A Machine Learning Framework for Concurrent Identification of Favorable, Deleterious, and Lethal Variants in Bovine Populations

N. Blair[1], D. Null[1], J. Cole[1,2,3,4], and E. Nicolazzi[1]

[1] Council on Dairy Cattle Breeding, Bowie, Maryland, 20716, USA

[2] Department of Animal Science, North Carolina State University, Raleigh, North Carolina, 27695-7621, USA

[3] Department of Animal Sciences, Donald Henry Barron Reproductive and Perinatal Biology Research Program, and the Genetics Institute, University of Florida, Gainesville, Florida, 32611-0910, USA

[4] Centre for Genetic Improvement of Livestock, University of Guelph, Guelph, Ontario, N1G 2W1, Canada

Corresponding author: john.cole@uscdcb.com

Abstract

Accurate genomic screening in livestock is essential for managing a complex spectrum of traits, including favorable production-related alleles, deleterious variants, and recessive lethal alleles. However, most organizations rely on legacy analytic pipelines that tend to be computationally slow and which suffer from high technical debt. These legacy systems are often difficult to maintain, lack modularity, and create significant bottlenecks as marker panels expand. We developed a high-throughput machine learning framework for identifying carrier status for many different haplotypes which will replace a legacy SAS program. By transitioning to a modular, Python-based architecture utilizing XGBoost and Polars, we streamlined the codebase, significantly reducing maintenance overhead and improving system extensibility. The new framework uses an automated feature-engineering pipeline to process genotype data, allowing for the concurrent classification of multiple favorable, sub-lethal deleterious, and recessive lethal variants in a single, reproducible environment. The ML-based approach demonstrated robust performance across all haplotypes tested, achieving an average concordance of 99.8% relative to the legacy system. In addition to matching the accuracy of the legacy system, the framework produced comparable runtimes on traditional CPU hardware and significantly reduced runtimes on modest GPU hardware. Crucially, the modernization of the codebase resulted in a 50% reduction in lines of code, enabling markedly faster updates as new genetic markers are discovered and integrated into the screening panel. Modernizing genomic screening with machine learning provides high-performance trait identification and a sustainable, maintainable engineering foundation with minimal drawbacks. By reducing technical debt and increasing throughput, this framework allows genomic labs to pivot quickly to new research requirements while maintaining rigorous operational standards for population health and genetic progress.

Key words: genomic screening, machine learning, XGBoost, haplotype carrier identification, dairy cattle, polled, recessive lethal haplotype

Introduction

Genomic screening is a routine component of dairy cattle breeding programs, enabling identification of animals which carry alleles of economic, health, or fertility relevance before mating and selection decisions are made. Variants of interest cover a spectrum:

favorable alleles such as polled (hornlessness), deleterious recessive conditions which reduce viability or performance, and recessive lethal haplotypes that cause embryonic loss when inherited in the homozygous state.

The Council on Dairy Cattle Breeding (CDCB; Bowie, Maryland, USA) currently assigns haplotype carrier status from an animal's genotype using a custom SAS-based software package called Hapcheck. Hapcheck, which was developed by scientists in the Animal Genomics and Improvement Laboratory of the United States Department of Agriculture (Beltsville, Maryland, USA), combines genotype and pedigree information to determine carrier status and has served reliably for years, but it has accrued substantial operational limitations. It is sensitive to segmentation in the input data; predictions for newly discovered haplotypes typically require additions to the SNP list and re-imputation of all genotypes; haplotypes are assigned serially, one after another; and numerous special cases have accumulated in the code that are increasingly difficult to validate as the number of genotypes grows.

To address these limitations we developed HapchIQ (pronounced "hap-sheek"), a modular, Python-based machine learning framework which reproduces Hapcheck's carrier calls, reduces technical debt, and scales gracefully as the number of haplotypes tracked grows. HapchIQ uses gradient-boosted decision trees (XGBoost) to classify carrier status for each haplotype, trains an independent model for each haplotype so that predictions can be produced in parallel, and returns a calibrated probability with each call. This paper describes the framework, evaluates it against Hapcheck and, where available, against direct gene-test results for favorable, deleterious, and recessive lethal variants in multiple breeds.

Materials and Methods

Genotype data

Evaluations were performed on CDCB genotype data across the Brown Swiss (BS), Holstein (HO), Jersey (JE), and Ayrshire (AY) breeds. Analyses reported here draw on populations ranging from roughly 20,000 Ayrshire animals to millions of Holstein genotype records, with training and evaluation runs spanning December 2024 through December 2025. The genotype records come from over 60 SNP chips of varying densities and missing SNP were imputed using findhap.f90 (VanRaden et al., 2015), our standard pedigree-based imputation process (Wiggans and Carrillo, 2022).

Target variants

The screening panel includes three categories of variant. Favorable: the polled locus in Brown Swiss (BHP), Holstein (HHP), and Jersey (JHP) (e.g., Aldersey et al., 2020). Recessive lethal haplotypes: Brown Swiss Haplotype 6 (BH6; Häfliger et al., 2021), Holstein Haplotypes 1, 2, and 5 (HH1, -2, -5; VanRaden et al., 2011; Cooper et al., 2013), and others. Deleterious recessive conditions: Holstein Cholesterol Deficiency (HCD; Kipp et al., 2015), Holstein Muscle Weakness (HMW; Dechow et al, 2022), Brown Swiss Weaver Syndrome (BHW; Leipold et al., 1973), Brown Swiss Spinal Dysmyelination (BHD; Hafner et al., 1993), Brown Swiss Haplotype 14 (BH14; Häfliger et al., 2021), Ayrshire Haplotype 2 (AH2; Null et al., 2017), Ayrshire Arthrogryposis Multiplex Congenita (AHC; Agerholm et al., 2016), and Jersey Neuropathy with Splayed Forelimbs (JNS; Al-Khudhair et al., 2022).

BH6 illustrates a key challenge and opportunity. Affected embryos die, reducing fertility; the causal allele lies on chromosome 2 (Ortega et al., 2022) and is not itself on the SNP list, but it can be detected through linkage disequilibrium because it falls between two SNPs that are tracked. Polled is challenging for a different reason: multiple mutations produce hornless animals, more than one variant segregates in some breeds, and many laboratory gene tests examine only the Friesian mutation.

Legacy system (Hapcheck)

Hapcheck assigns carrier status from genotype and pedigree information. In addition to non-carrier and carrier calls it emits ambiguous status codes ("3" and "4") when a call cannot be made confidently. Because it depends on the SNP list and on imputed genotypes, incorporating a newly discovered marker requires re-imputing all genotypes, and its serial, special-case-heavy design becomes progressively harder to maintain and validate as the database grows. Hapcheck is the primary baseline against which HapchIQ is compared.

Framework architecture

HapchIQ is implemented in Python 3.12 with a modular architecture that separates data ingestion, feature engineering, model inference, and reporting. Classification uses gradient-boosted decision trees (XGBoost 3.2.0), chosen because tree ensembles perform well on the large volumes of structured, tabular genotype data involved; one-dimensional convolutional neural networks were also evaluated and showed promise but were more resource-intensive. Large genotype flat files are parsed with Polars 1.36.1 for memory-efficient processing within reasonable RAM limits.

Automated feature engineering

An automated pipeline transforms raw genotype calls into model-ready features, encoding genotypes as allele dosages and selecting the markers relevant to each haplotype. Input is typically restricted to one or two chromosomes for efficiency, though this is not required. In practice, data engineering accounts for roughly 80% of the effort, driven by the difficulty of parsing very large data files efficiently.

Model training and class imbalance

A separate classifier is trained per haplotype using Hapcheck calls (and, where available, gene-test results) as labels. Training is sensitive to class imbalance: homozygous ("2") genotypes are very infrequent relative to non-carriers and heterozygotes for some haplotypes (e.g., HH2, BHW HCD), and the algorithm can discard them as likely genotyping or imputation errors. The data was segregated into train/validation/test partitions at 81%/9%/10%, respectively. Genotype rows exhibiting extreme class imbalances were discarded altogether.

Concurrent classification and certainty

Because a separate model is produced for each haplotype, predictions for all haplotypes can be generated in parallel within a single reproducible run. Each prediction carries a calibrated probability, providing a measure of certainty that removes the need for the ambiguous "3" and "4" status codes used by Hapcheck.

Validation and concordance

HapchIQ calls were compared with Hapcheck calls and, for BH6 and polled, with direct gene-test results, both overall and per haplotype. Concordance is reported as the percentage of animals receiving matching calls. Stability was also assessed by contrasting run-to-run variation (same

training set, later evaluation) against training-set-to-training-set variation (model retrained on newer data).

Hardware and runtime benchmarking

Runtimes were compared with Hapcheck on comparable inputs. GPU hardware produced a substantial reduction in training and inference time relative to CPU, at higher hardware cost. CPU comparisons were performed on an Intel Xeon Gold 6226 with 1TB of RAM; GPU hardware was a NVIDIA RTX 4000.

Results & Discussion

Recessive lethal detection: Brown Swiss Haplotype 6

On the official SNP list, HapchIQ reproduced gene-test BH6 status with 90.7% concordance, close to the 92.2% achieved by Hapcheck only after adding new DNA markers and re-imputing all Brown Swiss genotypes, and far above the 46.5% achieved by Hapcheck on the official SNP list without re-imputation (Table 1). HapchIQ thus recovers most of the BH6 signal from the existing SNP panel — via linkage disequilibrium with the untracked causal variant — without the costly re-imputation Hapcheck requires.

HapchIQ agreed with Hapcheck's BH6 calls for the large majority of animals, with disagreements confined to a small number of cases near the classification boundary (Table 2).

Table 1: Concordance of BH6 calls from Hapcheck and HapchIQ with gene-test results.

Program	SNP list	Concordance
Hapcheck	Official	46.5%
Hapcheck	Modified[1]	92.2%
HapchIQ	Official	90.7%

[1] Adding new DNA markers to the official SNP list requires that all genotypes be re-imputed; this was done only for Brown Swiss animals, for internal testing.

Table 2: Comparison of Hapcheck (official) and HapchIQ (research) results for Brown Swiss Haplotype 6.

		Hapchiq (research)	
		Non-carrier	Carrier
Hapcheck (official)	Non-carrier	87,932	32
	Carrier	51	4,656
	Homozygous	0	0

Predictions were stable over successive evaluation runs on a fixed training set, with only a handful of status changes; larger movement appeared only when the model was retrained on a substantially newer and larger training set (Table 3).

Favorable variant: polled in Brown Swiss, Holstein, and Jersey

Polled illustrates HapchIQ's largest advantage. On the official SNP list, HapchIQ reached 99.9%, 99.1%, and 99.3% concordance in Brown Swiss, Jersey, and Holstein, respectively, whereas Hapcheck performed well only in Holstein (95.4%) and poorly in Brown Swiss (85.3%) and Jersey (63.9%). After a correction was made to the algorithm used to infer polled carrier status from animal names ("Name fix"), all methods converged to near-perfect agreement (Table 4).

Table 3: Run-to-run variation versus training-set-to-training-set variation in BH6 calls[a].

TS1	ER1	TS2	ER2	Animals	Non-carrier to carrier	Carrier to non-carrier
2412	2412	—	2504	82,553	3	0
2412	2504	—	2508	83,829	1	1
2412	2508	2512	2512	90,062	78	46

[a]Training sets (TS) and evaluation runs (ER) are identified by year and month of evaluation, e.g., "2412" is the December 2024 run.

Table 4: Concordance of polled calls by breed, SNP list, and program (BS = Brown Swiss, JE = Jersey, HO = Holstein)[a].

Breed	SNP list	Program	Matches	Mismatches	Concordance
BS	Official	Hapcheck	17,679	3,047	85.3%
BS	Official	HapchIQ	20,697	29	99.9%
BS	Name fix	Hapcheck	20,661	65	99.7%
BS	Name fix	HapchIQ	20,723	3	100.0%
JE	Official	Hapcheck	17,559	9,928	63.9%
JE	Official	HapchIQ	27,231	256	99.1%
JE	Name fix	Hapcheck	26,784	671	97.6%
JE	Name fix	HapchIQ	27,470	15	99.9%
HO	Official	Hapcheck	55,943	2,714	95.4%
HO	Official	HapchIQ	58,248	409	99.3%
HO	Name fix	Hapcheck	58,079	577	99.0%
HO	Name fix	HapchIQ	58,590	66	99.9%

[a]"Name fix" indicates improvements made to the use of information from animal names to identify polled status (e.g., "P" to indicate heterozygous polled and "PP" to indicate homozygous polled.

HapchIQ therefore extracts the polled signal from the existing panel where Hapcheck cannot, without requiring panel updates or re-imputation.

Table 5: Comparison of Hapcheck and HapchIQ results for Holstein Haplotype 1 (HH1), Holstein Haplotype 2 (HH2), Holstein Haplotype 5 (HH5), Holstein Cholesterol Deficiency (HCD), and Holstein Muscle Weakness (HMW) using August 2025 data.

Haplotype	Mismatches / genotypes	Concordance
HH1	1,233 / 9,088,709	99.99%
HH2	5,963 / 9,089,160	99.93%
HH5	2,170 / 9,088,709	99.98%
HCDy	44,410 / 8,914,901	99.50%
HMW	67,024 / 8,173,668	99.18%

Table 6: Comparison of Hapcheck and HapchIQ results for other Ayrshire Haplotype 2 (AH2), Ayrshire Arthrogryposis Multiplex Congenita (AHC), Brown Swiss Weaver Syndrome (BHW), Brown Swiss Spinal Dysmyelination (BHD), Brown Swiss Haplotype 14 (BH14), and Jersey Neuropathy with Splayed Forelimbs (JNS) , using August 2025 data.

Haplotype	Mismatches / genotypes	Concordance
AH2	109 / 20,448	99.47%
AHC	48 / 20,448	99.77%
BHW	101 / 90,081	99.89%
BHD	136 / 90,081	99.85%
BH14	104 / 90,081	99.88%
JNS	479 / 930,718	99.95%

Other haplotypes across breeds

Across additional Holstein haplotypes, HapchIQ agreed with Hapcheck 99.18–99.99% of the time (Table 5). Holstein Muscle Weakness was the lowest at 99.18%, consistent with the difficulty of haplotypes with incomplete penetrance. Concordance was similarly high for haplotypes in other breeds, ranging from 99.47 to 99.95% (Table 6).

Codebase, runtime, and maintainability

Modernization reduced the codebase by approximately 50% in lines of code relative to Hapcheck, and the per-haplotype, parallelizable design markedly lowered the effort to add new markers. GPU hardware substantially reduced runtimes relative to CPU.

Discussion

HapchIQ matches Hapcheck on well-behaved haplotypes and clearly outperforms it in situations where Hapcheck is limited by the SNP list. For polled in Brown Swiss and Jersey, and for BH6, HapchIQ recovers most of the signal from the existing panel through linkage disequilibrium, avoiding the SNP-list

updates and full re-imputation Hapcheck needs. This is the central practical advantage: new or hard-to-genotype variants can be screened without the costly re-imputation cycle.

The engineering benefits reinforce the analytical ones. A modular, Python-based design with independent per-haplotype models supports parallel execution and rapid extension; calibrated probabilities supply a certainty measure that eliminates the ambiguous “3” and “4” calls; and Polars-based ingestion handles large flat genotype files within practical RAM limits. Together these reduce technical debt and shorten the path from a newly discovered marker to a deployed screen.

Several limitations remain. Training is sensitive to class imbalance, and rare homozygous genotypes for some haplotypes (HH2, BHW, HCD) can be discarded as likely genotyping or imputation errors. Training is computationally intensive — GPUs help considerably but add cost — and the rules encoded in the decision trees can be difficult to interpret, a consideration for a regulated screening context. Haplotypes with incomplete penetrance, such as HMW (99.18%), are harder to predict. Because HapchIQ is trained largely on Hapcheck calls, it can also inherit systematic errors from that baseline where gene-test labels are unavailable.

Ongoing work addresses these issues through multi-chromosome inputs and hyperparameter tuning, and research is underway to determine whether pedigree information — used by Hapcheck but not yet by HapchIQ — can be incorporated to further improve predictions.

Conclusions

HapchIQ, a modular XGBoost-based framework using Polars for efficient data handling, predicts haplotype carrier status as well as Hapcheck in some cases and better in others, notably for polled in Brown Swiss and Jersey and for BH6 on the official SNP list. A separate model is trained for each haplotype, so predictions run in parallel, and each prediction carries a measure of certainty that eliminates the need for ambiguous “3” and “4” calls. By recovering signal from the existing SNP panel without re-imputation, halving the codebase, and enabling GPU acceleration, HapchIQ delivers high-performance trait identification on a sustainable, maintainable engineering foundation, letting genomic laboratories integrate new markers quickly while maintaining rigorous standards for population health and genetic progress. Research is underway to incorporate pedigree information into HapchIQ predictions.

Acknowledgments

The authors thank the U.S. Dairy Producers, member sectors and collaborators for contributions to this project. The support and input from an internal CDCB team were essential throughout the review process.

The mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by CDCB. CDCB is an equal opportunity provider and employer.

References

Agerholm, J.S., McEvoy, F.J., Menzi, F., Jagannathan, V., and Drögemüller, C. 2016. A CHRNB1 frameshift mutation is associated with familial arthrogryposis multiplex congenita in Red dairy cattle. *BMC Genomics* 17, 479. https://doi.org/10.1186/s12864-016-2832-x.

Al-Khudhair, A., Null, D.J., Cole, J.B., Wolfe, C.W., Steffen, D.J., and VanRaden, P.M. 2022. Inheritance of a mutation causing neuropathy with splayed forelimbs in Jersey cattle. *J. Dairy Sci.* 105, 1338–1345. https://doi.org/10.3168/jds.2021-20600.

Aldersey, J.E., Sonstegard, T.S., Williams, J.L., and Bottema, C.D.K. 2020. Understanding the effects of the bovine POLLED variants. *Animal Genet.* 51, 166–176. https://doi.org/10.1111/age.12915.

Cooper, T.A., Wiggans, G.R., VanRaden, P.M., Hutchison, J.L., Cole, J.B., and Null, D.J. 2013. Genomic evaluation of Ayrshire dairy cattle and new haplotypes affecting fertility and stillbirth in Holstein, Brown Swiss and Ayrshire breeds. *J. Dairy Sci.* 96(E-Suppl. 1):74(Poster T206).

Dechow, C.D., Frye, E., and Maunsell, F.P.. 2022. Identification of a putative haplotype associated with recumbency in Holstein calves. *JDS Comm.* 3, 412–415. https://doi.org/10.3168/jdsc.2022-0224.

Häfliger, I.M., Seefried, F.R., Spengeler, M., and Drögemüller, C. 2021. Mining massive genomic data of two Swiss Braunvieh cattle populations reveals six novel candidate variants that impair reproductive success. *Genet. Sel. Evol.* 53, 95. https://doi.org/10.1186/s12711-021-00686-3.

Hafner, A., Dahme, E., Obermaier, G., Schmidt, P., and Dirksen, G. 1993. Spinal dysmyelination in new-born Brown Swiss x Braunvieh calves. *Zentralbl Veterinarmed B* 40, 413-22. https://doi.org/10.1111/j.1439-0450.1993.tb00158.x.

Kipp, S., Segelke, D., Schierenbeck, S., Reinhardt, F., Reents, R., Wurmser, C., Pausch, H., Fries, R., Thaller, G., Tetens, J., Pott, J., Piechotta, M., and Grünberg, W. 2015. A new Holstein haplotype affecting calf survival. *Interbull Bulletin* 49, 49-53.

Leipold, H.W., Blaugh, B., Huston, K., Edgerly, CG., and Hibbs, CM. 1973. Weaver syndrome in Brown Swiss cattle: clinical signs and pathology. *Vet. Med. Small Anim. Clin.* 68, 645-647.

Null, D.J., Hutchison, J.L., Bickhart, D.M., VanRaden, P.M., and Cole, J.B. 2017. Discovery of a haplotype affecting fertility in Ayrshire dairy cattle and identification of a putative causal variant. *J. Dairy Sci.* 100(Suppl. 2):199(abstr. 205).

Ortega, M.S., Bickhart, D.M., Lockhart, K.N., Null, D.J., Hutchison, J.L., McClure, J.C., and Cole, J.B. 2022. Truncation of IFT80 causes early embryonic loss in Holstein cattle associated with Holstein haplotype 2. *J. Dairy Sci.* 105, 9001–9011. https://doi.org/10.3168/jds.2022-21853.

VanRaden, P.M., Olson, K.M., Null, D.J., and Hutchison, J.L. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. *J. Dairy Sci.* 94, 6153–6161. https://doi.org/10.3168/jds.2011-4624.

VanRaden, P.M., C. Sun, and J.R. O'Connell. 2015. Fast imputation using medium or low-coverage sequence data. *BMC Genetics* 16:82. https://doi.org/10.1186/s12863-015-0243-7.

Wiggans, G.R., and J.A. Carrillo. 2022. Genomic selection in United States dairy cattle. *Front. Genet.* 13:994466. https://doi.org/10.3389/fgene.2022.994466