Deep Learning for Crossbred Genomic Prediction

F. Shokor [1,2], P. Croiseau [2], H. Gangloff [3], S. Fritz [1,2], A.A.A. Martin[1,2], T. Mary-Huard[3,4], and B.C.D. 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, UMR MIA Paris-Saclay, 91120 Palaiseau, France

[4]Université Paris Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, 91190 Gif-sur-Yvette, France

76th EAAP Annual Meeting
Innsbruck – Austria

Background

- Crossbreeding: maximize strengths and minimize weaknesses from different breeds,
- Heterosis (hybrid vigor) can increase productivity, fertility, and health,
- Terminal crossbreeding systems greatly benefit from these aspects,
- Non-terminal crossbreeding systems: selection of the crossbred animals to serve as parents for future generations.

Challenges

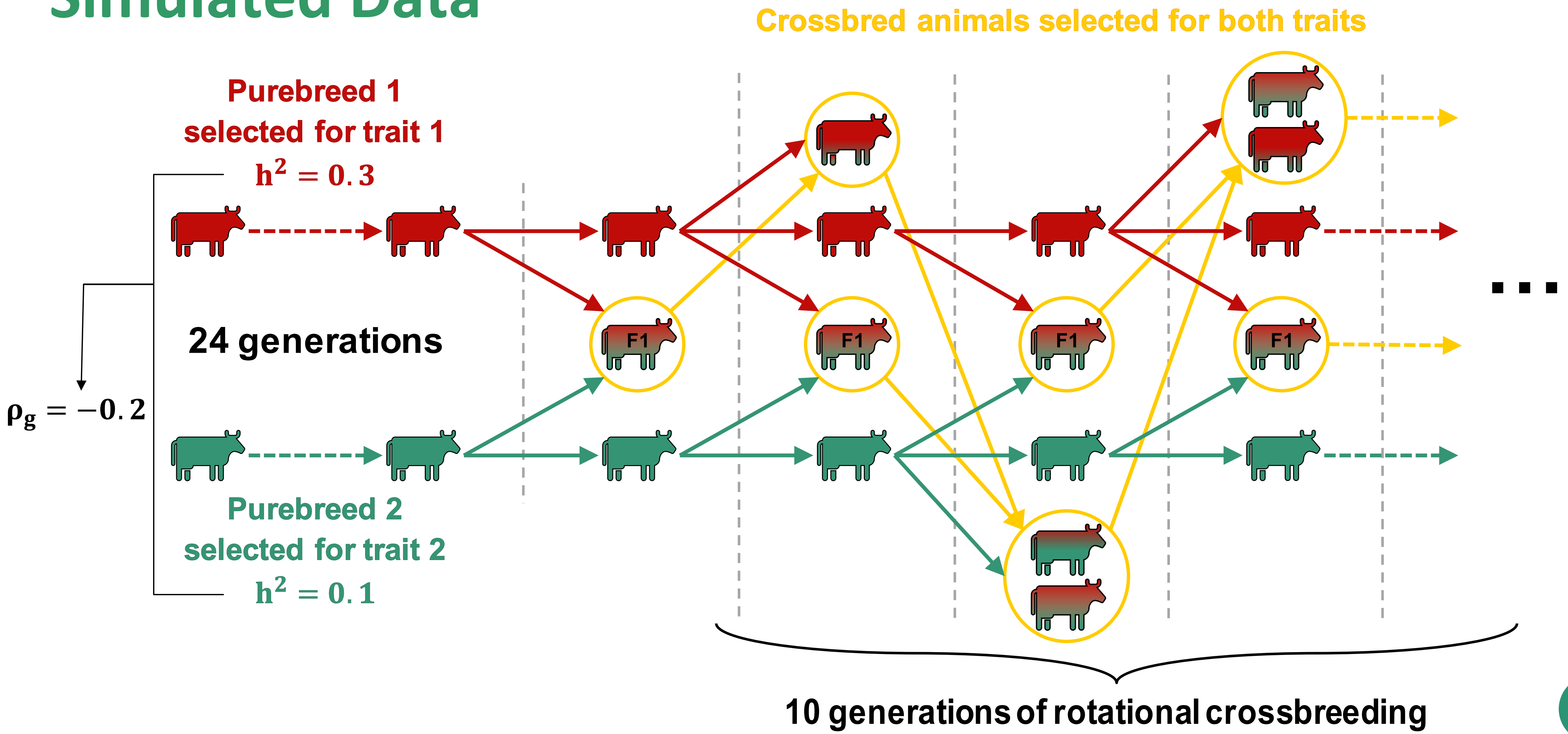
- Predicting genetic merit in crossbreds is harder than in purebreds,
- Different LD patterns, confounding of heterosis with additive effects, …
- Purebred-based prediction often lose accuracy in crossbreds,
  - BUT, accounting for the breed of origin of the alleles improves crossbred prediction,
- Traditional genomic models focus include heterosis as a mean crossbred effect.

Can Deep Learning improve crossbred prediction accuracy, and provide individual estimates of heterosis?

Objectives

- Use deep learning for crossbred prediction under varying heterosis levels,
- Combine additive genetic effects and heterosis for better prediction accuracy.

Simulated Data



Crossbred animals selected for both traits

Purebreed 1
selected for trait 1
$h^2 = 0.3$

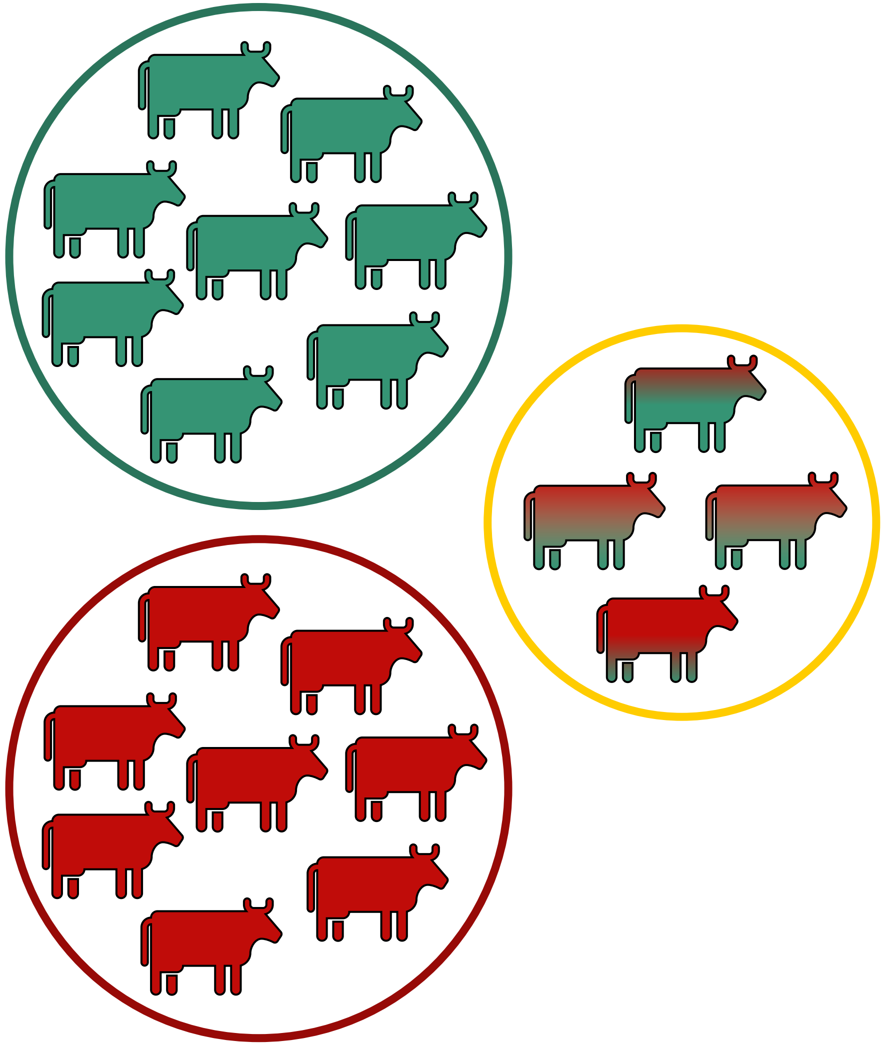
24 generations

$\rho_g = -0.2$

Purebreed 2
selected for trait 2
$h^2 = 0.1$

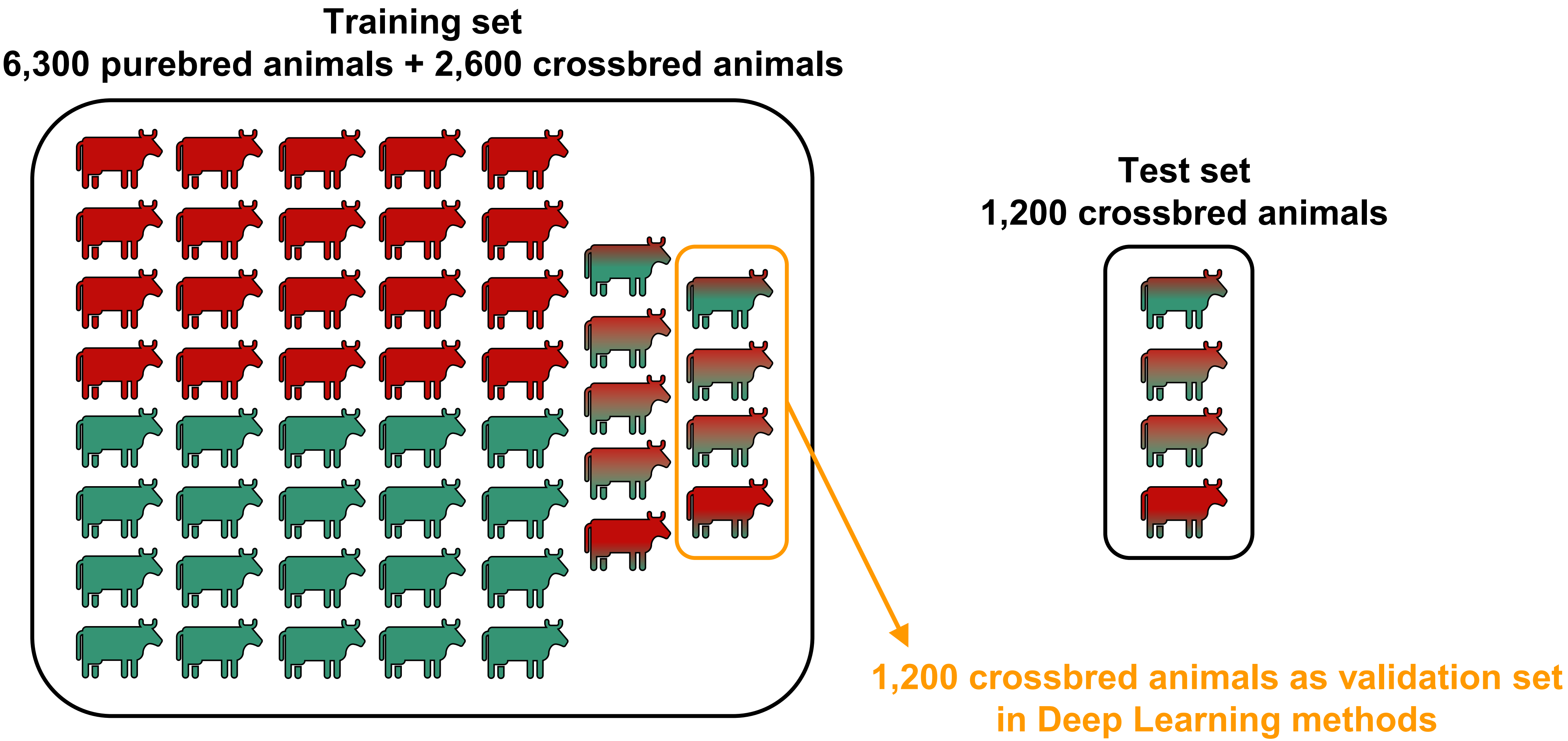
10 generations of rotational crossbreeding

Simulated Data



- Simulated 10,000 SNPs genotypes (500 QTL)
- Breed-of-origin of alleles (BOA) tracked for all crossbred animals.
- Heterosis introduced for crossbred animals at heterozygous loci in which alleles were from different breeds.
- Different portions of total genetic variance due to heterosis variance (heterosis ratio):
  - 0 (none), 0.1 (low), 0.2 (moderate), 0.3 (high).

Genomic Prediction



Training set
6,300 purebred animals + 2,600 crossbred animals

Test set
1,200 crossbred animals

1,200 crossbred animals as validation set in Deep Learning methods

Statistical Models

- **GBLUP** without considering BOA.
- **Admixture Model** considering BOA and SNP effects of purebred animals at each SNP.

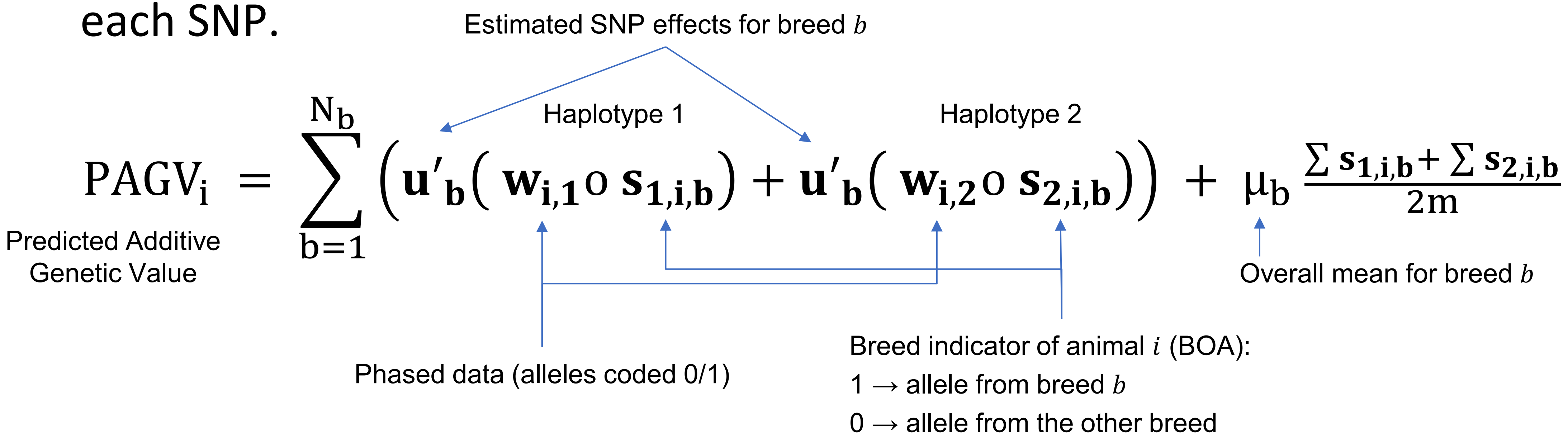
$$\mathrm{PAGV_i} = \sum_{b=1}^{N_b} \left(\mathbf{u'_b} \left(\mathbf{w_{i,1}} \circ \mathbf{s_{1,i,b}} \right) + \mathbf{u'_b} \left(\mathbf{w_{i,2}} \circ \mathbf{s_{2,i,b}} \right) \right) + \mu_b \frac{\sum \mathbf{s_{1,i,b}} + \sum \mathbf{s_{2,i,b}}}{2m}$$

- N_b = number of breeds (2)
- $\boldsymbol{u_b}$ = the vector of estimated SNP effects for breed b
- $\boldsymbol{w_{i,j}}$ = phased haplotype j of animal i (alleles coded 0/1)
- $\boldsymbol{s_{j,i,b}}$ = breed indicator vector for haplotype j of animal i (BOA)
  - 1 → allele from breed b
  - 0 → allele from the other breed
- μ_b = the overall mean for breed b
- $PAGV$ = predicted additive genetic value

Eiríksson JH, Karaman E, Su G, Christensen OF. Breed of origin of alleles and genomic predictions for crossbred dairy cows. Genet Sel Evol. 2021 Nov 6;53(1):84.

Statistical Models

- **GBLUP** without considering BOA.
- **Admixture Model** considering BOA and SNP effects of purebred animals at each SNP.



Estimated SNP effects for breed *b*

Haplotype 1

Haplotype 2

$$\mathrm{PAGV_i} = \sum_{b=1}^{N_b} \left(\mathbf{u'_b} \left(\mathbf{w_{i,1}} \circ \mathbf{s_{1,i,b}} \right) + \mathbf{u'_b} \left(\mathbf{w_{i,2}} \circ \mathbf{s_{2,i,b}} \right) \right) + \mu_b \frac{\sum \mathbf{s_{1,i,b}} + \sum \mathbf{s_{2,i,b}}}{2m}$$

Predicted Additive Genetic Value

Overall mean for breed *b*

Phased data (alleles coded 0/1)

Breed indicator of animal *i* (BOA):
1 → allele from breed *b*
0 → allele from the other breed

Eiríksson JH, Karaman E, Su G, Christensen OF. Breed of origin of alleles and genomic predictions for crossbred dairy cows. Genet Sel Evol. 2021 Nov 6;53(1):84.

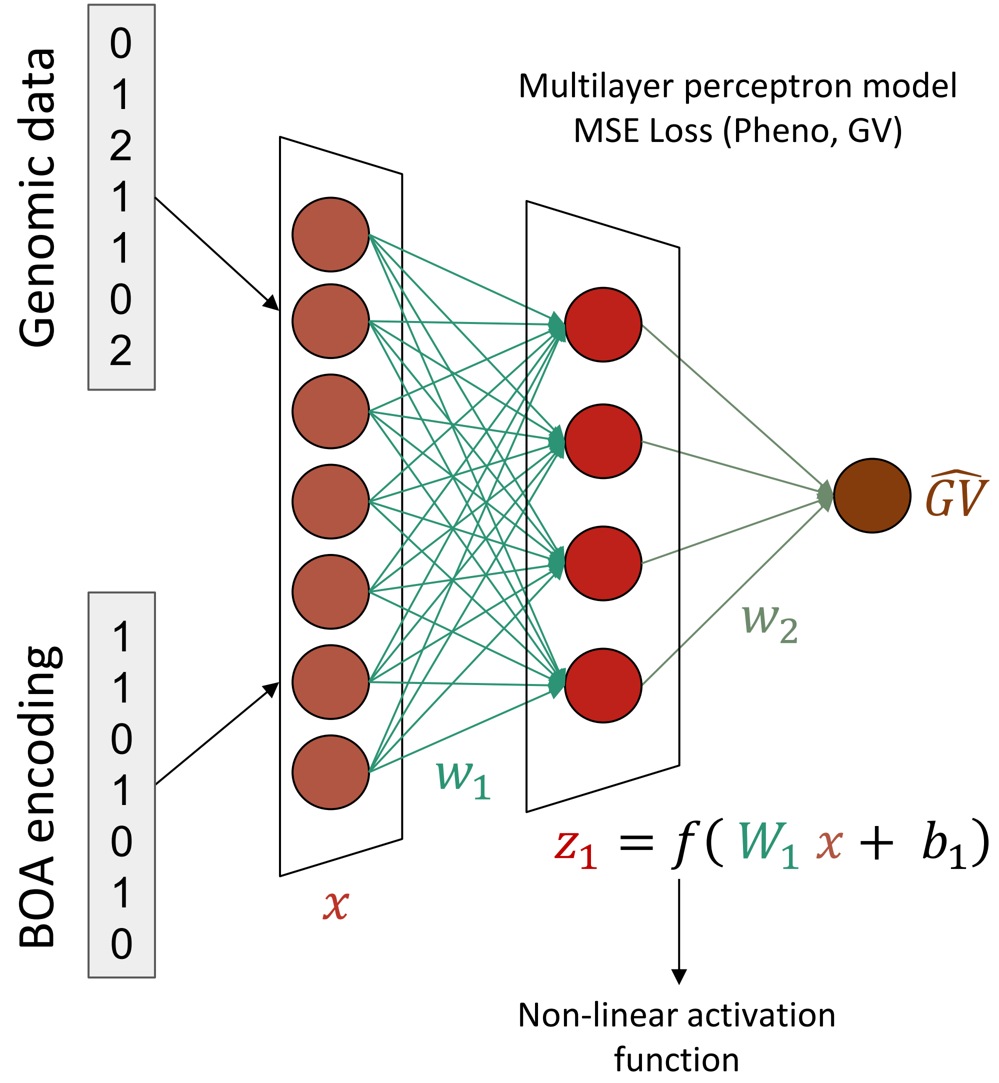
Deep Learning Models

- **DL:** implicit heterosis encoding from genomic data.

$$\widehat{\mathrm{PGV}} = \mathrm{W}_2\, \mathrm{z}_1 + \mathrm{b}_2$$

- **DL heterosis model**: explicit heterosis encoding using BOA
  - highlight heterosis information at each SNP. BOA encoding: different breed → 1 same breed → 0

$$\widehat{\mathrm{PHGV}} = \mathrm{W}_2\, \mathrm{z}_1 + \mathrm{b}_2$$



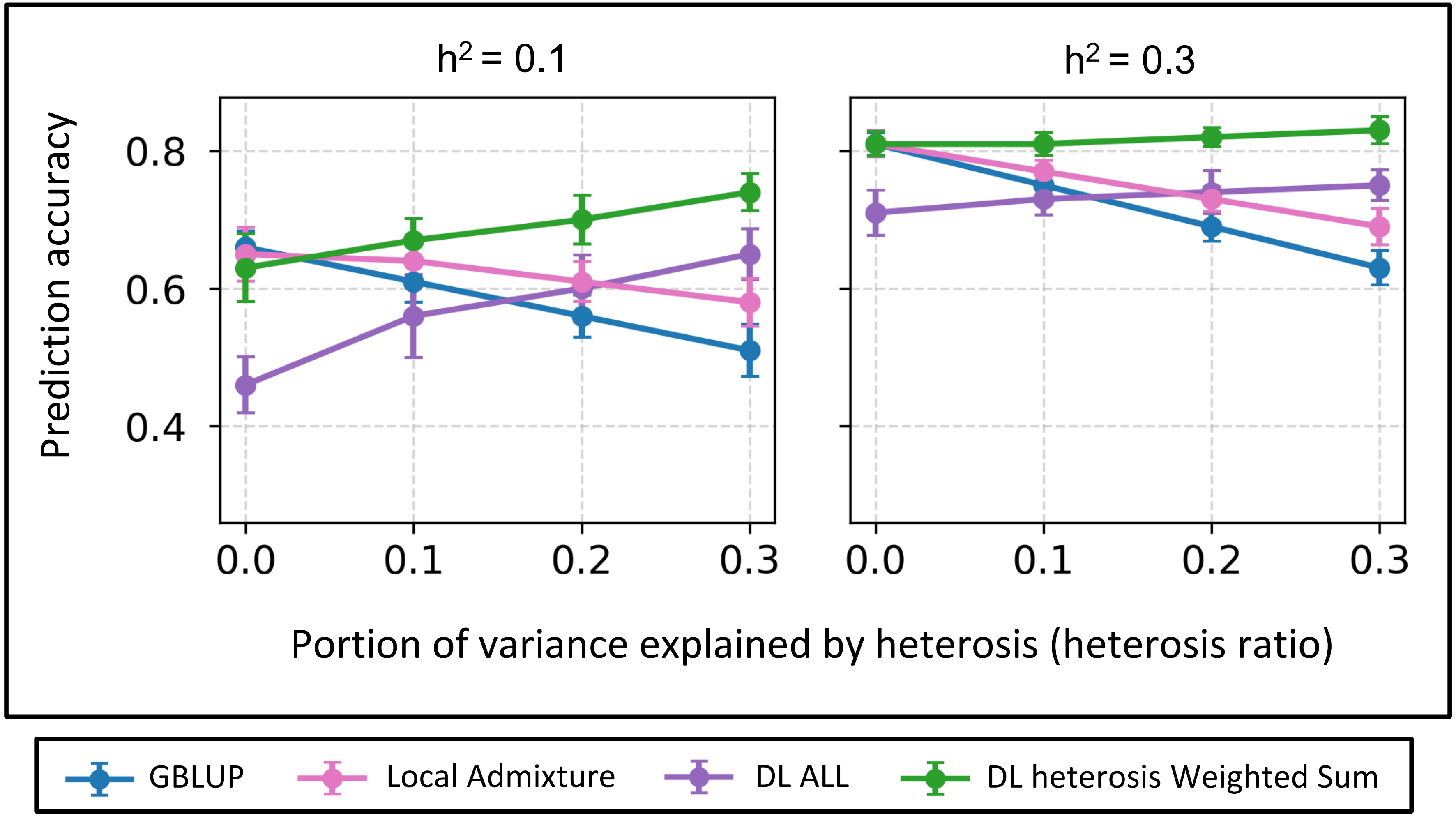
Weighted Sum Approach

- Predicts additive ($\widehat{\mathrm{PAGV}}$) and heterosis ($\widehat{\mathrm{PHGV}}$) genetic values separately.
- Using the phenotypes ($\boldsymbol{y}$), estimated weights for PAGV and PHGV using linear regression:

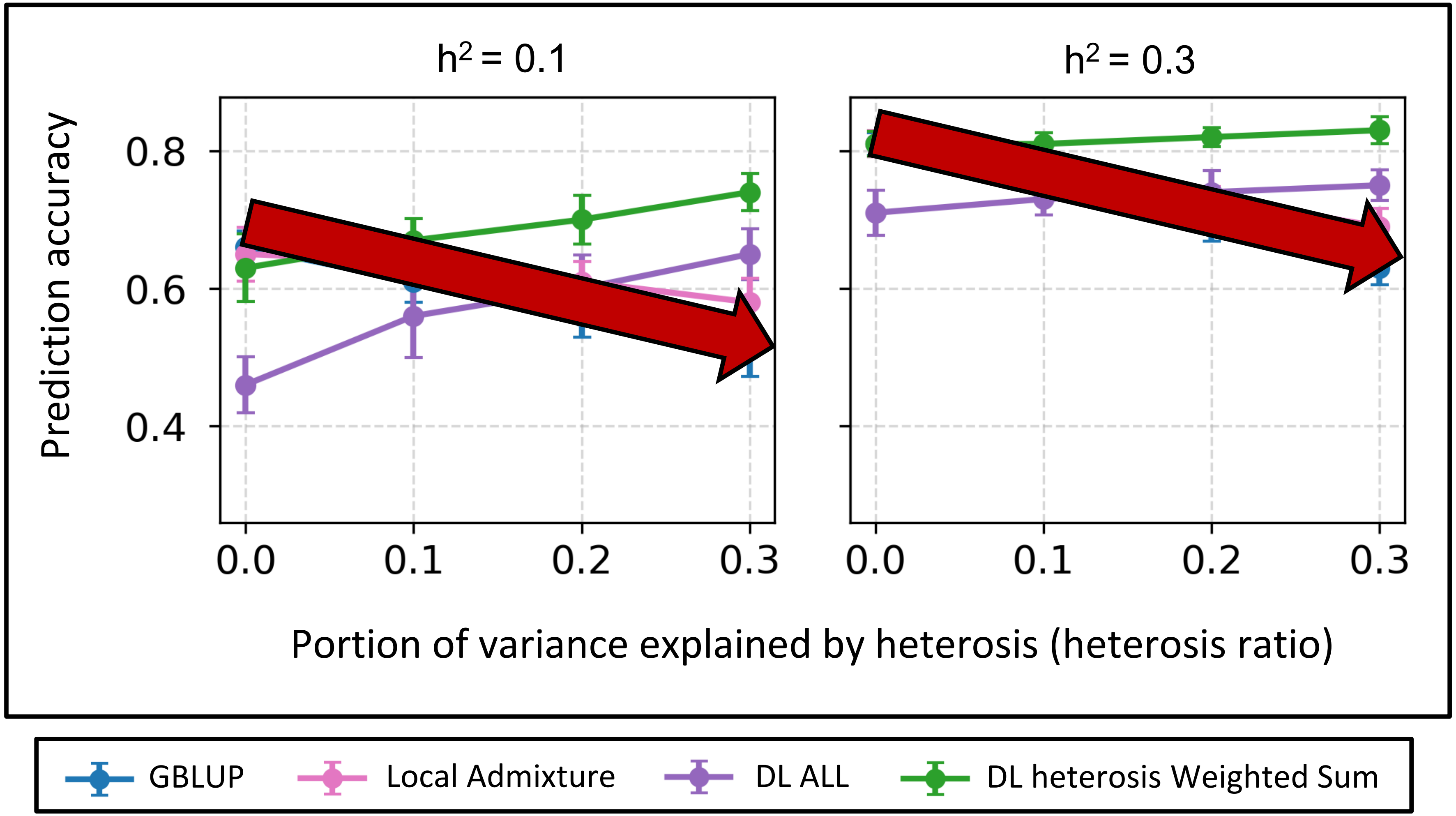
$$\mathbf{y} = \theta_1 \times \mathbf{PAGV} \; + \; \theta_2 \times \mathbf{PHGV} + \mathbf{e}$$

- Finally: $\widehat{\mathbf{PTGV}} = \hat{\theta}_1 \times \widehat{\mathbf{PAGV}} + \hat{\theta}_2 \times \widehat{\mathbf{PHGV}}$

Results


h² = 0.1
h² = 0.3
Prediction accuracy
0.8
0.6
0.4
0.0
0.1
0.2
0.3
Portion of variance explained by heterosis (heterosis ratio)
GBLUP
Local Admixture
DL ALL
DL heterosis Weighted Sum

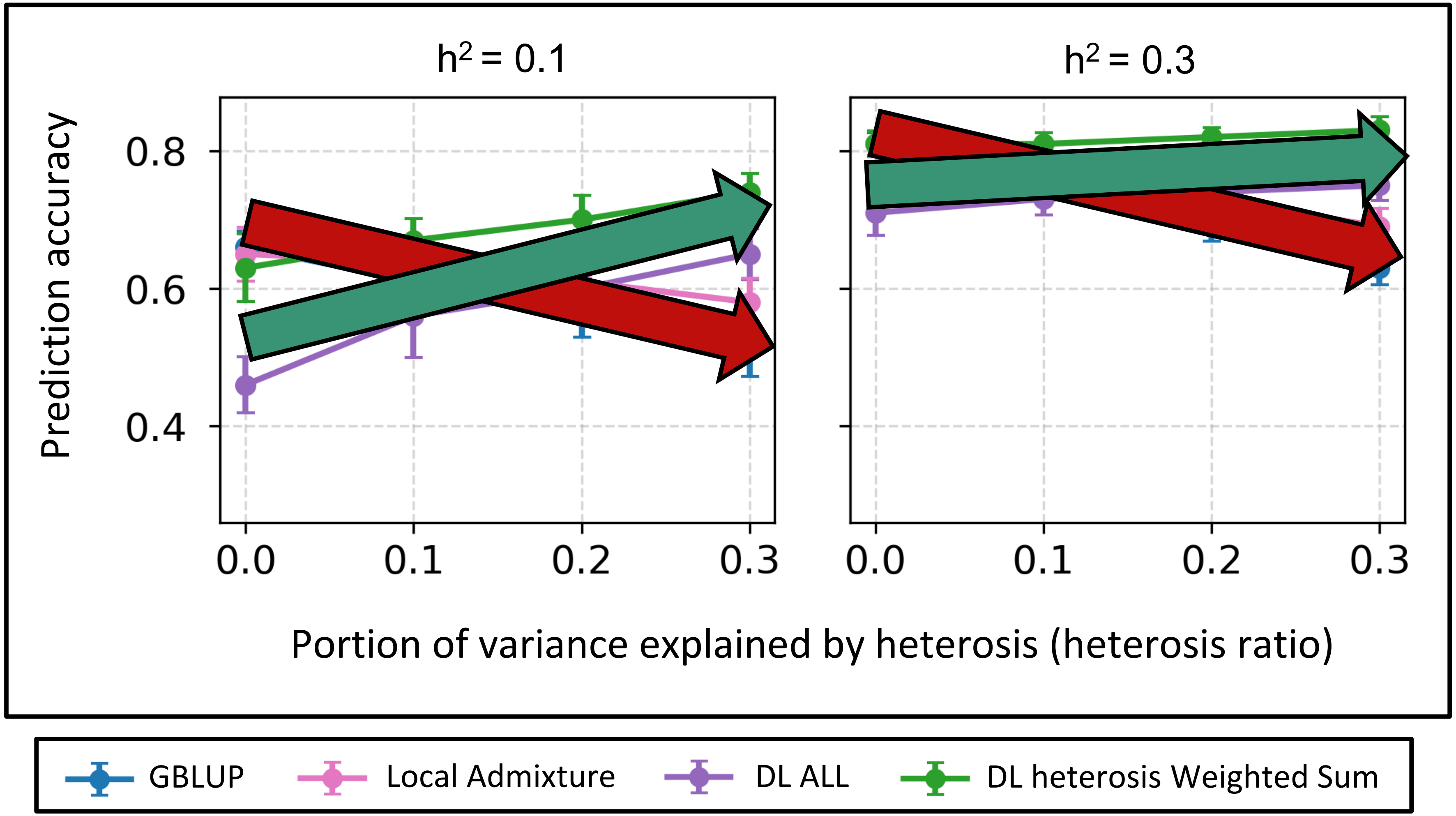
Results



Statistical models

Loss of prediction accuracy as heterosis increases

Results



Statistical models

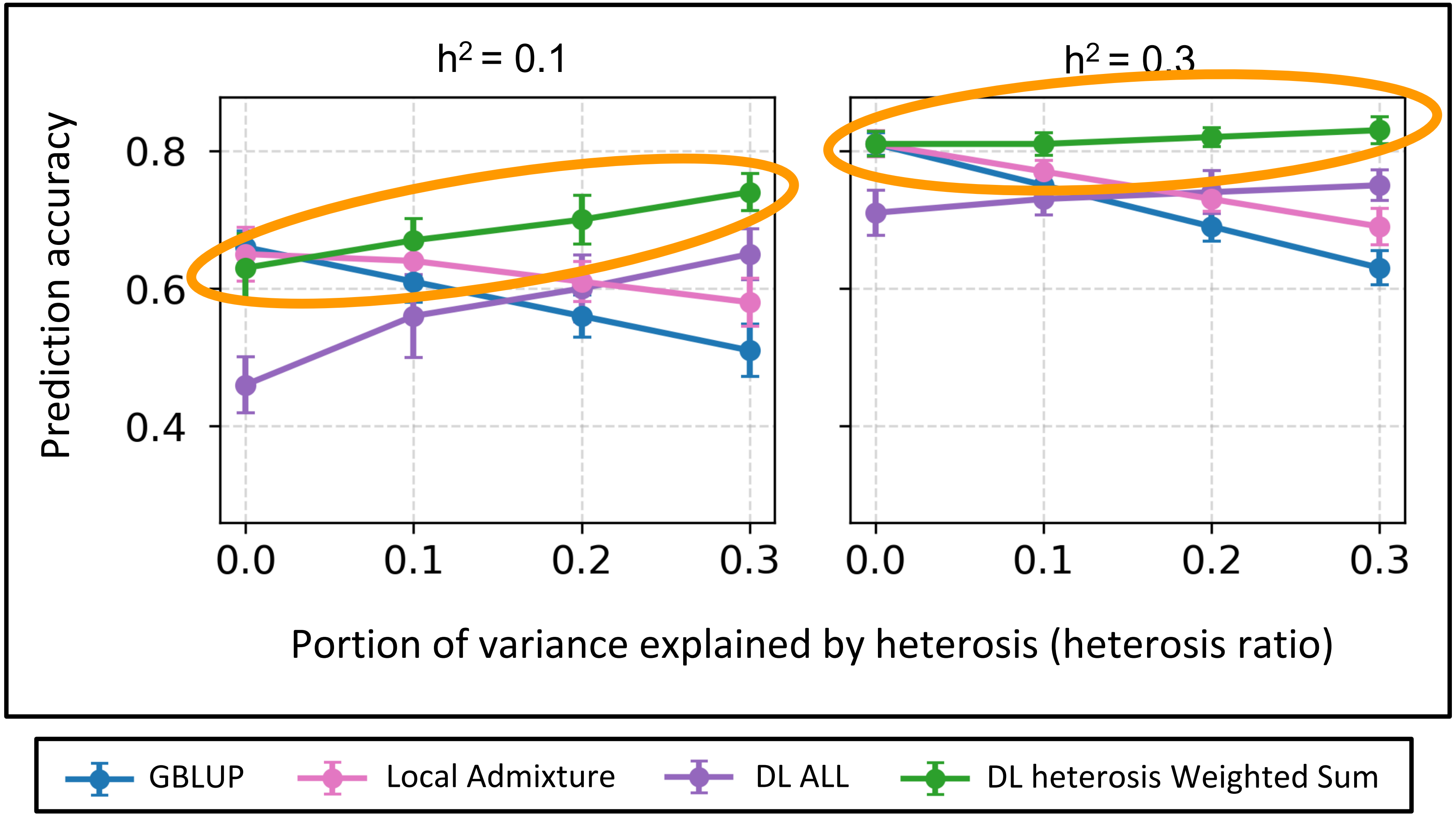
Loss of prediction accuracy as heterosis increases

Deep Learning

Improvement of prediction accuracy as heterosis increases

Greater gain for lower h^2

Results



Best predictions with DL heterosis weighted sum

Conclusions

- DL effectively captures heterosis effects for each individual.
- Incorporating local admixture information (BOA) further improves predictions.
- Weighted sum of additive and heterosis genetic values gives the best overall results.
- Our method offers a robust and flexible approach that can be applied for genomic selection in crossbreeding programs.

For more:

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

THANK YOU!