Deep Learning for Advancing Animal Breeding - A Study on Austrian Fleckvieh Cattle

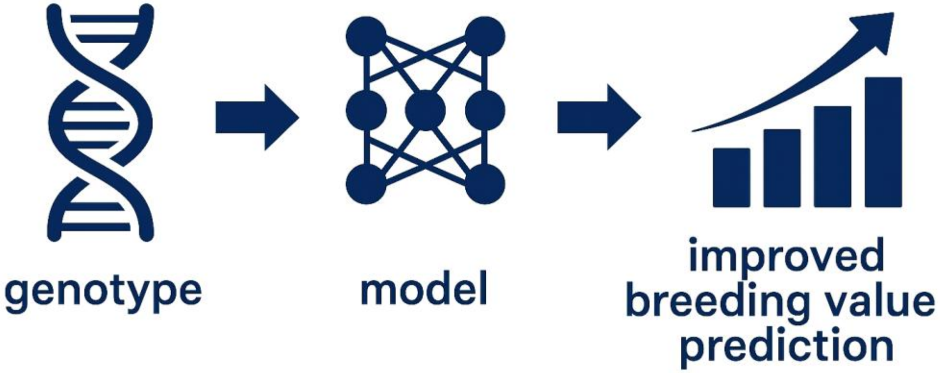
Jakob Ganitzer[1], Judith Himmelbauer[1], Hermann Schwarzenbacher[1], Maximilian Tschuchnigg[2]

[1]ZuchtData EDV-Dienstleistungen GmbH, Dresdner Str. 89, 1200 Vienna, Austria

[2]Salzburg University of Applied Sciences, Urstein Süd 15, 5412 Puch/Salzburg, Austria

Introduction and Motivation

- Genomic selection has revolutionized livestock breeding
- Breeding programs need accurate EBVs in young animals
- Hypothesis: Deep learning captures SNP-trait patterns better than GBLUP and can enhance genomic prediction accuracy

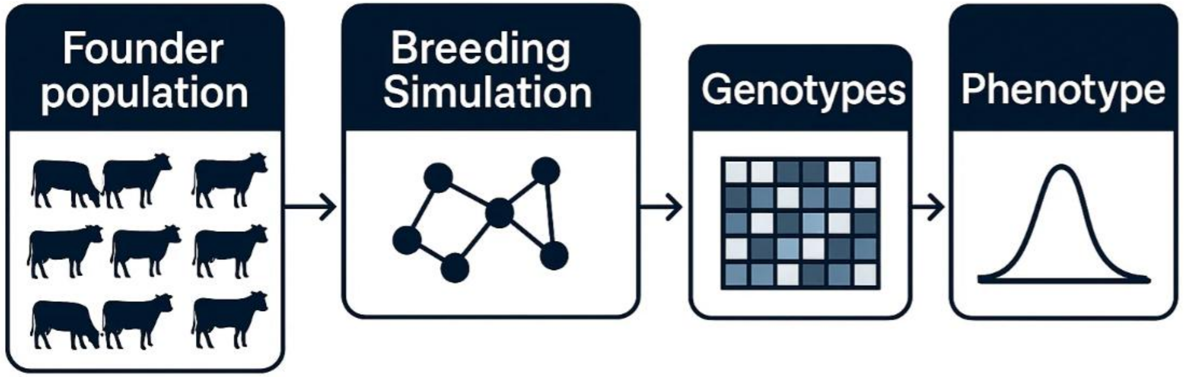


Research Objectives

- Benchmark deep learning architectures against Single step GBLUP (ssGBLUP) and XGBoost
- Develop large-scale simulated datasets
- Validation of model robustness across 5 simulated datasets
- Application of explainable AI (XAI) methods to uncover relevant SNP effects

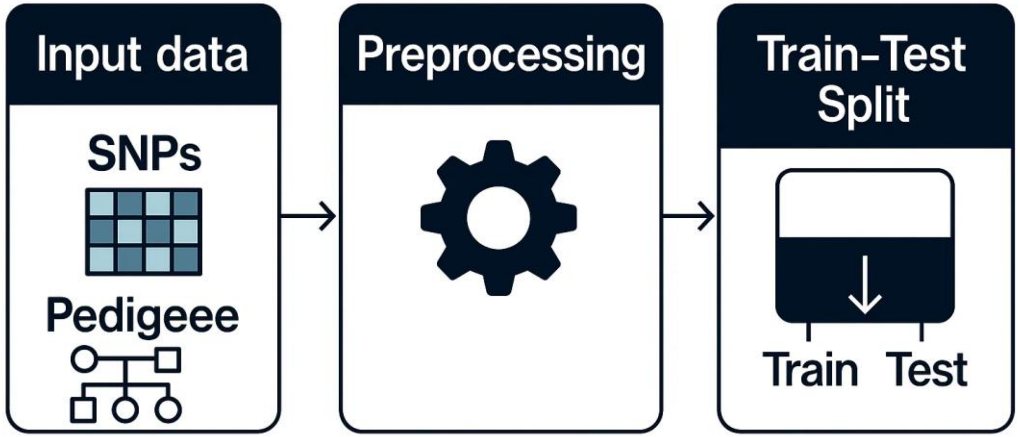
Data Simulation

- 5 simulated datasets
- Simulated breeding program with 30 overlapping generations, 1.3 million animals per dataset
- Genotyped individuals: 462,000 (Generation 20 to 30)
- 210,000 genotyped and phenotyped females in each training set
- SNP data: 49,800 markers across 30 chromosomes
- Target trait: sex-limited (female-only) trait with heritability of 0.40
- AlphaSimR used for breeding program simulation


Founder population → Breeding Simulation → Genotypes → Phenotype

Data Preprocessing

- Input:
  - SNP data: 49,800 markers per individual
  - Pedigree structure for graph model
- Preprocessing Steps:
  - Removal of monomorphic SNPs
  - Filtering by minor allele frequency (1%)
    - 1. Set: 26282 SNPs; 2. Set: 40638 SNPs; 3. Set: 42003 SNPs;
    - 4. Set: 41801 SNPs; 5. Set: 41825 SNPs;
  - Population Graph Generation
  - Encoding missing SNP data as Zero-Sequences
- Train-Test Split
  - Generation 20 – 29 for Training (210,000 females)
  - Generation 30 for Testing (42,000 animals)

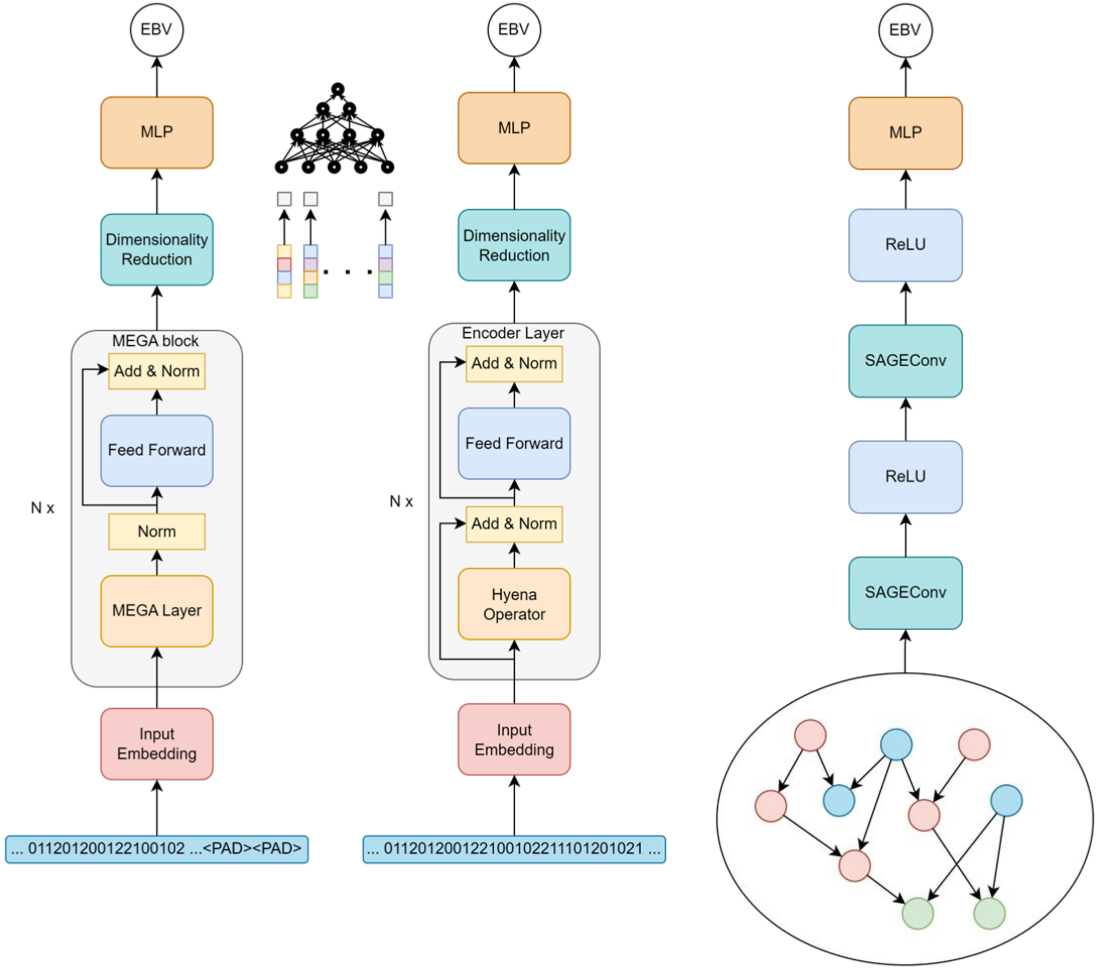


Model Architectures

- **Transformer-based Models**
  - Hyena-regressor (6 Layers)
  - MEGA-regressor (2 stacked MEGA blocks)

- **Graph-based Model**
  - GraphSAGE using pedigree structure

- **Baselines**
  - ssGBLUP (single-step GBLUP)
  - XGBoost



Model Architectures

Why Hyena and MEGA Instead of Standard Self-Attention?

Self-Attention in Transformers has quadratic complexity: limits sequence length, costly on 49K+ SNPs

Need for linear-scaling or sub-quadratic alternatives for long genomic sequences

- **Hyena Hierarchy [Poli et al., 2023]**
  - Sub-quadratic drop-in replacement for attention
  - Uses implicitly parametrized long convolutions and data-controlled gating
  - Allows longer context lengths and lower time complexity

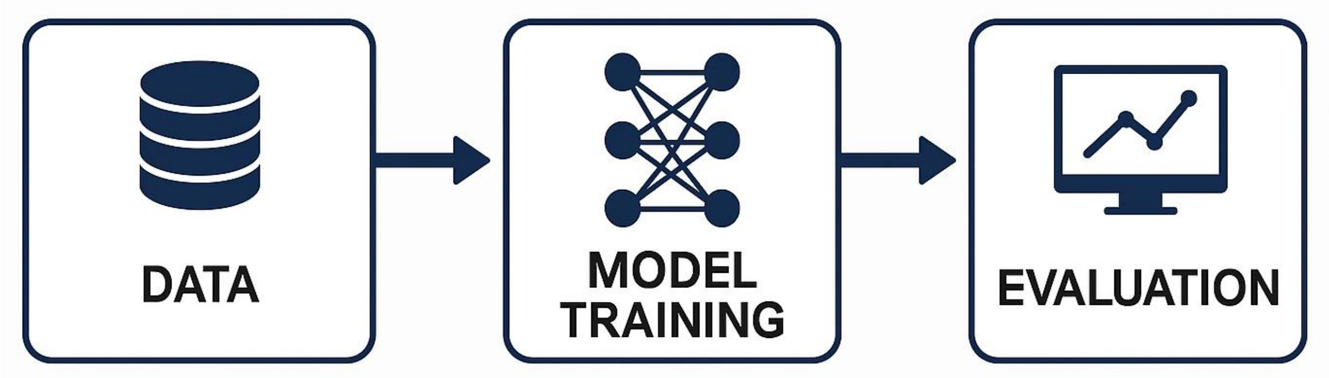
- **MEGA (Moving Average Equipped Gated Attention) [Ma et al., 2023]**
  - Drop-in replacement for regular multi-head attention.
  - Combines Exponential Moving Average (EMA) with Gated Attention
  - Linear time and space via chunking mechanism

M. Poli, S. Massaroli, E. Nguyen, et al., "Hyena hierarchy: Towards larger convolutional language models", in Proceedings of the 40th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 28 043–28 078.

X. Ma, C. Zhou, X. Kong, et al., "Mega: Moving average equipped gated attention", in Proceedings of the 11th International Conference on Learning Representations (ICLR-2023), 2023.

Training Methodology

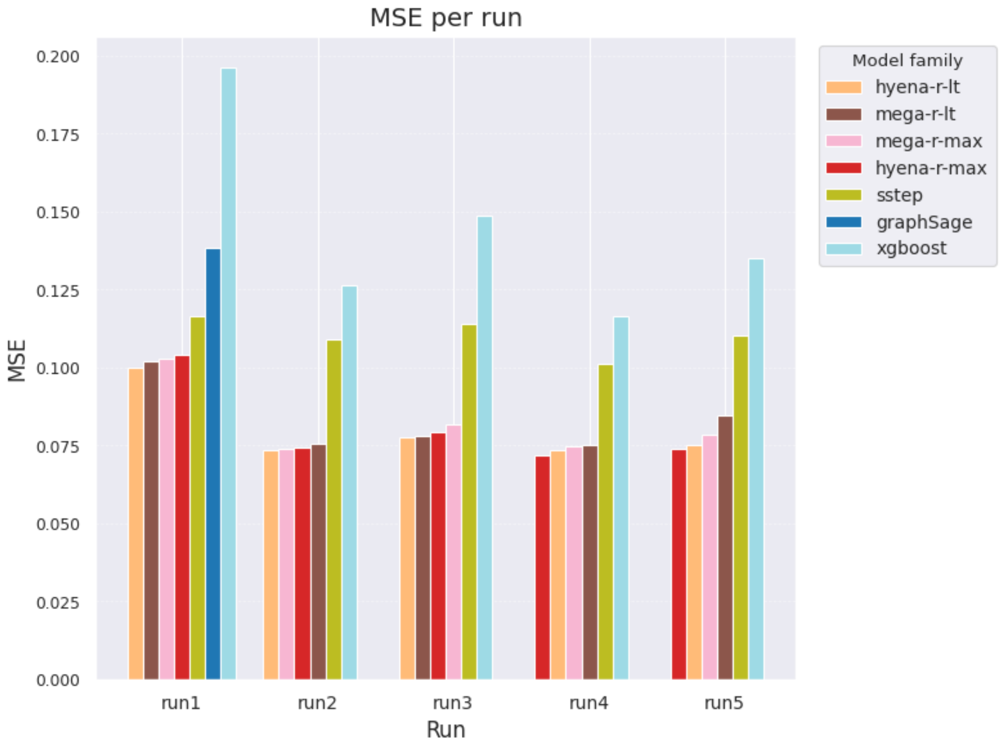
- Loss function: Mean Squared Error (MSE)
- Optimizer: AdamW, batch size = 64, epochs = 40
- Linear learning rate: initially 0.001 with linear scheduler
- Regularization: Lasso (L1 penalty)

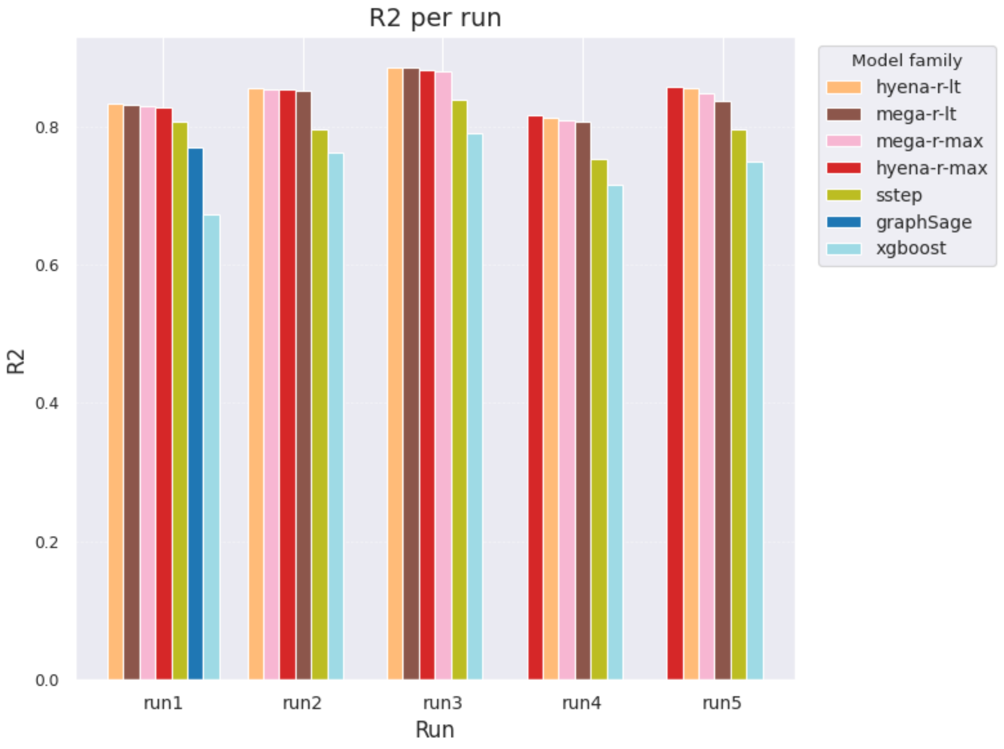


Model Performance Across Datasets

- Transformer-based models consistently outperformed benchmark models across all datasets

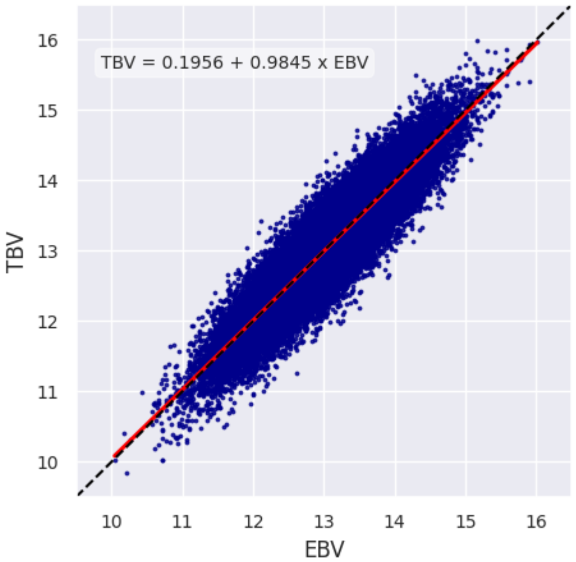
- Performance improves with more SNPs



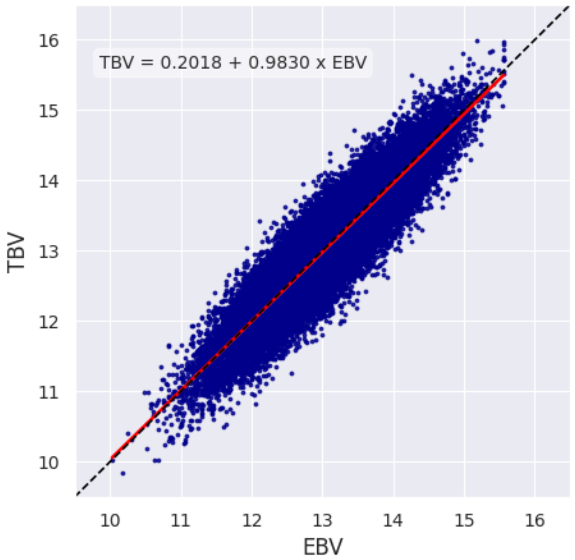


Model Evaluation

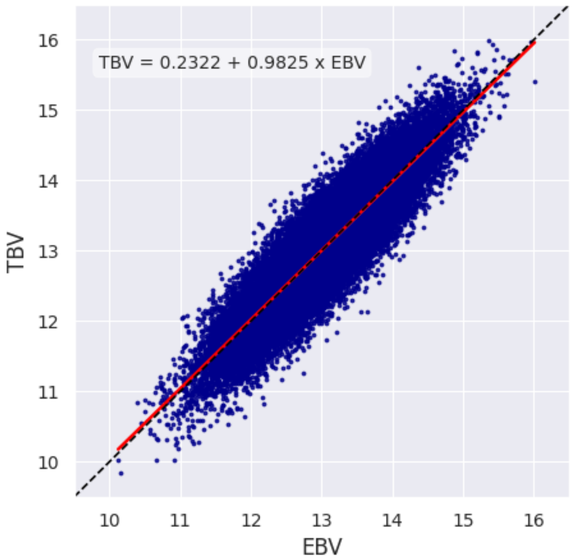
Hyena-R-LT_run1



Mega-R-LT_run1

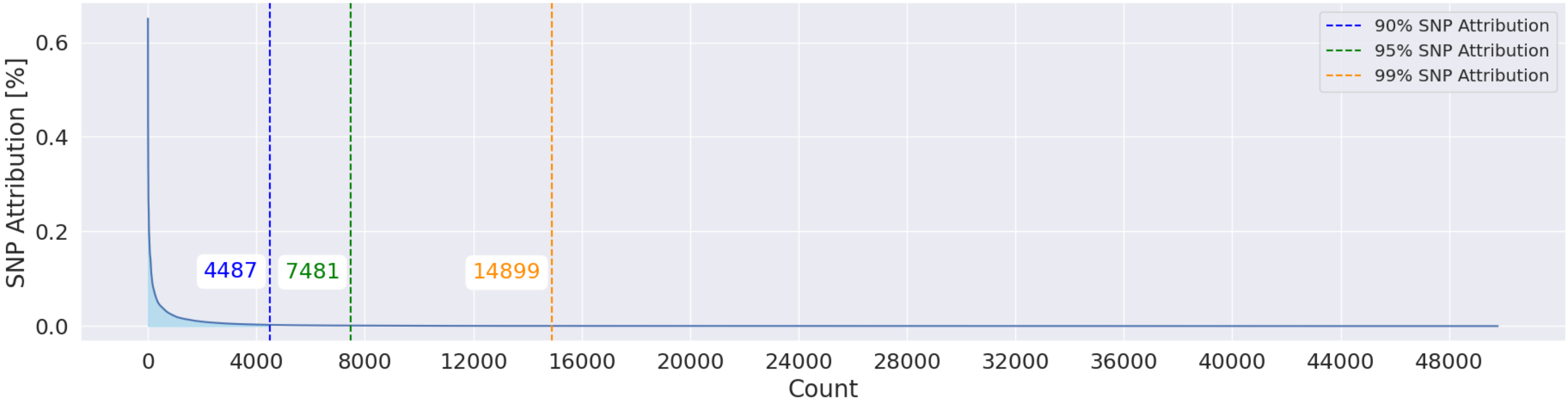


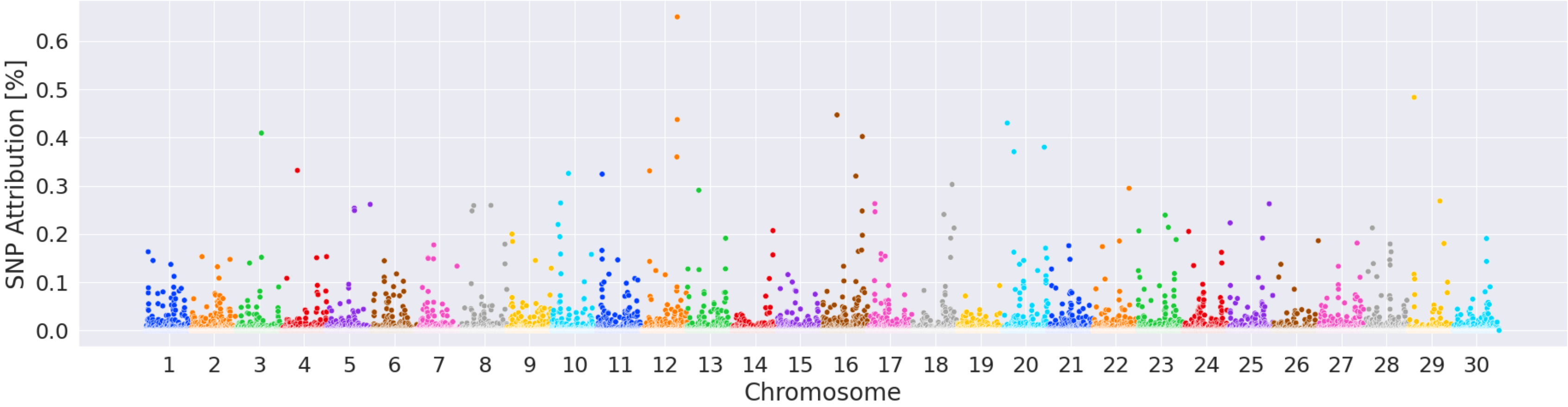
SSTEP_run1



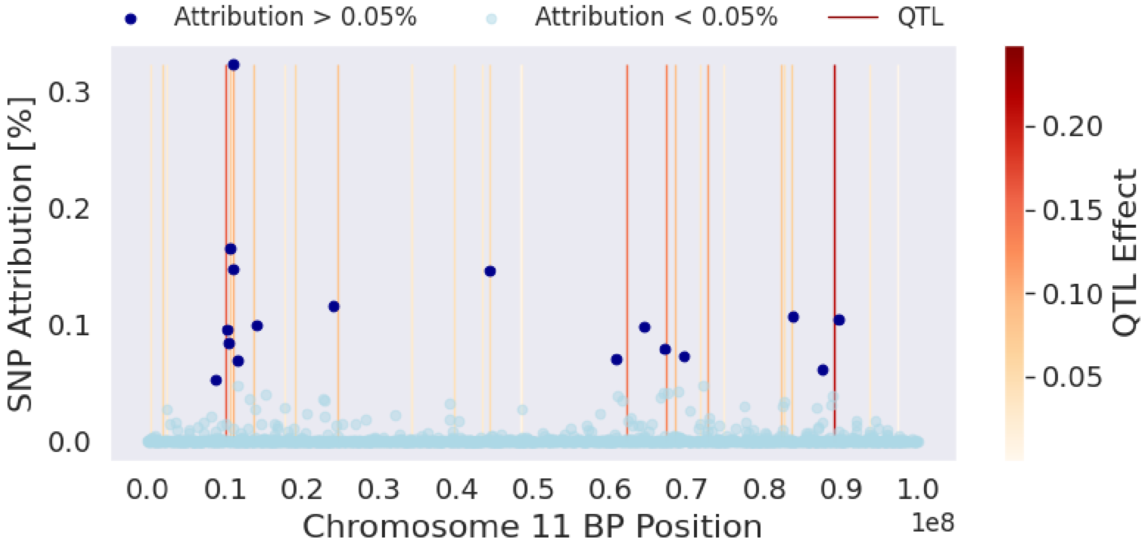
Explainable AI Results

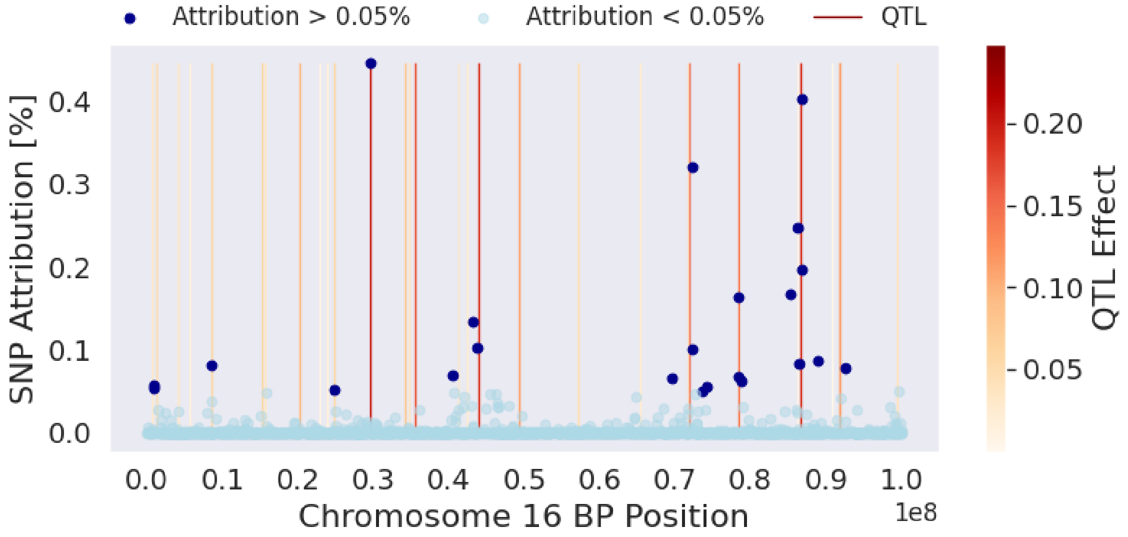
- Integrated gradients analysis: A small SNP subset explains ~90% of model attribution.
- High attribution SNPs aligned with known QTL locations.
- Demonstrates validity of the deep learning model.





Explainable AI Results





Conclusion

- Machine learning models are promising for genomic evaluation
- Limitation:
    - No environmental effects simulated
- Future research
    - Extension to real-world genomic datasets.
- Implication for routine application
    - Accuracy and unbiasedness
    - Computational cost
    - Acceptance by breeders

Thank You

RINDERZUCHT
AUSTRIA
Innovation

ZUCHT
DATA

Training Resource Requirements

- ssGBLUP (2 x AMD EPYC 9554 3.1GHz 64-core):
  - Peak RAM usage: 163 GB
  - CPU time: ~43h
- Heyna Model (NVIDIA RTX 6000 Ada Generation)
  - Train: ~19h 30 min
  - Test: 1 min 49 s
- MEGA Model (NVIDIA RTX 6000 Ada Generation)
  - Train: ~25h 30 min
  - Test: 2 min 42 s