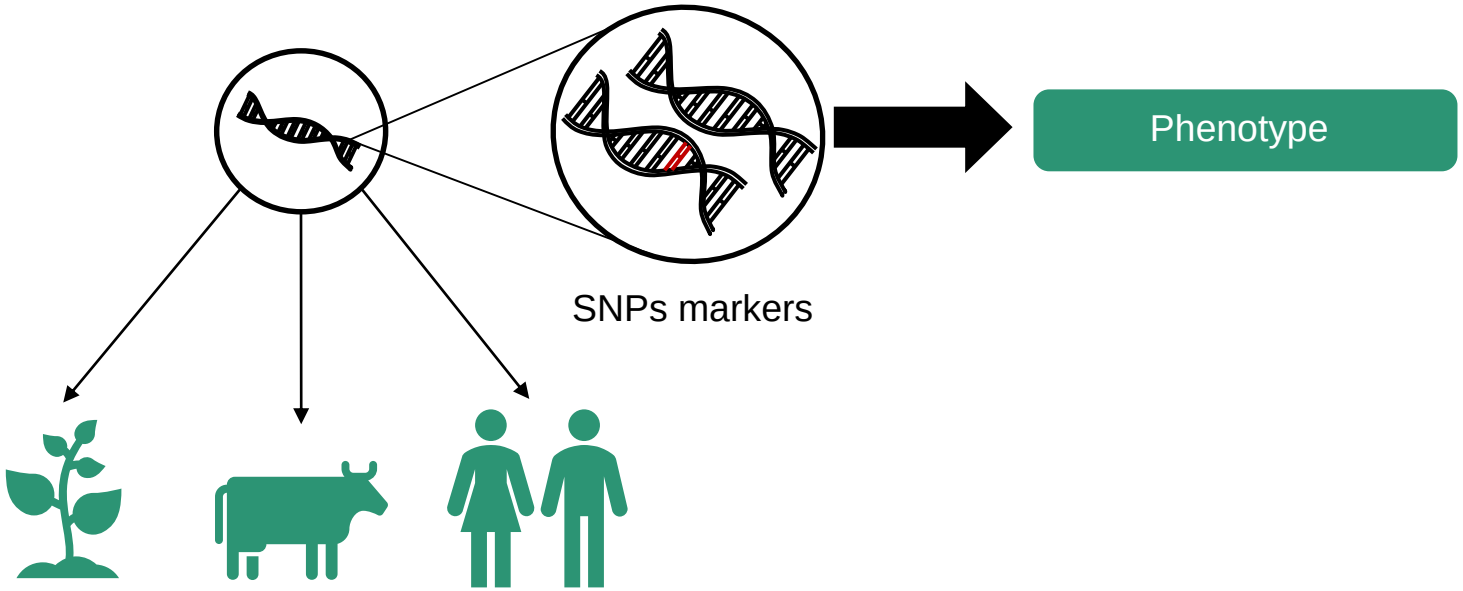
[2]Université Paris Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France

[3]Université Paris Saclay, INRAE, AgroParisTech, MIA, 91400 Palaiseau, France

74th EAAP Annual Meeting

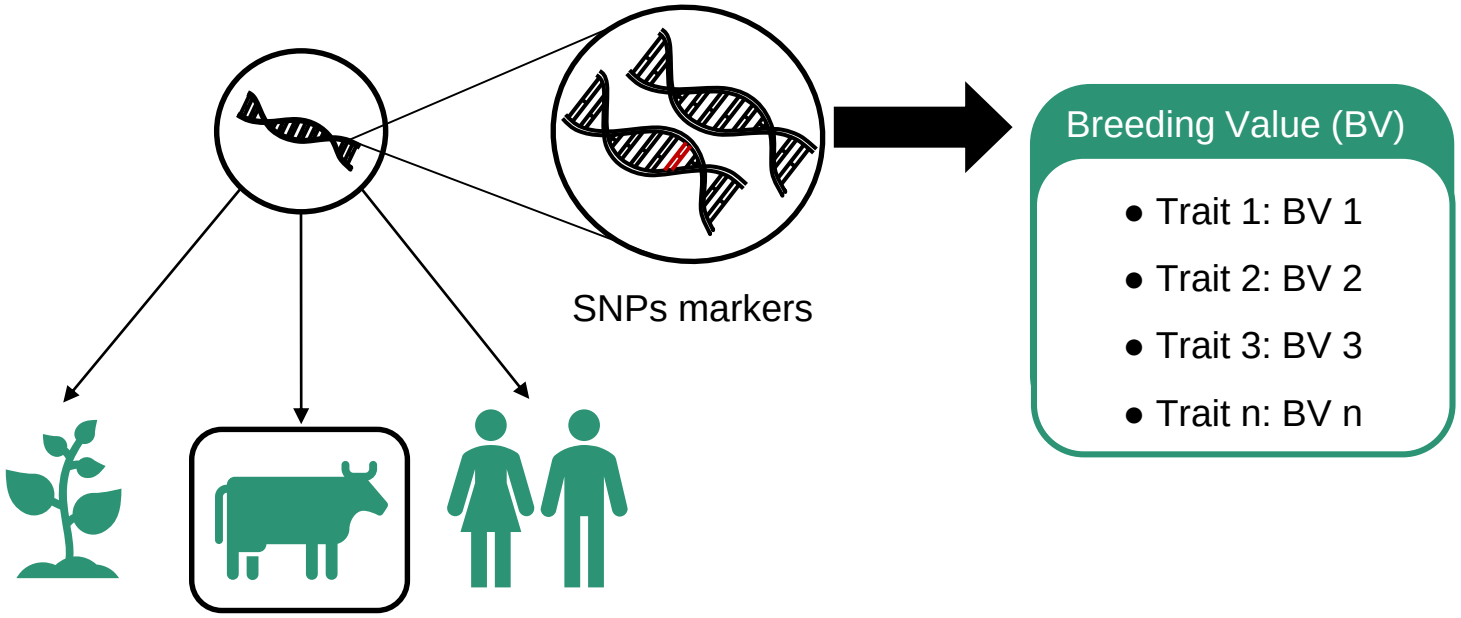
INTRODUCTION/GENETIC EVALUATION

Genomic Prediction



INTRODUCTION/GENETIC EVALUATION

Genomic Prediction

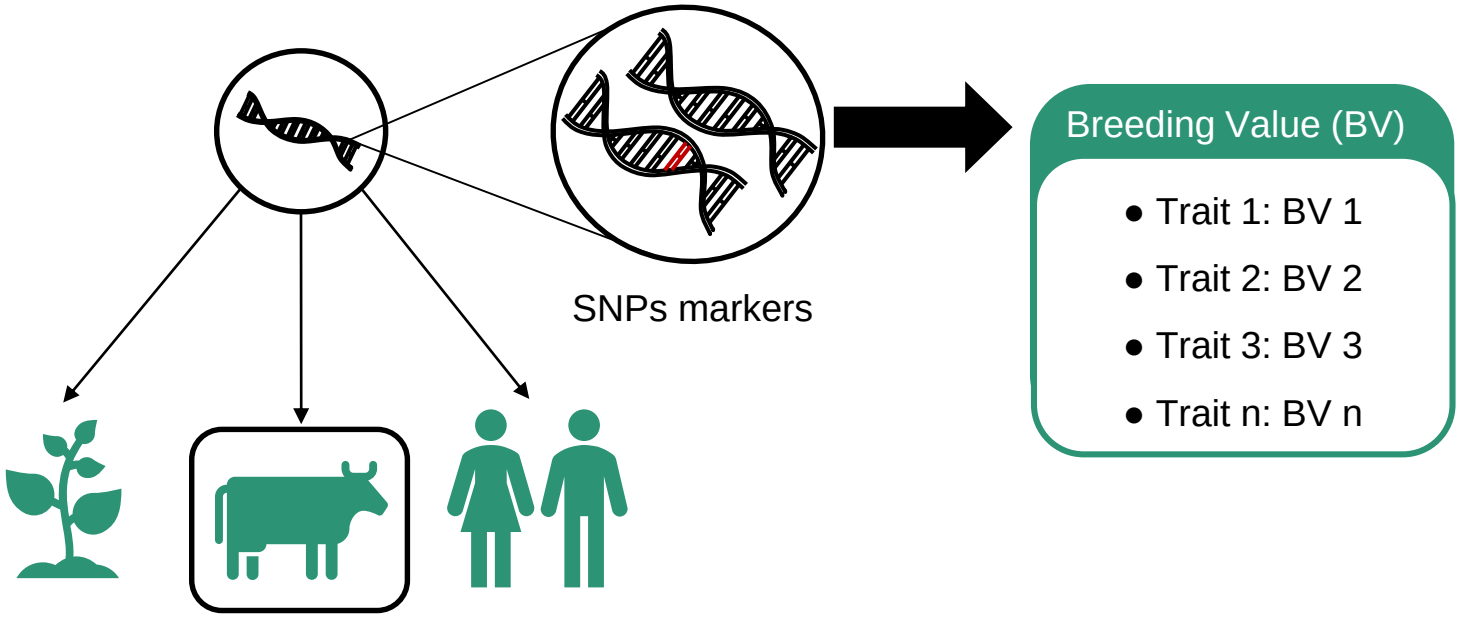


Breeding Value (BV)

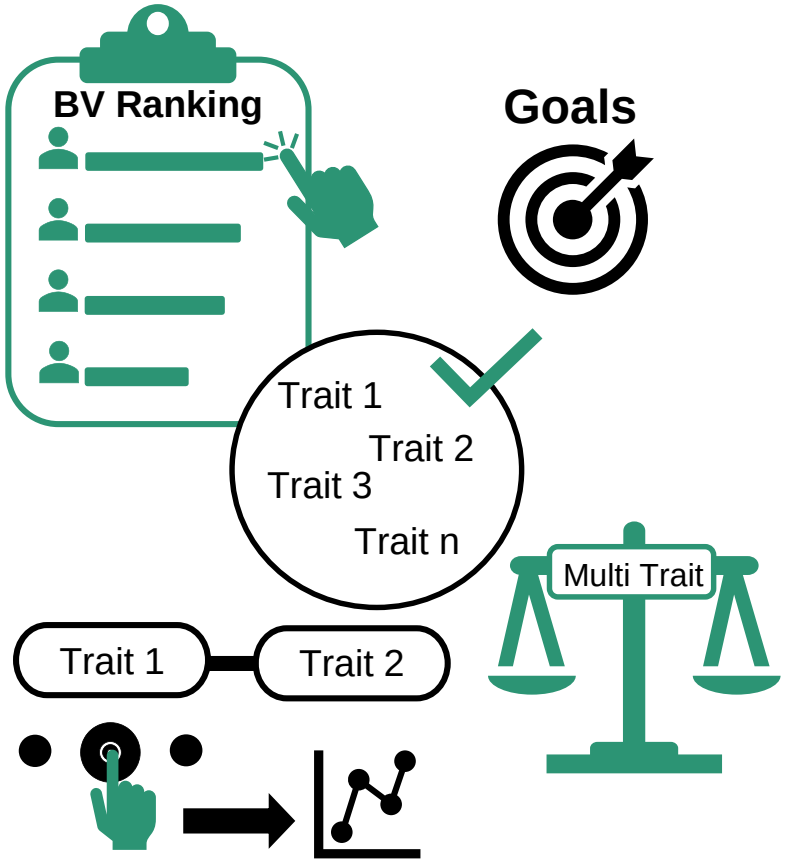
- Trait 1: BV 1
- Trait 2: BV 2
- Trait 3: BV 3
- Trait n: BV n

INTRODUCTION/GENETIC EVALUATION

Genomic Prediction



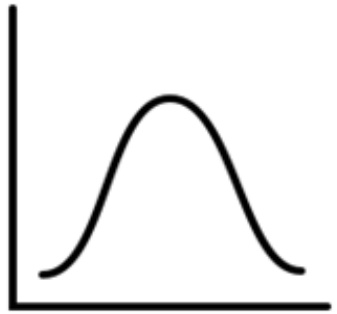
Genomic Selection



INTRODUCTION/STATISTICAL METHODS

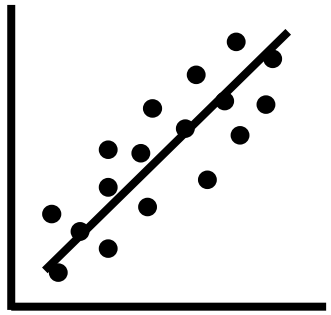
Statistical methods like **GBLUP** and Bayesian make the assumption of:

The genetic architecture follows a normal distribution



Linearity :

- SNPs and the trait of interest
- Between traits

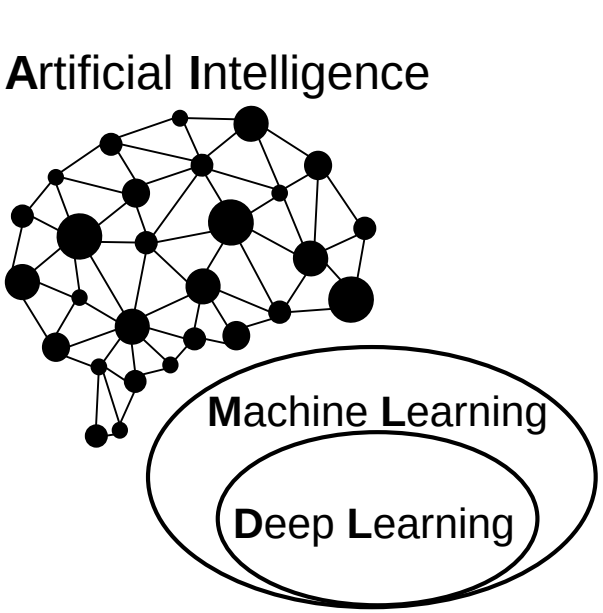


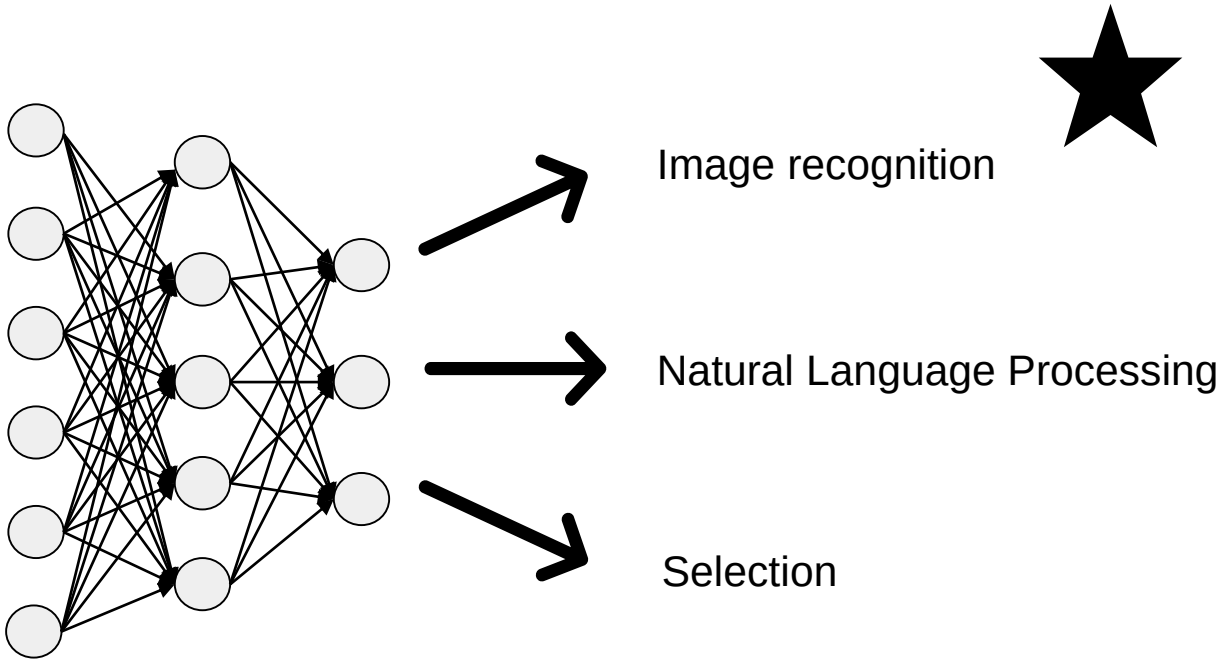
➔ True in many cases so it gives Good results on population level

But Not always

INTRODUCTION/DEEP LEARNING

Special Interest in **Deep Learning**





Artificial Neural Networks

Learn arbitrarily complex functions from input/output data

INTRODUCTION/DEEP LEARNING

Special Interest in **Deep Learning**

Advantages:

- Can handle large and complex genomic data sets
- Can identify complex & non linear patterns in genomic data that may be missed by traditional statistical methods

⇒ Hoping to IMPROVE PREDICTIONS

BUT NO…

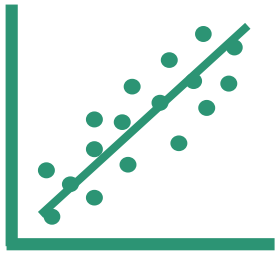
SAME or LESS than Statistical methods

Method or Usage?

PROBLEM/HYPOTHESIS

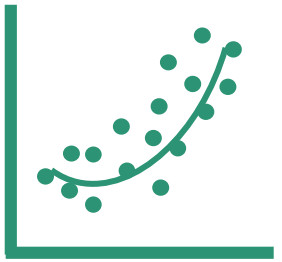
Genetic correlation between traits:

Linear: according to statistical methods



What if …

Nonlinear



→ **Loss** of prediction accuracy

→ **Errors** in selection

OBJECTIVES

1. Understand the consequences of non linear correlation to GBLUP prediction using two-trait model
2. Use DL to model non-linear correlation, and compare the results to GBLUP, with respect to:
    1. the ability of identifying non-linear genetic correlations
    2. the accuracy of GEBVs

SIMULATED DATA

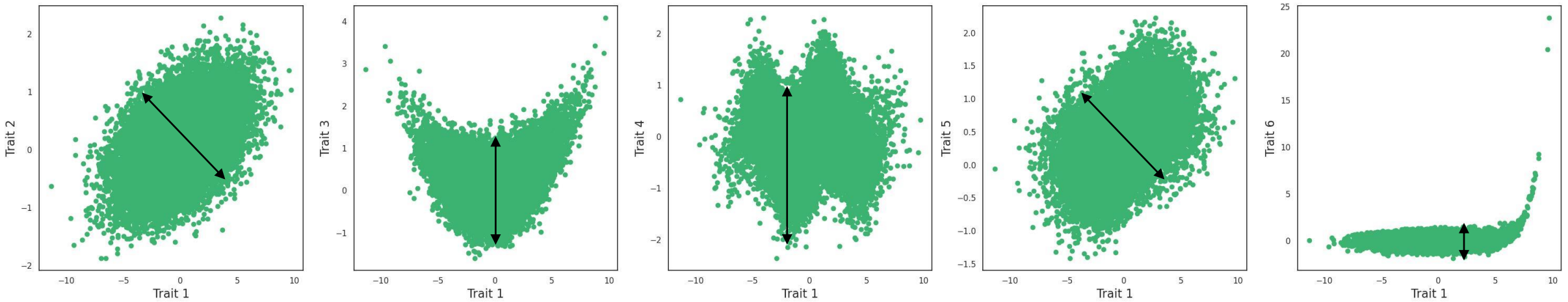
SAMPLES: 25,000 (Training: 20000/ Validation_DL: 2500/ Test: 2500) (20 Replicas)

QTL: 512

Phenotypes: Heritability = 0.3

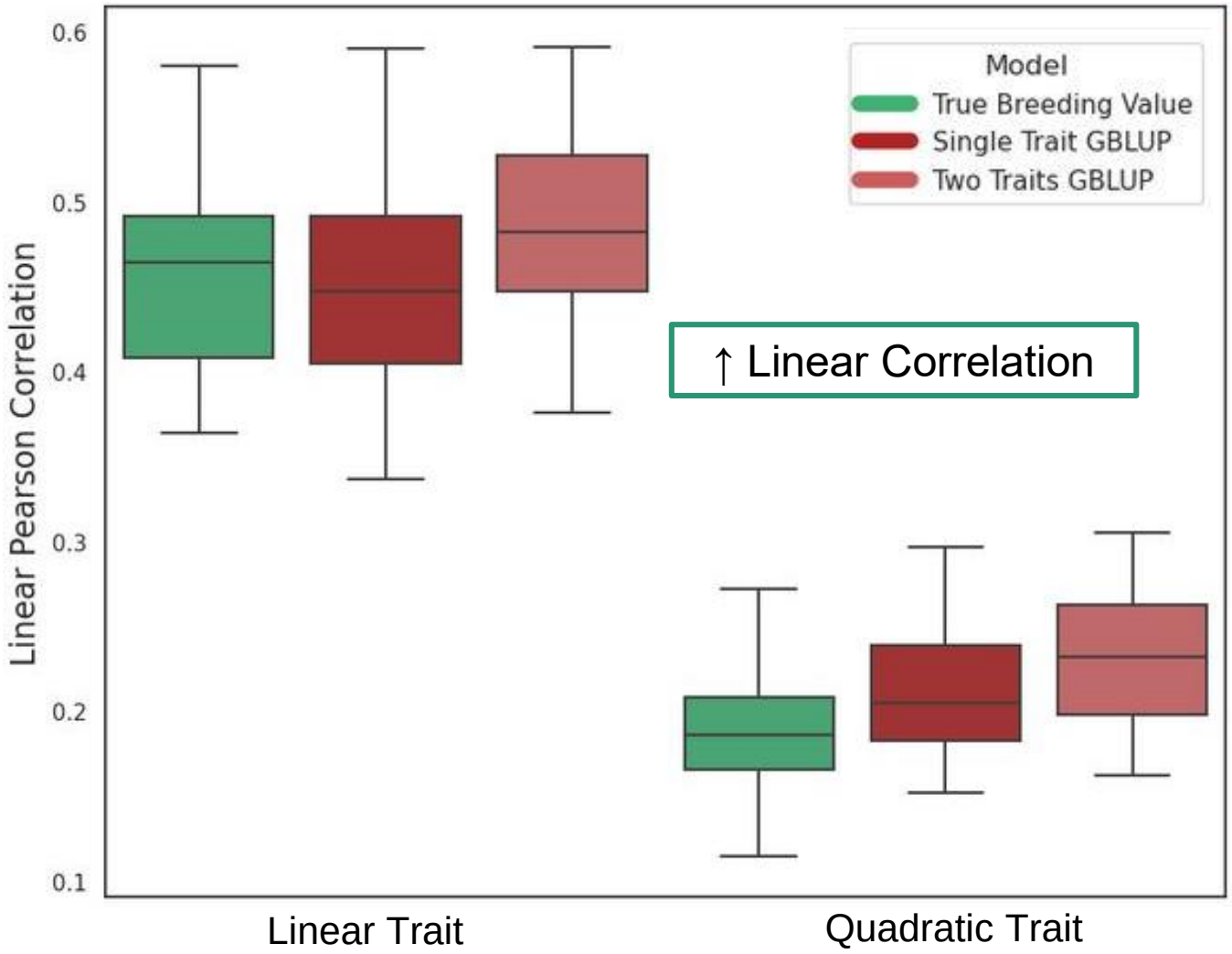
- Reference trait +

Linear Quadratic Sinusoidal Logistic Exponential

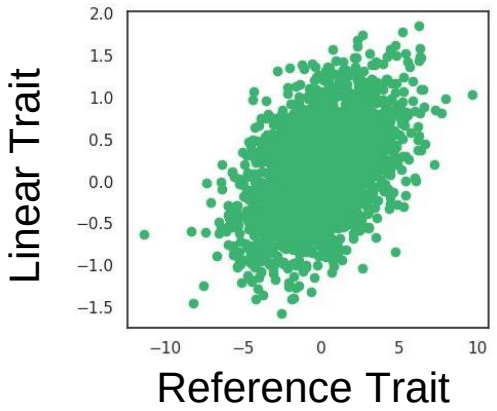


Genetic correlation = 0.5 ↕

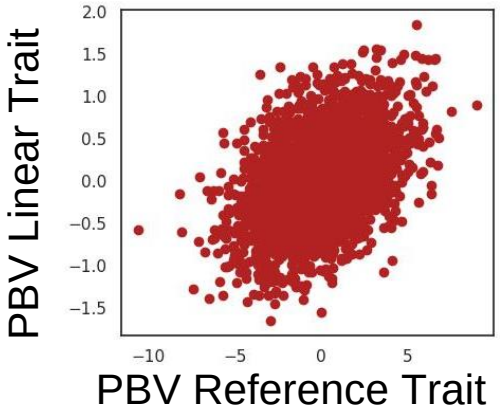
GBLUP RESULTS

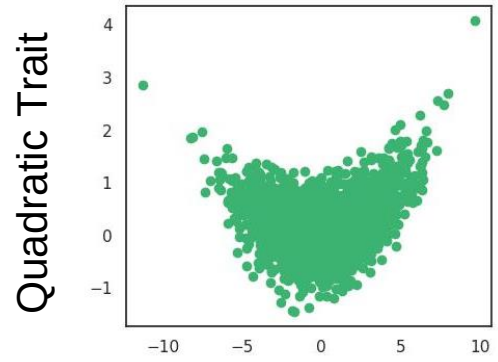


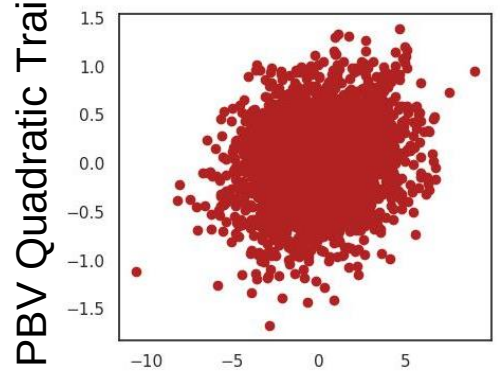
True Breeding Value



GBLUP

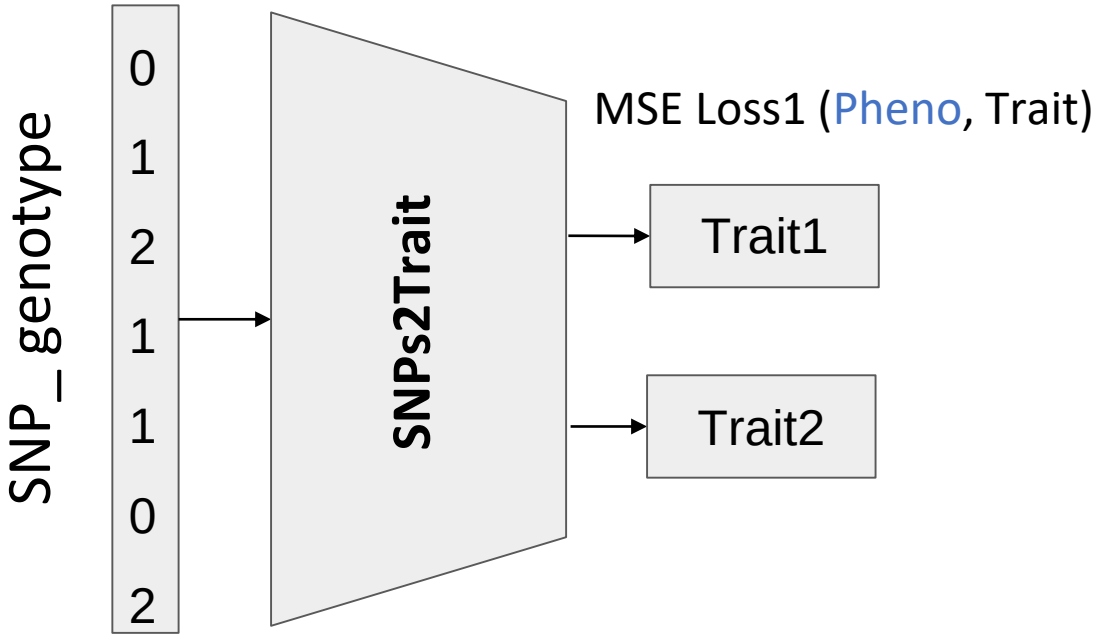






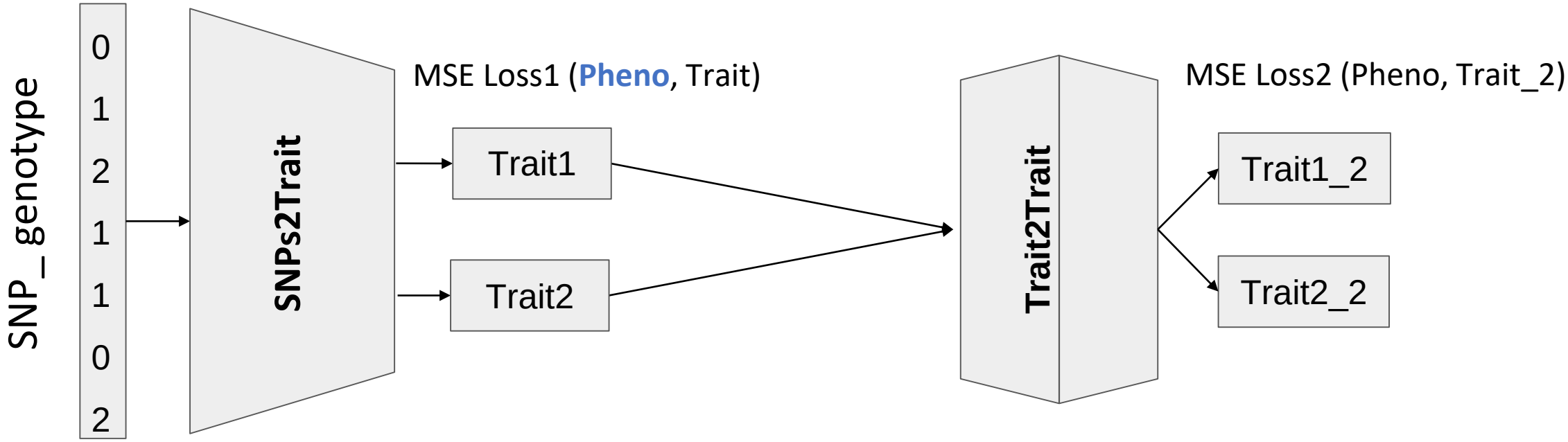
Linear Correlation just found

PURE DL 2-TRAITS MODEL



Predicts the BV from genomic data
by accounting on additive effects

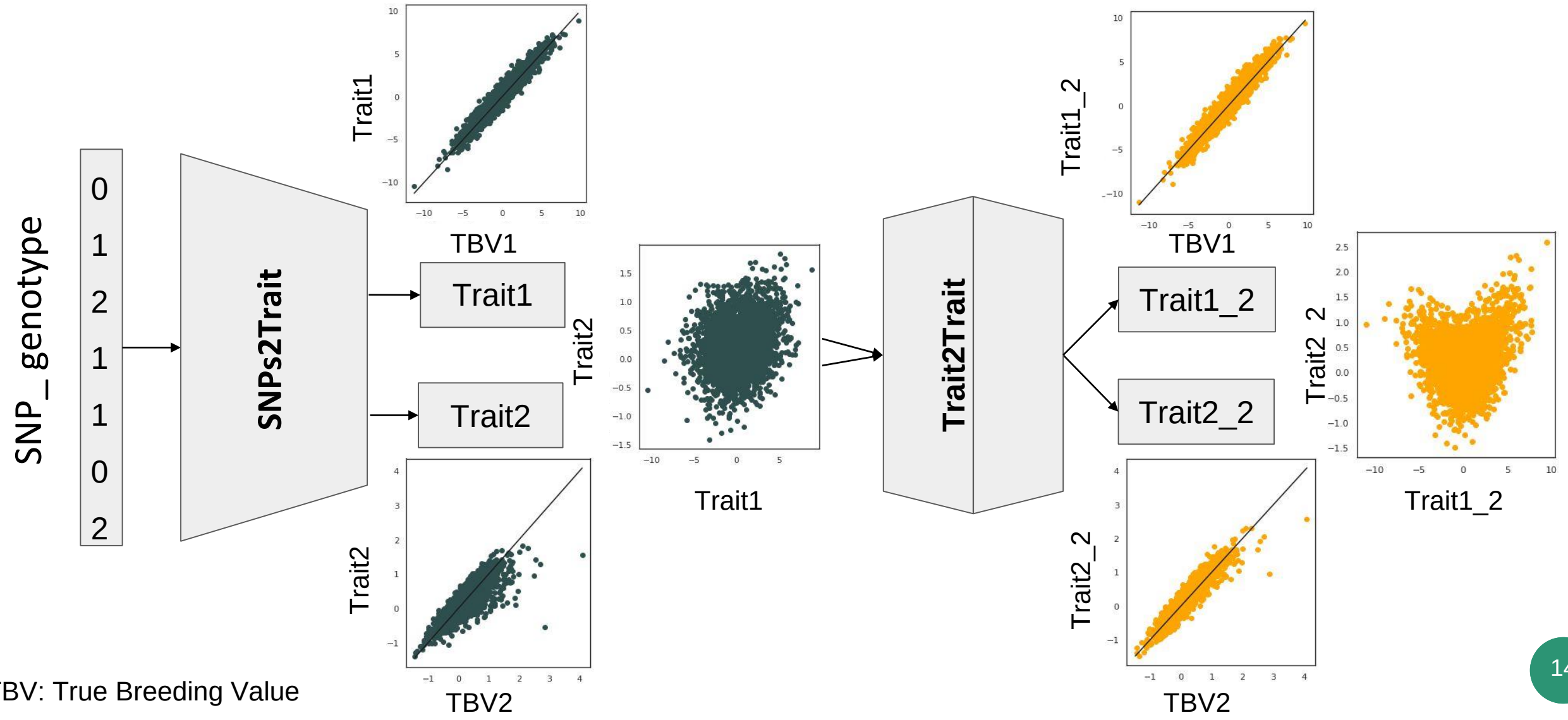
PURE DL 2-TRAITS MODEL



Predicts the BV from genomic data by accounting on additive effects

Predicts the BV according to the relationships with other traits

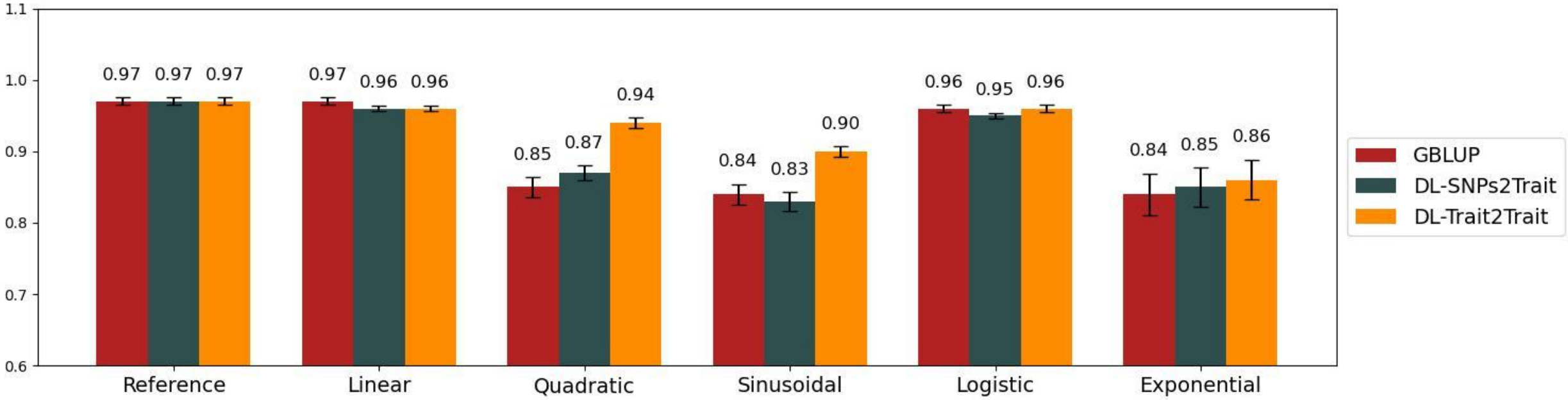
PURE DL 2-TRAITS RESULTS



TBV: True Breeding Value

PURE DL 2-TRAITS VS GBLUP

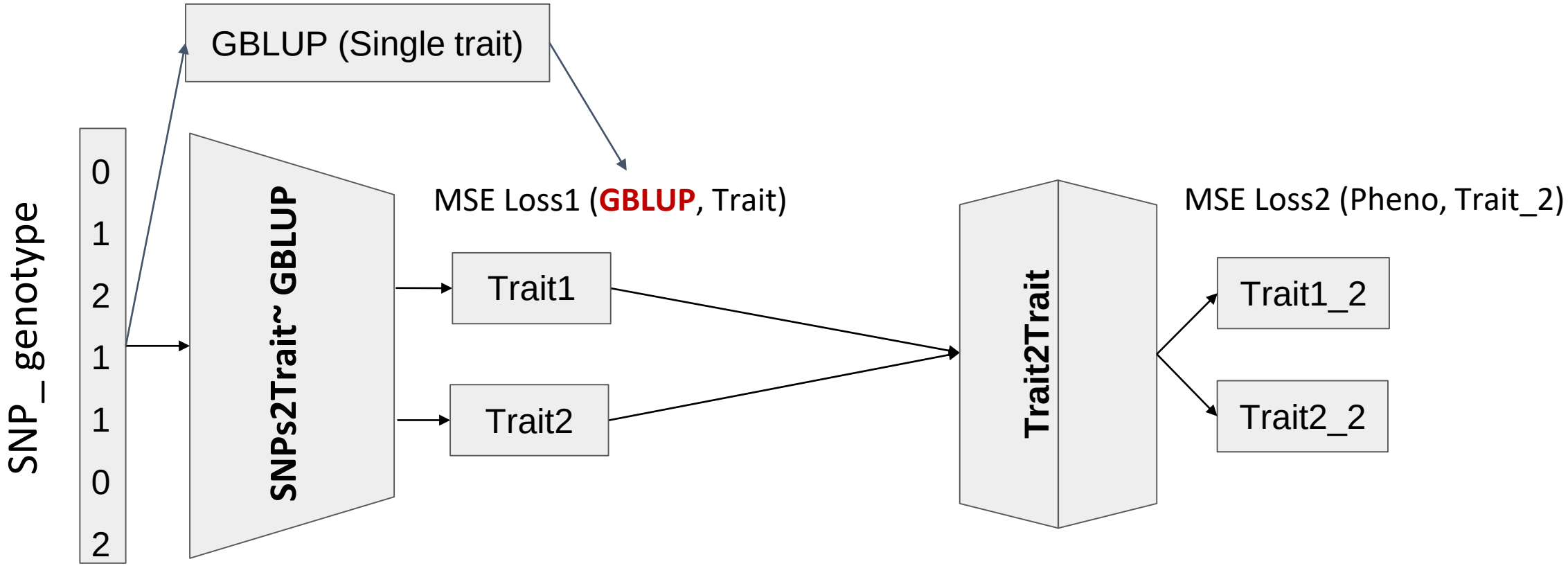
Prediction accuracy (cor(TBV, PBV))



Prediction from additive effects:
GBLUP > DL for some traits

Prediction from Traits relationships:
GBLUP < DL for Non linear traits

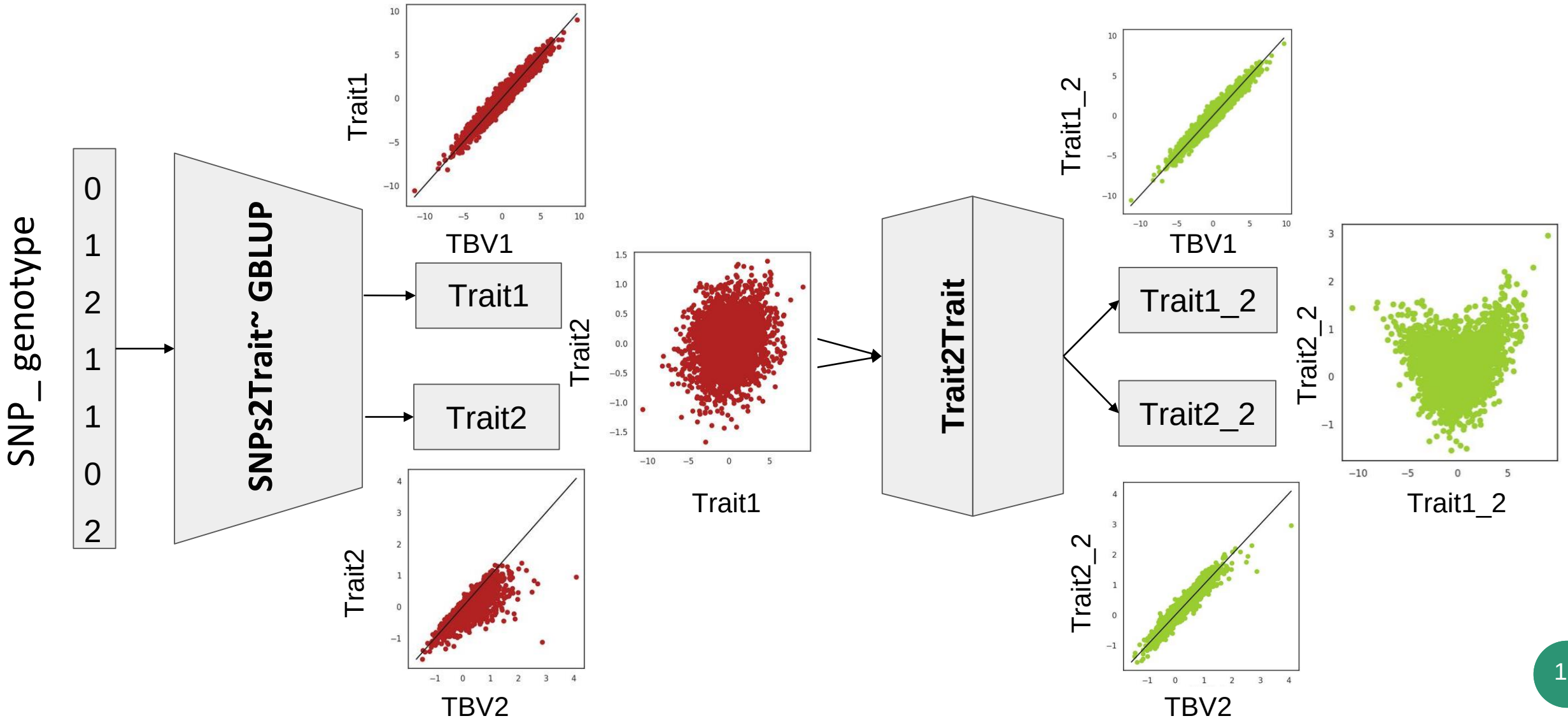
DL-GBLUP 2-TRAITS MODEL



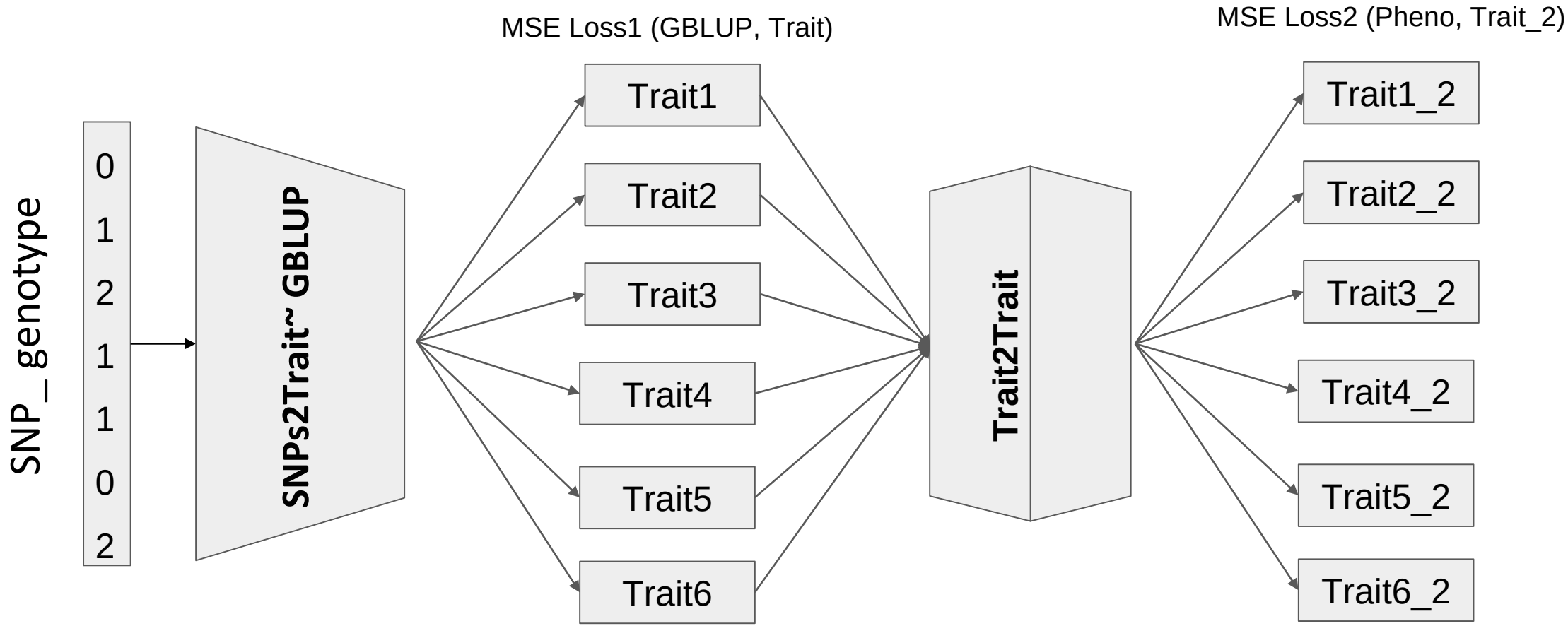
Learns to predict the output of GBLUP

Predicts the BV according to the relationships with other traits

DL-GBLUP 2-TRAITS RESULTS

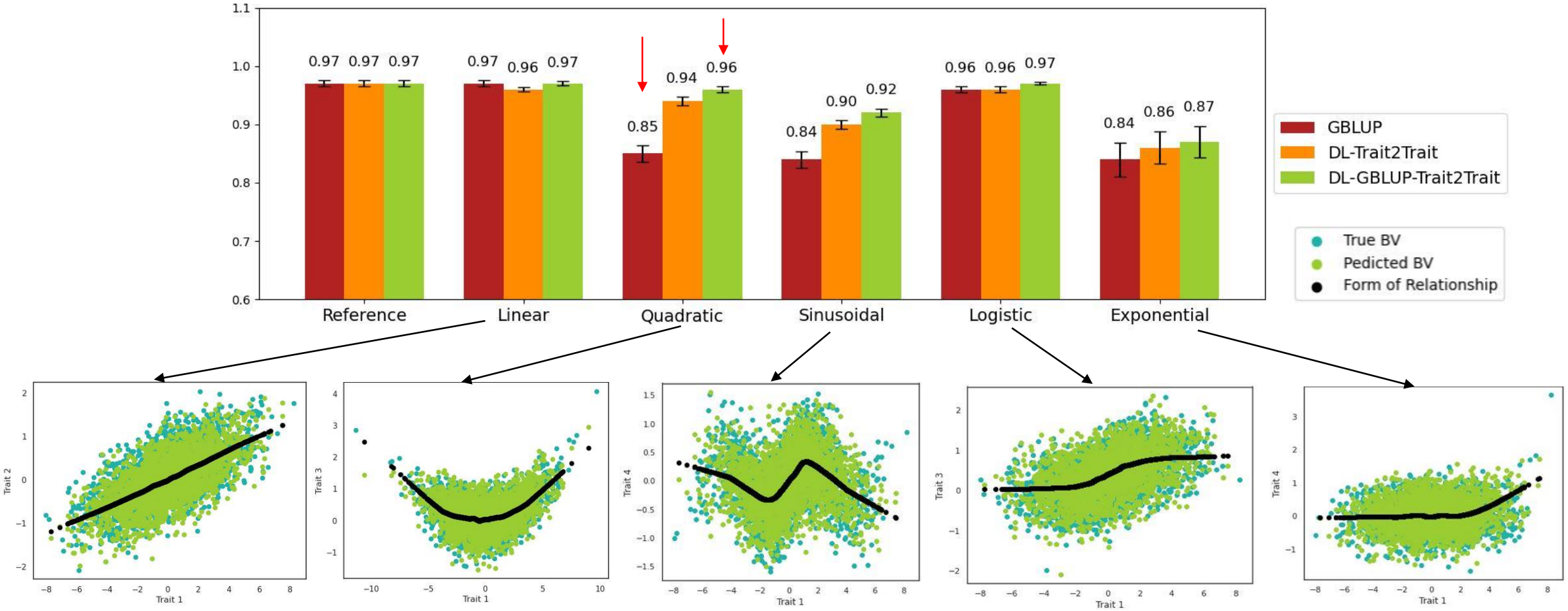


MULTITRAIT MODEL



RESULTS

Prediction accuracy (cor(TBV, PBV))



512 SNPs all as QTL and same across traits

CONCLUSION

Statistical methods (here GBLUP):

- powerful in performing predictions
- miss non-linear patterns

Deep Learning is good at identifying non-linear patterns in genomic data

Multi-trait DL-GBLUP combines the advantages of both methods, reducing their disadvantages:

➔ Increases the accuracy of GEBVs ➔ Provides more accurate selection

	GBLUP	Deep Learning	DL-GBLUP
Linear	✔	✔	✔
Non Linear	X	✔	✔✔

THANK YOU

ELIANCE
Des éleveurs. Une ambition.

INRAE

VetBIS

INSTITUT DE L'ELEVAGE idele

APIS-GENE
Investir Innover Valoriser

anrt
ASSOCIATION NATIONALE RECHERCHE TECHNOLOGIE

Cifre

Ecole Doctorale ABIES

université PARIS-SACLAY

Metaprogramme DIGIT-BIO

MODEL'S PARAMETERS

GBLUP:

- BGLR package (R), 25K iterations
- Training set: 22,500/ Validation set: 2,500

DL:

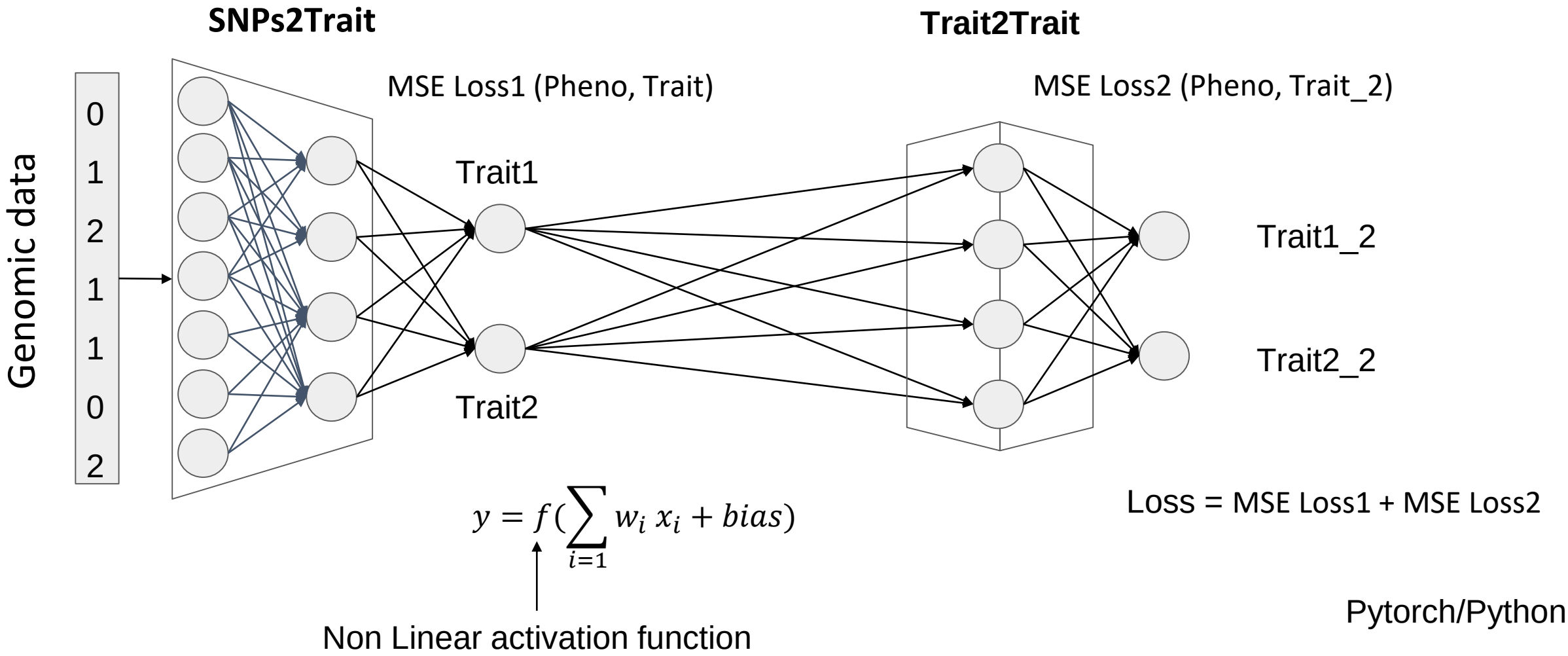
- MLP, batch size: 200, epochs: 100, Adam optimizer
- Encoder: Linear layer (512 → 400) → LeakyReLU → Linear layer (400 → nb of traits)
- Estimator: Linear layer (nb of traits → 400) → LeakyReLU → Linear layer (400 → nb of traits)

SIMULATION EQUATION

- *Reference trait :*
  - $y_1 = M_{1,i}\,\alpha_1 + e_{1i} = g_{1\,i} + e_{1\,i}$
  - *g : True breeding values*
  - *M : Centered QTL_genotypes*
  - α *: QTL effects*
- *Dependent Trait:*
  - $y_2 = g_{1,2\,i} + e_{1\,i}$
  - $g_{1,2\,i} \sim N(\mu = \rho_{1,2}\ (\sigma_{g2}/\sigma_{g1})\ f(g_{1\,i}),\ \sigma = G \times \text{heritability} \times (1 - \rho_{1,2}^2))$
  - *G = Genomic Relationship Matrix*
  - *f: Linear/Quadratic/Sinusoidal/Exponential/logistic*

DL 2-TRAITS MODEL

Multi Layer Perceptron: MLP



ADDING NO QTL SNPS EFFECT