J. Dairy Sci. 106
https://doi.org/10.3168/jds.2022-22119

Predicting milk protein fraction using infrared spectroscopy and a gradient boosting machine for breeding purposes in Holstein cattle

L. F. Macedo Mota,[1] V. Bisutti,[1] A. Vanzin,[1] S. Pegolo,[1]* A. Toscano,[1] S. Schiavon,[1] F. Tagliapietra,[1] L. Gallo,[1] P. Ajmone Marsan,[2] and A. Cecchinato[1]

[1]Department of Agronomy, Food, Natural Resources, Animals and Environment (DAFNAE), University of Padova, Viale dell' Università 16, 35020 Legnaro, Italy
[2]Department of Animal Science, Food and Nutrition (DIANA) and Research Center Romeo and Enrica Invernizzi for Sustainable Dairy Production (CREI), Faculty of Agricultural, Food and Environmental Sciences, Università Cattolica del Sacro Cuore, 29122 Piacenza, Italy

ABSTRACT

In recent years, increasing attention has been focused on the genetic evaluation of protein fractions in cow milk with the aim of improving milk quality and technological characteristics. In this context, advances in high-throughput phenotyping by Fourier-transform infrared (FTIR) spectroscopy offer the opportunity for large-scale, efficient measurement of novel traits that can be exploited in breeding programs as indicator traits. We took milk samples from 2,558 Holstein cows belonging to 38 herds in northern Italy, operating under different production systems. Fourier-transform infrared spectra were collected on the same day as milk sampling and stored for subsequent analysis. Two sets of data (i.e., phenotypes and FTIR spectra) collected in 2 different years (2013 and 2019–2020) were compiled. The following traits were assessed using HPLC: true protein, major casein fractions [α_{S1}-casein (CN), α_{S2}-CN, β-CN, κ-CN, and glycosylated-κ-CN], and major whey proteins (β-lactoglobulin and α-lactalbumin), all of which were measured both in grams per liter (g/L) and proportion of total nitrogen (% N). The FTIR predictions were calculated using the gradient boosting machine technique and tested by 3 different cross-validation (CRV) methods. We used the following CRV scenarios: (1) random 10-fold, which randomly split the whole into 10-folds of equal size (9-folds for training and 1-fold for validation); (2) herd/date-out CRV, which assigned 80% of herd/date as the training set with independence of 20% of herd/date assigned as the validation set; (3) forward/backward CRV, which split the data set in training and validation set according with the year of milk sampling (FTIR and gold standard data assessed in 2013 or 2019–2020) using the "old" and "new" databases for training and validation, and vice-versa with independence among them; (4) the CRV for genetic parameters (CRV-gen), where animals without pedigree as assigned as a fixed training population and animals with pedigree information was split in 5-folds, in which 1-fold was assigned to the fixed training population, and 4-folds were assigned to the validation set (independent from the training set). The results (i.e., measures and predictions) of CRV-gen were used to infer the genetic parameters for gold standard laboratory measurements (i.e., proteins assessed with HPLC) and FTIR-based predictions considering the CRV-gen scenario from a bi-trait animal model using single-step genomic BLUP. We found that the prediction accuracies of the gradient boosting machine equations differed according to the way in which the proteins were expressed, achieving higher accuracy when expressed in g/L than when expressed as % N in all CRV scenarios. Concerning the reproducibility of the equations over the different years, the results showed no relevant differences in predictive ability between using "old" data as the training set and "new" data as the validation set and vice-versa. Comparing the additive genetic variance estimates for milk protein fractions between the FTIR predicted and HPLC measures, we found reductions of −19.7% for milk protein fractions expressed in g/L, and −21.19% expressed as % N. Although we found reductions in the heritability estimates, they were small, with values ranging from −1.9 to −7.25% for g/L, and −1.6 to −7.9% for % N. The posterior distributions of the additive genetic correlations (r_a) between the FTIR predictions and the laboratory measurements were generally high (>0.8), even when the milk protein fractions were expressed as % N. Our results show the potential of using FTIR predictions in breeding programs as indicator traits for the selection of animals to enhance milk protein fraction contents. We expect acceptable responses to selection due to the high genetic correlations between HPLC measurements and FTIR predictions.

Received March 25, 2022.
Accepted October 10, 2022.
*Corresponding author: sara.pegolo@unipd.it

Key words: cross-validation strategies, genetic parameters, milk protein fraction prediction, prediction accuracy

INTRODUCTION

Milk protein composition is a key of milk that has an important biological effect on its quality and technological traits, such as milk coagulation and cheesemaking aptitude (Silva and Malcata, 2005; Amalfitano et al., 2019). The major caseins (α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN) and whey proteins (α-LA and β-LG) represent approximately 90% of the milk protein content (Walstra, 1999) and are related to important sources of active peptides with different physiological and nutritional effects (Silva and Malcata, 2005). Milk protein fractions have, therefore, been identified as a selection criterion in dairy cattle to improve milk technological traits (Amalfitano et al., 2019). However, assessing milk protein fractions at the individual level in dairy cattle is difficult and time-consuming due to high phenotyping costs, which are a deterrent to large-scale quantification.

From a technological point of view, advances in milk Fourier-transform infrared spectroscopy (**FTIR**) for high-throughput phenotyping of dairy cattle allows the assessment of complex traits that are difficult and expensive to measure on a large scale. Milk FTIR spectra have been used for direct prediction of different phenotypes in milk, such as fat (Rutten et al., 2010), fatty acids (Soyeurt et al., 2011), protein fractions (Baba et al., 2021), lactoferrin (Soyeurt et al., 2020), minerals (Zaalberg et al., 2021), and highly detailed milk composition traits (Grelet et al., 2016; Bonfatti et al., 2017b; Mota et al., 2021b). Recently, increasing attention has been directed to the potential usefulness of milk FTIR spectroscopy for large-scale phenotyping, as the technique is cost-effective, fast, nondestructive, and able to phenotype a large number of animals (Baba et al., 2021). In this context, high-throughput measurement by milk FTIR can be considered a suitable method for application in dairy breeding programs (Rutten et al., 2011; Bittante et al., 2013; Cecchinato et al., 2013, 2020).

The main concern regarding the use of FTIR spectroscopy to predict milk composition is its predictive ability. However, appropriate statistical methods using rank-reduction and variable selection can be used to identify the relevant FTIR wavelengths and capture the nonlinear relationships between predictor variables and target traits (Soyeurt et al., 2020; Mota et al., 2021b), which can lead to improvements in predictive ability. These factors are, therefore, important in dairy cattle selection (Karoui et al., 2010; Zaalberg et al., 2019) and provide helpful support for farm managers to make decisions on several aspects of management of the farm. Further improvements in FTIR prediction ability have come from statistical approaches that better capture and describe the complex relationship between chemical bonds and milk components related to specific wavelengths (Soyeurt et al., 2020; Pegolo et al., 2021; Mota et al., 2022). Moreover, Grelet et al. (2015) have pointed out that differences in the spectrometers used to measure the FTIR spectra could result in prediction bias and less accurate predictions. Furthermore, prediction accuracy is also affected by time as a result of changes in the FTIR spectrometer parameters, such as light source intensity, detector sensitivity, and laser stability, although zero-set calibration and weekly calibration adjustments for milk components (i.e., fat, lactose, protein, and TS) can minimize these changes in the signal intensity over time (Young, 1978; Nieuwoudt et al., 2021).

The potential application of FTIR-predicted traits (i.e., indicator traits) for breeding purposes depends on their genetic correlations with measured traits. Several authors have reported high genetic correlations between gold standard measurements (i.e., measured by HPLC) and FTIR predictions for different traits, such as milk coagulation aptitude, fatty acid profiles, and other milk components (Cecchinato et al., 2009, 2015; Soyeurt et al., 2010; Sanchez et al., 2017). Nevertheless, even moderate predictive ability provides valuable information for breeding programs as the breeding value of a sire is based on a rather large amount of data on many relatives that allows noise estimated breeding value correction (Poulsen et al., 2014). Furthermore, Rutten et al. (2010) showed that the size of training set data strongly affects the FTIR predictive ability and the correlation between prediction and gold standard phenotype measurement. As a solution, Mota et al. (2021a) used pooled multibreed data to increase the training set size.

In this work, therefore, we investigated the potential use of FTIR predictions of milk protein fractions in Holstein cattle as indicator traits for breeding purposes. The specific aims were (1) to assess the predictive ability of gradient boosting machine (**GBM**) using random 10-fold and leave-one-batch-out CRV methods for true proteins (**TP**), specifically the casein fractions β-CN, κ-CN, α_{S1}-CN, and α_{S2}-CN, total casein (**TCN**), the whey proteins α-LA and β-LG, and total whey proteins (**TWP**), expressed as proportions of total nitrogen (**% N**) and contents in milk (**g/L**); (2) to measure FTIR predictive ability using calibration and validation databases collected in different years, thereby testing the reproducibility of GBM equations over time; and (3) to estimate the genetic parameters for FTIR predictions

Table 1. Schematic representation regarding the number of animals with phenotypic and Fourier-transform infrared information across the herd explored in this study

		Herd/location							
Trait[1]		Herd 1: Lombardy	Herd 2: Emilia-Romagna	Herd 3: Emilia-Romagna	Herd 4: Veneto	Herd 5: Veneto	Herd 6: Veneto	Herds 7–38: Trentino	Total
TP	g/L	22	70	927	133	17	67	1,174	2,410
	% N	21	69	917	129	17	67	1,168	2,388
Casein									
αs_1-CN	g/L	20	70	927	131	17	67	1,169	2,401
	% N	19	69	921	128	17	66	1,170	2,390
αs_2-CN	g/L	21	69	911	133	17	64	1,188	2,403
	% N	21	70	908	133	17	67	1,183	2,399
β-CN	g/L	21	70	920	133	17	67	1,183	2,411
	% N	20	69	905	130	17	66	1,188	2,395
κ-CN	g/L	21	69	921	133	17	65	1,177	2,403
	% N	20	69	905	133	17	67	1,190	2,401
Glyco-κ-CN	g/L	20	69	911	133	17	65	1,154	2,369
	% N	20	69	908	133	17	67	1,166	2,380
TCN	g/L	22	70	924	132	17	66	1,172	2,403
	% N	21	70	919	127	17	67	1,178	2,399
Whey protein									
α-LA	g/L	20	70	915	133	17	67	1,193	2,415
	% N	21	71	910	133	17	67	1,186	2,405
β-LG	g/L	22	70	929	133	17	67	1,164	2,402
	% N	20	68	918	131	17	66	987	2,207
TWP	g/L	22	70	923	133	17	67	1,168	2,400
	% N	21	70	921	132	17	67	1,169	2,397

[1] TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total whey protein; g/L = protein fraction contents in grams per liter of milk; % N = protein fraction in percentage of nitrogen.

and milk protein fractions measured by the gold standard method (i.e., HPLC), based on bi-trait genomic model analysis.

MATERIALS AND METHODS

The animal procedures in this study were approved by the Organismo Preposto al Benessere Degli Animali (OPBA; Organization for Animal Welfare) of the Università Cattolica del Sacro Cuore (Piacenza, Italy) and by the Italian Ministry of Health (protocol number 510/2019-PR of 19/07/2019). The study was carried out also following ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines.

Field Data

For this study, we integrated data from previous research projects. The data set was compiled by the LATSAN and BENELAT projects, whose aims are to develop new strategies and innovative tools to improve animal welfare and milk quality in dairy farming (Pegolo et al., 2021), the COWPLUS project, which is aimed at evaluating multibreed dairy production systems in mountain areas (Stocco et al., 2017), and the AGER project, within which several farm-level interventions supporting dairy industry innovation were developed (Bisutti et al., 2022). The phenotypic information from COWPLUS project were obtained from specialized (Holstein and Brown Swiss) and dual-purpose breeds (Alpine Grey, Rendena, and Simmental) belonging to 32 multibreed dairy farms (which showed 2 or 5 breeds in the herd) located in the province of Trentino (northeastern Italy).

Milk samples were collected once during the evening milking from specialized dairy breeds, including Holstein (1,618 cows from 30 herds) and Brown Swiss (586 cows from 30 herds), and dual-purpose breeds Alpine Grey (80 cows from 14 herds), Rendena (116 cows from 9 herds), and Simmental (158 cows from 16 herds), which the cows belonged to 38 herds managed under different dairy systems in northern Italy (Table 1). The cows were housed mostly in sand-bedded freestalls and fed twice a day on TMR based on corn and sorghum silage or hay (Emilia-Romagna and Trentino Region herds) supplemented with concentrates. The cows were sampled once after health checks; specifically, animals with clinical disease or on medical treatment were excluded from the study. Further details on the multibreed data set measured in 2013 are available in Stocco et al. (2017) and on the Holstein data set measured in 2019 and 2020 in Pegolo et al. (2021).

Phenotypic and Infrared Information

Milk samples were collected in 55 batches (i.e., herd/date combinations, where each cow was sampled once and each herd was sampled on a specific date): 32 batches in 2013 (1,197 cows), 17 in 2019 (1,011 cows), and 6 in 2020 (350 cows). The large herd (herd 3; Table 1) was sampled in 2019 in 14 batches (856 cows) and in 2020 in 2 batches (80 cows), considering an experimental design where each batch was included in the analysis of milk coagulation properties on fresh milk, and the laboratory could only process a maximum of roughly 65 milk samples per day, as described in Pegolo et al. (2021). All procedures and protocols were identical for both databases assessed in 2013 and 2019 to 2020. The individual milk samples were maintained at 4°C until laboratory analysis (within 24 h). Each sample was divided into the following 2 aliquots: bronopol preservative was added to 1 sample, which was then transferred to the laboratories of the Breeders' Association of the Veneto or of the Province of Trento for analyses of milk quality and composition, and the other sample, without preservative, was transported to the University of Padova (Legnaro, Padova, Italy) for analysis of milk protein fractions by validated reversed-phase HPLC (Maurmayr et al., 2013).

The following milk protein traits were measured: true protein (TP), the major casein fractions α_{S1}-CN, α_{S2}-CN, κ-CN, and β-CN, total casein (TCN; the sum of all casein fractions), the major whey proteins β-LG and α-LA, and TWP (the sum of all whey protein fractions). The milk protein fraction traits were expressed as grams per liter of milk (g/L), calculated by multiplying each milk protein fraction determined by HPLC by the milk casein contents obtained by FTIR and as a percentage of the total milk nitrogen content (% N).

Milk FTIR spectra were recorded on 2,558 Holstein cows and analyzed with a MilkoScan FT6000 (Foss Electric); specifically, they consisted of the transmittance values measured at 1,060 wavenumbers ranging from 5,011 to 925 (cm^{-1}). The 2 spectra obtained were averaged before the data analysis, expressed as an absorbance value [log(1/transmittance)], and standardized to mean zero and standard deviation equal to one. Principal component analysis of the FTIR spectral information was performed, and the Mahalanobis distance was calculated; particularly, animals were considered outliers when they exhibited a Mahalanobis distance based on FTIR information from the average spectral population greater than 3.5 standard deviations (Figure 1). The principal component analysis results pointed out a similarity between old and new FTIR files, indicating that no preprocessing was required to mitigate possible biases due to differences in baseline absorbance over time. After FTIR quality control, milk spectral data from 2,496 Holstein cows remained in the data set. Following phenotypic quality control of the milk protein fractions, observations outside the interval between 3 standard deviations below and above the mean of each batch were removed. After phenotypic quality control, 2,437 cows remained for the analysis; specifically, we had 1,197 cows sampled in 2013, 1,011 cows sampled in 2019, and 229 in 2020, all under similar conditions. A summary of the records for the milk protein fractions by the herd is shown in Table 1. The average (± SD) DIM was 188.26 ± 112.47, parity 2.3 ± 1.51, milk yield 29.30 ± 10.01 kg, and the number of cows per herd/date ranged from 17 to 930. Descriptive statistics for the milk protein fractions expressed in g/L and % N are summarized in Table 2; the boxplots for the milk protein fractions across herds are shown in Supplemental Figure S1 for g/L and Supplemental Figure S2 for % N (https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023).

The 1,067 cows whose phenotypic information was acquired in 2019 and 2020 were genotyped with the Geneseek Genomic profiler Bovine 100K SNP Chip assay (Neogene). The non-autosomal regions were excluded from the subsequent genotypic quality control. Autosomal markers presenting minor allele frequencies of less than 0.05, deviating significantly from the Hardy–Weinberg equilibrium ($P \leq 10^{-5}$), and with a call rate lower than 0.95, were removed. After quality control, 1,056 cows and 81,274 SNP markers remained in the data set.

CRV Scenarios

The FTIR prediction for each milk protein fraction was assessed using random 10-fold cross-validation (**CRV**) and 3 batch-independent CRV scenarios [i.e., herd/date-out, forward/backward (**F/B**), and 5-fold genetic parameters].

Random 10-Fold. In a random 10-fold CRV, the data set considering an admixture of breeds was split into 10-folds of equal size within each breed, with 9-folds used as the training set and the remaining 1-fold as the validation set to assess model performance. This CRV scenario was replicated 10 times, with the average value of these 10 replications used to determine the predictive ability of the model.

Herd/Date Out. In the herd/date-out CRV, which was based on the herd and date on which samples were collected, 80% of the population was randomly assigned to the training set (44 herd/dates), and the other 20% to the validation set (11 herd/dates). Given the variability in herd size, random sampling was carried out to ensure greater homogeneity in the number of animals in

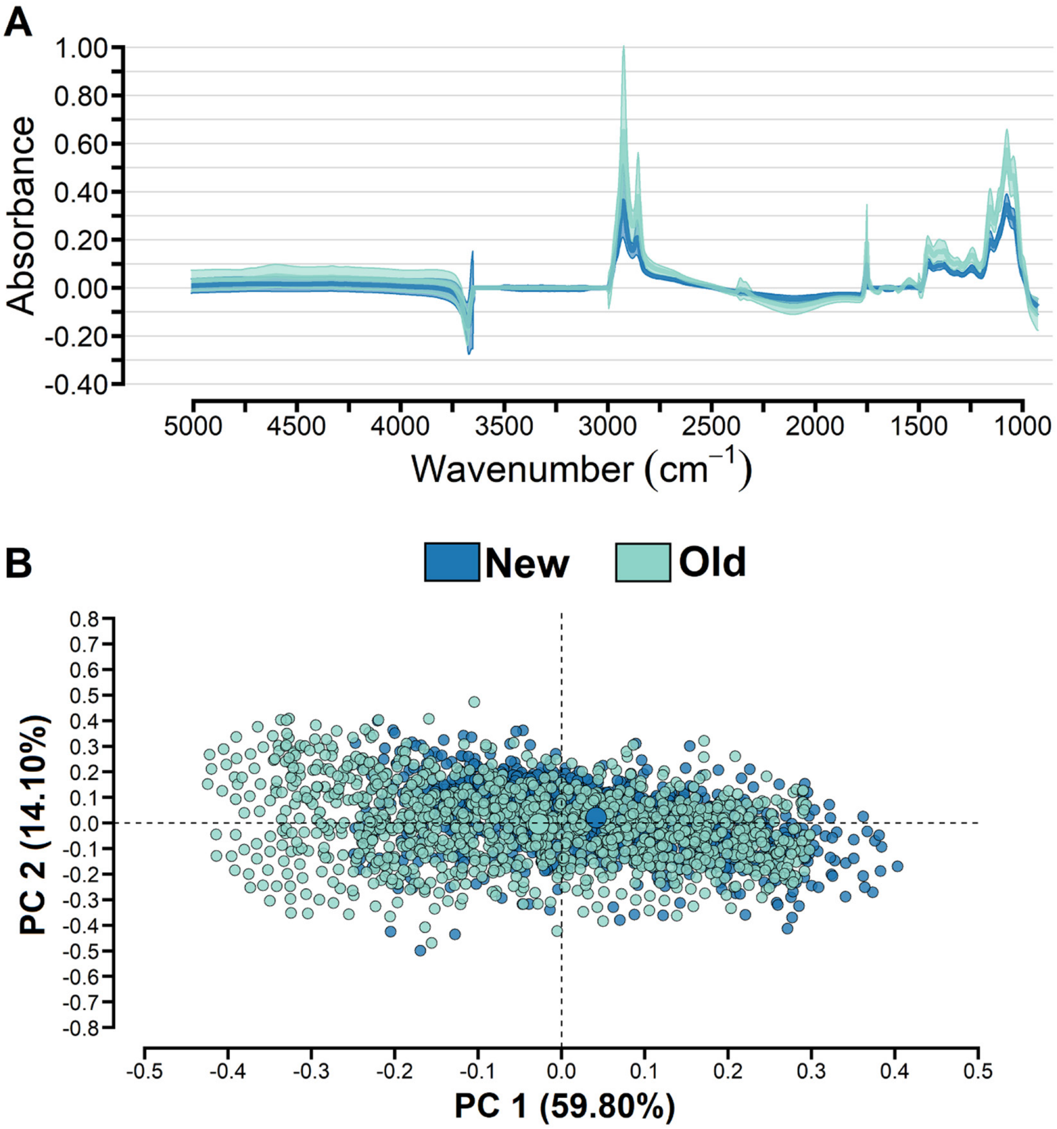

Figure 1. (A) Average value for Fourier-transform infrared (FTIR) information expressed as absorbance (solid line represents the average, and color region represents the mean ± 3 SD) and (B) principal component (PC) for the FTIR spectral data of milk samples. Colors represent the years of FTIR assessment; old population (2013; n = 1,197 cows) and new population (2019 and 2020; n = 1,241 cows).

the training and validation sets. For this, batches were grouped into 5 similar groups based on the number of cows (ranging from 43 to 45 cows) and then divided into 80% for training and 20% validation within each group. This CRV scenario was repeated 10 times as for the random 10-fold. The 80% herd/dates (i.e., the entire herd/date which encompasses the production level) considered in the training set were independent of that 20% of herd/date considered in the validation set.

F/B. In this CRV scenario, we wanted to assess the predictive ability of the GBM equations across the different years of sampling to test their reproducibility over time. The training and validation were subsets of animals classified according to the year the FTIR spectral data were recorded, and the herds in the "old" (2013; multibreed herds, 1,197 cows) and "new" data set (2019–2020, 1,240 Holstein cows) were completely independent among them. For the Forward CRV, "old" FTIR data were used as the training set, whereas the "new" was assigned as the validation set. For the Backward CRV, the "new" FTIR data were assigned to the training set and the "old" FTIR data as the validation set. The farms considered in training and validation sets were completely independent among them.

CRV for Genetic Parameters. We used CRV for genetic parameters (**CRV-gen**) to assess the useful-

Table 2. Descriptive statistics for milk protein fractions expressed in different ways in Holstein cows after quality control[1]

Trait[2]	N	Mean	SD	Minimum	Maximum	IQR[3]
Protein fraction content (g/L)						
TP	2,429	32.48	3.849	22.7	43.32	5.04
Casein						
α_{S1}-CN	2,421	9.24	1.18	5.92	12.73	1.56
α_{S2}-CN	2,424	3.12	0.71	1.29	4.98	0.93
β-CN	2,431	9.82	1.53	5.68	14.10	2.13
κ-CN	2,423	5.32	1.10	2.55	8.24	1.55
Glyco-κ-CN	2,388	1.79	0.66	0.44	3.88	0.92
TCN	2,422	27.45	3.11	19.30	36.30	4.10
Whey protein						
α-LA	2,436	0.87	0.17	0.39	1.36	0.25
β-LG	2,421	3.99	1.14	1.42	7.14	1.55
TWP	2,419	4.99	1.13	2.21	8.20	1.51
Protein fraction proportion (% N)						
TP	2,407	92.46	2.94	84.84	100.35	4.00
Casein						
α_{S1}-CN	2,411	26.28	1.84	20.85	31.68	2.58
α_{S2}-CN	2,420	8.91	1.86	4.13	13.81	2.50
β-CN	2,416	28.01	2.86	20.19	35.73	3.97
κ-CN	2,422	15.12	2.39	8.97	21.58	3.38
Glyco-κ-CN	2,401	5.08	1.71	1.36	10.24	2.39
TCN	2,419	78.30	1.16	75.07	81.43	1.56
Whey protein						
α-LA	2,426	2.50	0.53	1.06	3.95	0.78
β-LG	2,224	9.13	2.56	3.16	16.00	3.90
TWP	2,417	14.17	2.70	7.02	21.44	3.54

[1] For descriptive trait by herd, see Supplemental Figures S1 and S2 (https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023).

[2] TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein; g/L = protein fraction contents in grams per liter of milk; % N = protein fraction in the percentage of nitrogen.

[3] IQR = interquartile interval.

ness of FTIR predictions as a potential indicator trait for breeding purposes. In the first step, we assigned a fixed training population considering animals without pedigree information from the multibreed data set (138 Holstein cows, 537 Brown Swiss, 74 Alpine Grey, 101 Rendena, and 107 Simmental) to exploit all the available FTIR information efficiently. Next, the data set that considers animals with known pedigree and genomic information (i.e., the new FTIR data set sampled between 2019 and 2020; Supplemental Figure S3, https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023), which encompasses 23 herds/date, and we used it as the base to split the population into 5-folds based on herd-date randomly. The entire herd-date was considered within each fold, approximately 4 herd-date for 2-fold and 5 herd-date for 3-folds. From these 5-folds, 1-fold was assigned to the fixed training population, and 4-folds were assigned to the validation set (independent from the training set), aiming to guarantee a large number of animals in the validation set. Thus, the training set comprised the fixed population (957 cows without pedigree information) plus 1-fold, and the validation set 4-folds. Finally, we repeated the process 5 times, and predictions obtained on each validation set (a total of 5 different validation folds) were used to estimate the genetic parameters using a bi-trait animal model for the FTIR predictions and the measurements using the HPLC approach for milk protein fractions.

FTIR Calibration Equations

We selected the GBM statistical method because previous results indicated that this method achieved the highest accuracy of FTIR-based prediction of different phenotypic traits compared with the partial least squares (**PLS**; Mota et al., 2021b). We compared the GBM performance against the gold-standard method (PLS) for milk protein fractions in grams per liter (g/L) and percentage of nitrogen (% N) for the following CRV scenarios: 10-folds, herd/date-out, F/B (Supplemental Table S1, https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023). The PLS regression was implemented using the PLS R package, version 2.8.1 (Mevik and Wehrens, 2007).

The milk protein fractions were predicted using the GBM, an ensemble learning approach that converts weak learners into strong learners through a sequential combination of different regression tree models, with bias and variance reduced by shrinkage and variable selection (Hastie et al., 2009). This method was chosen because of its greater accuracy with FTIR-based predictions of different phenotypic traits and validation scenarios (Mota et al., 2021b). We implemented GBM with a random tuning of the 4 main hyperparameters [i.e., the number of trees, learning rate, maximum tree depth, and minimum samples per leaf; Natekin and Knoll, 2013]. We performed the random tuning by splitting the training subset into 5-fold to optimize the hyperparameters and increase model performance [i.e., higher accuracy and lower root mean square error (**RMSE**)]. The number of trees values were determined in the range 100 to 5,000 in intervals of 100, the learning rate in the range 0.01 to 1 in intervals of 0.01, the maximum tree depth in the range 5 to 80 in intervals of 5, and minimum samples per leaf in the range 10 to 100 in intervals of 10. The GBM model was built with a random search using the h2o.grid function in the R h2o package (https://cran.r-project.org/web/packages/h2o). The relative importance of the FTIR wavelength predictor was determined by calculating the relative influence of the FTIR wavelength on predictive ability improvements during the regression tree building process, this being the sum of the squared improvements over all internal nodes of the tree for which the FTIR wavelength was chosen as the partitioning variable (Hastie et al., 2009). The predictive ability of the GBM approach was assessed by Pearson correlation (r) between the observed and predicted phenotypes, and RMSE were assessed in the validation set across the CRV scenarios. The RMSE was calculated as $\sqrt{\frac{\sum(y_{lab}-y_{pred})^2}{n}}$, where y_{lab} is the measured phenotype and y_{pred} is the predicted phenotype in the validation set. We assessed the model unbiasedness by the slope of the linear regression of the gold standard laboratory measurements on predicted values in each CRV scenario for milk protein fractions.

Genetic Parameters

The genetic parameters for gold-standard laboratory measurements (y_{lab}) and FTIR-based predictions from CRV-gen scenario (y_{pred}; i.e., 5 different validation sets), for milk protein fractions expressed in g/L and % N with a bi-trait animal model using a single-step genomic BLUP, separately for each fold, as follows:

$$\begin{bmatrix} \boldsymbol{y_{lab}} \\ \boldsymbol{y_{pred}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_{lab}} & 0 \\ 0 & \boldsymbol{X_{pred}} \end{bmatrix} \begin{bmatrix} \boldsymbol{b_{lab}} \\ \boldsymbol{b_{pred}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{W_{lab}} & 0 \\ 0 & \boldsymbol{W_{pred}} \end{bmatrix} \begin{bmatrix} \boldsymbol{a_{lab}} \\ \boldsymbol{a_{pred}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e_{lab}} \\ \boldsymbol{e_{pred}} \end{bmatrix},$$

where $\boldsymbol{y_{lab}}$ is the gold standard laboratory measurement and $\boldsymbol{y_{pred}}$ is the FTIR-based prediction from the CRV-gen scenario for milk protein fractions; $\boldsymbol{b_{lab}}$ and $\boldsymbol{b_{pred}}$ are the vectors of the fixed effects of DIM (6 classes: class 1, less than 60 d; class 2, 60–120 d; class 3, 121–180 d; class 4, 181–240 d; class 5, 241–300 d; class 6, more than 300 d), parity (4 classes: 1, 2, 3, ≥4), and herd-date for gold standard laboratory measurement and FTIR-based prediction, respectively; $\boldsymbol{a_{lab}}$ and $\boldsymbol{a_{pred}}$ are the vectors of additive genetic effects for gold standard laboratory measurement and FTIR-based prediction, respectively; $\boldsymbol{X_{lab}}$, $\boldsymbol{X_{pred}}$, $\boldsymbol{W_{lab}}$, and $\boldsymbol{W_{pred}}$ are the incidence matrices relating $\boldsymbol{y_{lab}}$ and $\boldsymbol{y_{pred}}$ to the fixed effects ($\boldsymbol{b_{lab}}$ and $\boldsymbol{b_{pred}}$) and the additive effects ($\boldsymbol{a_{lab}}$ and $\boldsymbol{a_{pred}}$), respectively; and $\boldsymbol{e_{lab}}$ and $\boldsymbol{e_{pred}}$ are the residual effect for gold standard laboratory measurement and FTIR-based prediction, respectively. The single-step genomic BLUP model was fitted under the following assumptions for the random effects:

$$a=\left\{a_j\right\}\sim N\left(0,\ \mathbf{H}\otimes\begin{bmatrix} \sigma^2_{a\,lab} & \sigma_{a\,lab,pred} \\ \sigma_{a\,lab,pred} & \sigma^2_{a\,pred} \end{bmatrix}\right)$$

and $e=\left\{e_{ij}\right\}\sim N(0,\mathbf{I}\otimes\mathbf{R})$, where $\tilde{A}^2_{a\,lab}$, $\tilde{A}^2_{a\,pred}$, and $\tilde{A}_{a\,lab,pred}$ are the genetic variances in the gold standard measurements, the FTIR-based predictions, and the covariances between them, respectively; $\mathbf{R}$ is the (co)variance residual matrix

$$\begin{bmatrix} \sigma^2_{e\,lab} & \sigma_{e\,lab,pred} \\ \sigma_{e\,lab,pred} & \sigma^2_{e\,pred} \end{bmatrix},$$

where $\sigma^2_{e\,lab}$, $\sigma^2_{e\,pred}$, and $\sigma_{e\,lab,pred}$ are the residual variances in the gold standard measurements, the FTIR-based predictions, and the covariances between them, respectively; $\mathbf{I}$ is the identity matrix, and the symbol ⊗ represents the Kronecker product. $\mathbf{H}$ is a matrix that combines pedigree and genomic information (Aguilar et al., 2010), and its inverse ($\mathbf{H}^{-1}$) is given by

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{A}^{-1}$ is the inverse of the pedigree relationship matrix, $\mathbf{A}_{22}^{-1}$ is the inverse of the pedigree relationship matrix for genotyped animals, and $\mathbf{G}^{-1}$ is the inverse of the genomic relationship matrix obtained according to VanRaden (2008). We assumed a flat prior distribution for the fixed effects and used an inverse Wishart distribution as a prior for the random effects.

Heritability (h^2) was calculated based on the posterior (co)variance estimates for each trait as $h^2 = \sigma_a^2 \Big/ \left(\sigma_a^2 + \sigma_e^2\right)$, where σ_a^2 is the additive genetic variance, and σ_e^2 is the residual variance for the gold standard measurements ($\boldsymbol{y_{lab}}$) or FTIR-based predictions ($\boldsymbol{y_{pred}}$) of milk protein fractions expressed in % N and g/L. Genetic ($\boldsymbol{r_g}$) and phenotypic ($\boldsymbol{r_p}$) correlation estimates were calculated as follows:

$$r_g = \frac{\sigma_{a\,lab,pred}}{\sqrt{\sigma^2_{a\,lab} \times \sigma^2_{a\,pred}}} \text{ and } r_p = \frac{\sigma_{p\,lab,pred}}{\sqrt{\sigma^2_{p\,lab} \times \sigma^2_{p\,pred}}},$$

where $\sigma^2_{p\,lab}$ and $\sigma^2_{p\,pred}$ denote the phenotypic variance calculated as the sum of σ_a^2 and σ_e^2, and $\sigma_{p\,lab,pred}$ is the phenotypic covariance between traits calculated as the sum of the additive genetic and residual covariance for the gold standard measurements ($\boldsymbol{y_{lab}}$) or FTIR-based predictions ($\boldsymbol{y_{pred}}$).

The model was implemented in the R package BGLR 1.0.9 (Pérez and de los Campos, 2014). The genetic parameters were obtained from the posterior distribution using the Markov Chain Monte Carlo method via Gibbs sampling. We ran a single chain of 500,000 cycles, with a burn-in of the first 50,000 iterations, with samples stored every 10 cycles. Hence, the posterior means were obtained from 45,000 samples, and the analysis converged through visual inspection using the Bayesian output analysis (Smith, 2007), and for the Geweke (Geweke, 1992), the convergence was attained for the evaluated traits with a P-value > 0.15.

RESULTS

FTIR Prediction

The fitting statistics of the GBM model for milk protein fractions using the different CRV strategies are shown in Table 3, expressed in g/L, and Table 4, expressed in % N. Predictive ability was lower when the milk protein fractions were expressed in % N than in g/L (Supplemental Figure S4, https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023). These reductions in prediction accuracy ranged from −27 to −51% using the 10-fold CRV scenario [except total casein (TCN), which increased by 3.61%], −2.47 to −51% using herd/date-out, −7.59 to 52.31 for F/B, and −1.2 to −45% using CRV-gen (Supplemental Figure S4). On the other hand, comparing the average RMSE for each average was observed a reduction of 36% with values ranging from −1 to −88%, whereas an increased 23% for α_{S2}-CN, 28% for glycosylated (**glyco**)-κ-CN, 19% for α-LA, 26% for β-LG, and 20% for TWP (Table 3 and 4).

Predictive ability for milk protein fractions expressed in g/L ranged from 0.63 ± 0.069 to 0.88 ± 0.020 with random 10-fold CRV, 0.62 ± 0.063 to 0.83 ± 0.077 with herd/date-out, 0.60 ± 0.063 to 0.78 ± 0.053 with F/B, and 0.62 ± 0.067 to 0.87 ± 0.031 with CRV-gen (Table 3). True protein and TCN showed better prediction accuracy than α-LA, which had the lowest predictive ability across all the CRV scenarios evaluated. With F/B, the predictive ability of the models based on FTIR milk spectra was lower than those obtained from a random 10-fold CRV scenario, with reductions ranging from −4.8% for α-LA to −19.8% for β-LG (Table 3). Unbiased estimation based on the regression slope of the gold standard measures of milk protein fractions on the FTIR-predicted values indicated a great difference across the CRV scenarios. The slope values obtained with the F/B CRV scenario indicated a prediction bias greater than 1, with values ranging from 10% for glyco-κ-CN to 27% for blood β-CN and a decrease of 8% for TP. The slope coefficient estimates obtained with the random 10-fold and 5-fold genetic CRV scenarios were less biased than those obtained with the herd/date-out and F/B CRV scenarios. These results agree with the assessment of model fit by RMSE and showed that the random 10-fold CRV and CRV-gen scenarios led to lower residual parameters compared with herd/date-out and F/B, showing a greater reduction in the RMSE ranging from 4% for κ-CN to 55% for α-LA, and from 8% for β-CN to 73% for α-LA, respectively (Table 3).

The predictive ability of milk protein fractions expressed in % N ranged from 0.34 ± 0.034 to 0.86 ± 0.023 with random 10-fold CRV, 0.33 ± 0.094 to 0.79 ± 0.071 with herd/date-out, 0.31 ± 0.024 to 0.73 ± 0.084 with F/B, and 0.36 ± 0.036 to 0.81 ± 0.022 with CRV-gen (Table 4). The best predictive abilities were obtained for TCN (R^2 = 0.73–0.86) and TP (R^2 = 0.60–0.64), and the lowest for β-CN (R^2 = 0.31–0.36) across all the CRV scenarios evaluated (Table 4). With the F/B CRV scenario, the predictive ability of FTIR predictions was lower than that of the random 10-fold CRV scenario, with reductions ranging from −6.3% for TP to −15.79% for α_{S1}-CN (Table 4). On the other hand, CRV-gen exhibited lower predictive ability than random 10-fold CRV, ranging from −1.56% for TP to −8.33 for glyco-κ-CN. However, the milk protein fractions β-CN and TWP showed an increased predictive ability of 4.7 and 2.63%, respectively. Inflation, estimated as the regression slope of the measured milk protein fractions on the FTIR-predicted values, indicated a slight variation in slope values between random 10-fold

Table 3. Average model predictive performance (R^2) and SD values for milk protein fractions in grams per liter of milk considering different cross-validation (CRV) scenarios

Trait[1]	CRV scenario[2]	R^2	RMSE[3]	Slope[4]	RMSE/mean[5] (%)
TP	Random 10-fold	0.88 ± 0.020	1.35 ± 0.117	0.99 ± 0.025	4.16
	Herd/date-out	0.83 ± 0.077	1.53 ± 0.371	1.08 ± 0.028	4.71
	Forward/backward	0.78 ± 0.053	1.74 ± 0.387	0.92 ± 0.096	5.36
	CRV-gen	0.87 ± 0.031	1.36 ± 0.159	0.98 ± 0.029	4.19
Casein					
α_{S1}-CN	Random 10-fold	0.70 ± 0.017	0.63 ± 0.018	1.02 ± 0.033	6.82
	Herd/date-out	0.68 ± 0.048	0.66 ± 0.069	1.08 ± 0.081	7.14
	Forward/backward	0.66 ± 0.066	0.72 ± 0.050	1.15 ± 0.049	7.79
	CRV-gen	0.69 ± 0.035	0.65 ± 0.038	1.05 ± 0.048	7.03
α_{S2}-CN	Random 10-fold	0.68 ± 0.042	0.41 ± 0.014	1.03 ± 0.063	13.14
	Herd/date-out	0.65 ± 0.087	0.46 ± 0.036	1.11 ± 0.091	14.74
	Forward/backward	0.61 ± 0.091	0.48 ± 0.023	1.22 ± 0.115	15.38
	CRV-gen	0.65 ± 0.048	0.43 ± 0.008	1.08 ± 0.083	13.78
β-CN	Random 10-fold	0.70 ± 0.027	0.84 ± 0.054	1.03 ± 0.043	8.55
	Herd/date-out	0.68 ± 0.081	0.89 ± 0.104	1.13 ± 0.095	9.06
	Forward/backward	0.65 ± 0.054	0.91 ± 0.108	1.27 ± 0.104	9.27
	CRV-gen	0.69 ± 0.037	0.86 ± 0.036	1.09 ± 0.073	8.76
κ-CN	Random 10-fold	0.71 ± 0.054	0.74 ± 0.146	1.04 ± 0.035	13.91
	Herd/date-out	0.68 ± 0.066	0.77 ± 0.160	1.12 ± 0.077	14.47
	Forward/backward	0.65 ± 0.067	0.81 ± 0.228	1.26 ± 0.085	15.23
	CRV-gen	0.69 ± 0.046	0.75 ± 0.153	1.05 ± 0.056	14.10
Glyco-κ-CN	Random 10-fold	0.72 ± 0.027	0.33 ± 0.019	1.04 ± 0.027	18.44
	Herd/date-out	0.71 ± 0.090	0.36 ± 0.033	0.92 ± 0.028	20.11
	Forward/backward	0.66 ± 0.052	0.38 ± 0.054	1.10 ± 0.074	21.23
	CRV-gen	0.70 ± 0.039	0.34 ± 0.026	1.02 ± 0.028	18.99
TCN	Random 10-fold	0.83 ± 0.019	1.29 ± 0.072	0.99 ± 0.036	4.70
	Herd/date-out	0.81 ± 0.027	1.39 ± 0.099	1.13 ± 0.057	5.06
	Forward/backward	0.79 ± 0.044	1.69 ± 0.245	1.25 ± 0.103	6.16
	CRV-gen	0.82 ± 0.023	1.29 ± 0.077	1.02 ± 0.049	4.70
Whey protein					
α-LA	Random 10-fold	0.63 ± 0.069	0.11 ± 0.008	1.06 ± 0.045	12.64
	Herd/date-out	0.62 ± 0.051	0.17 ± 0.013	1.10 ± 0.059	19.54
	Forward/backward	0.60 ± 0.063	0.19 ± 0.014	1.24 ± 0.089	21.84
	CRV-gen	0.62 ± 0.067	0.11 ± 0.011	0.95 ± 0.039	12.64
β-LG	Random 10-fold	0.81 ± 0.094	0.55 ± 0.064	1.10 ± 0.046	13.78
	Herd/date-out	0.72 ± 0.121	0.63 ± 0.113	1.15 ± 0.047	15.79
	Forward/backward	0.65 ± 0.137	0.60 ± 0.076	1.19 ± 0.052	15.04
	CRV-gen	0.76 ± 0.097	0.56 ± 0.066	1.11 ± 0.055	14.04
TWP	Random 10-fold	0.79 ± 0.089	0.57 ± 0.036	1.03 ± 0.062	11.42
	Herd/date-out	0.69 ± 0.128	0.66 ± 0.144	1.07 ± 0.075	13.23
	Forward/backward	0.67 ± 0.013	0.69 ± 0.085	1.12 ± 0.119	13.83
	CRV-gen	0.75 ± 0.099	0.60 ± 0.054	1.02 ± 0.073	12.02

[1] TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein.

[2] CRV-gen = CRV for genetic parameters.

[3] RMSE = root mean square error.

[4] Slope = the slope of the linear regression of the gold standard laboratory measurements on predicted values across the CRV scenarios for each trait.

[5] RMSE/mean (%) = RMSE expressed as a ratio of the mean for each trait.

CRV and CRV-gen, with slope values ranging from 0.97 ± 0.038 for TWP to 1.12 ± 0.011 for α_{S2}-CN, and from 0.97 ± 0.024 for β-CN to 1.12 ± 0.01 for glyco-κ-CN, respectively (Table 4). In contrast, the slope of the F/B CRV scenario showed a tendency to biased predictions, with values ranging from 1.08 ± 0.034 for TP to 1.26 ± 0.060 for TWP. Overall, FTIR prediction using the herd/date-out and F/B CRV scenarios produced more biased predictions than random 10-fold CRV and 5-fold CRV-gen.

Associations Between FTIR Wavelength Absorbance and Milk Protein Fractions

Overall, the milk protein fractions fell in the same FTIR wavelength regions, whether measured g/L (Figure 2) or % N (Figure 3). For milk protein fractions expressed in g/L, 3 main regions were found to explain more than 0.5% importance in the GBM approach (Figure 2). Significant individual FTIR wavelengths ranged from 37 for glyco-κ-CN to 68 for α_{S2}-CN

Table 4. Average model predictive performance (R^2) and SD values for milk protein fractions in the percentage of nitrogen considering different cross-validation (CRV) scenarios

Trait[1]	CRV scenario[2]	R^2	RMSE[3]	Slope[4]	RMSE/mean[5] (%)
TP	Random 10-fold	0.64 ± 0.041	1.76 ± 0.087	0.98 ± 0.013	1.90
	Herd/date-out	0.62 ± 0.045	1.88 ± 0.098	1.05 ± 0.025	2.03
	Forward/backward	0.60 ± 0.046	1.93 ± 0.038	1.08 ± 0.034	2.09
	CRV-gen	0.63 ± 0.047	1.81 ± 0.082	0.98 ± 0.018	1.96
Casein					
α_{S1}-CN	Random 10-fold	0.38 ± 0.028	1.44 ± 0.113	1.05 ± 0.021	5.48
	Herd/date-out	0.35 ± 0.077	1.53 ± 0.059	0.93 ± 0.043	5.82
	Forward/backward	0.32 ± 0.021	1.58 ± 0.102	1.19 ± 0.056	6.01
	CRV-gen	0.38 ± 0.038	1.49 ± 0.011	1.07 ± 0.025	5.67
α_{S2}-CN	Random 10-fold	0.38 ± 0.029	1.47 ± 0.047	1.02 ± 0.011	16.50
	Herd/date-out	0.35 ± 0.104	1.58 ± 0.099	0.97 ± 0.019	17.73
	Forward/backward	0.34 ± 0.019	1.68 ± 0.124	1.19 ± 0.031	18.86
	CRV-gen	0.37 ± 0.026	1.51 ± 0.032	1.03 ± 0.016	16.95
β-CN	Random 10-fold	0.34 ± 0.034	2.10 ± 0.063	0.99 ± 0.025	7.50
	Herd/date-out	0.33 ± 0.094	2.37 ± 0.109	1.11 ± 0.042	8.46
	Forward/backward	0.31 ± 0.024	2.56 ± 0.069	1.16 ± 0.049	9.14
	CRV-gen	0.36 ± 0.036	2.24 ± 0.071	0.97 ± 0.024	8.00
κ-CN	Random 10-fold	0.47 ± 0.035	1.74 ± 0.079	1.02 ± 0.020	11.51
	Herd/date-out	0.45 ± 0.065	1.79 ± 0.091	0.96 ± 0.033	11.84
	Forward/backward	0.43 ± 0.022	1.88 ± 0.025	1.15 ± 0.039	12.43
	CRV-gen	0.46 ± 0.035	1.77 ± 0.061	0.98 ± 0.023	11.71
Glyco-κ-CN	Random 10-fold	0.48 ± 0.029	1.23 ± 0.037	1.08 ± 0.009	24.21
	Herd/date-out	0.43 ± 0.034	1.28 ± 0.045	1.17 ± 0.017	25.20
	Forward/backward	0.41 ± 0.029	1.36 ± 0.056	1.22 ± 0.028	26.77
	CRV-gen	0.44 ± 0.030	1.23 ± 0.042	1.11 ± 0.010	24.21
TCN	Random 10-fold	0.86 ± 0.023	0.44 ± 0.044	1.06 ± 0.011	0.56
	Herd/date-out	0.79 ± 0.071	0.59 ± 0.062	1.14 ± 0.019	0.75
	Forward/backward	0.73 ± 0.084	0.66 ± 0.089	1.17 ± 0.026	0.84
	CRV-gen	0.81 ± 0.022	0.52 ± 0.043	1.04 ± 0.017	0.66
Whey protein					
α-LA	Random 10-fold	0.44 ± 0.074	0.39 ± 0.063	0.99 ± 0.025	15.60
	Herd/date-out	0.41 ± 0.108	0.40 ± 0.098	1.09 ± 0.039	16.00
	Forward/backward	0.40 ± 0.172	0.43 ± 0.073	1.15 ± 0.047	17.20
	CRV-gen	0.43 ± 0.079	0.36 ± 0.066	1.06 ± 0.031	14.40
β-LG	Random 10-fold	0.57 ± 0.046	1.59 ± 0.109	0.98 ± 0.010	17.42
	Herd/date-out	0.52 ± 0.073	1.72 ± 0.154	1.14 ± 0.015	18.84
	Forward/backward	0.51 ± 0.011	1.79 ± 0.322	1.22 ± 0.023	19.61
	CRV-gen	0.57 ± 0.047	1.63 ± 0.109	0.98 ± 0.011	17.85
TWP	Random 10-fold	0.38 ± 0.062	2.08 ± 0.107	0.97 ± 0.038	14.68
	Herd/date-out	0.36 ± 0.083	2.14 ± 0.178	1.12 ± 0.056	15.10
	Forward/backward	0.34 ± 0.037	2.23 ± 0.301	1.26 ± 0.060	15.74
	CRV-gen	0.39 ± 0.063	2.10 ± 0.102	1.06 ± 0.041	14.82

[1] TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein.

[2] CRV-gen = CRV for genetic parameters.

[3] RMSE = root mean square error.

[4] Slope = the slope of the linear regression of the gold standard laboratory measurements on predicted values across the CRV scenarios for each trait.

[5] RMSE/mean = RMSE expressed as a ratio of the mean of each trait.

(Supplemental Figure S5, https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023), covering 3 main regions (3,200–2,900 cm^{-1}, 1,750–1,500 cm^{-1}, and 1,250–950 cm^{-1}). Twenty-nine wavelengths were shared by at least 5 milk proteins expressed in g/L, including 1,680 cm^{-1} (9 traits), 1,727, 2,972, 2,975, and 3,149 cm^{-1} (8 traits), 1,677 and 2,979 cm^{-1} (7 traits), 1,619, 1,653, 1,684, 2,968, 2,983, 3,018, 3,022, 3,184, and 3,191 cm^{-1} (6 traits), and 1,006, 1,561, 1603, 1,615, 1,646, 1,715, 2,918, 2,987, 2,991, 2,995, 3,029, 3,122, and 3,241 cm^{-1} (5 traits). These regions contributed between 0.61 and 2.21% of predictive ability in the GBM approach. Consistent with these results, the main milk FTIR wavelength regions were highly correlated with the target milk protein traits expressed in g/L, with values ranging from −0.25 to −0.97 and from 0.23 to 0.99 (r; Supplemental Figure S6, https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023).

For milk proteins expressed in % N, 4 main regions (in the case of TP, α_{S2}-CN, κ-CN, glyco- κ-CN, β-LG, and TWP) and a further 5 (in the case of α_{S1}-CN, β-CN, TCN, and α-LA) were found to explain more than 0.5%

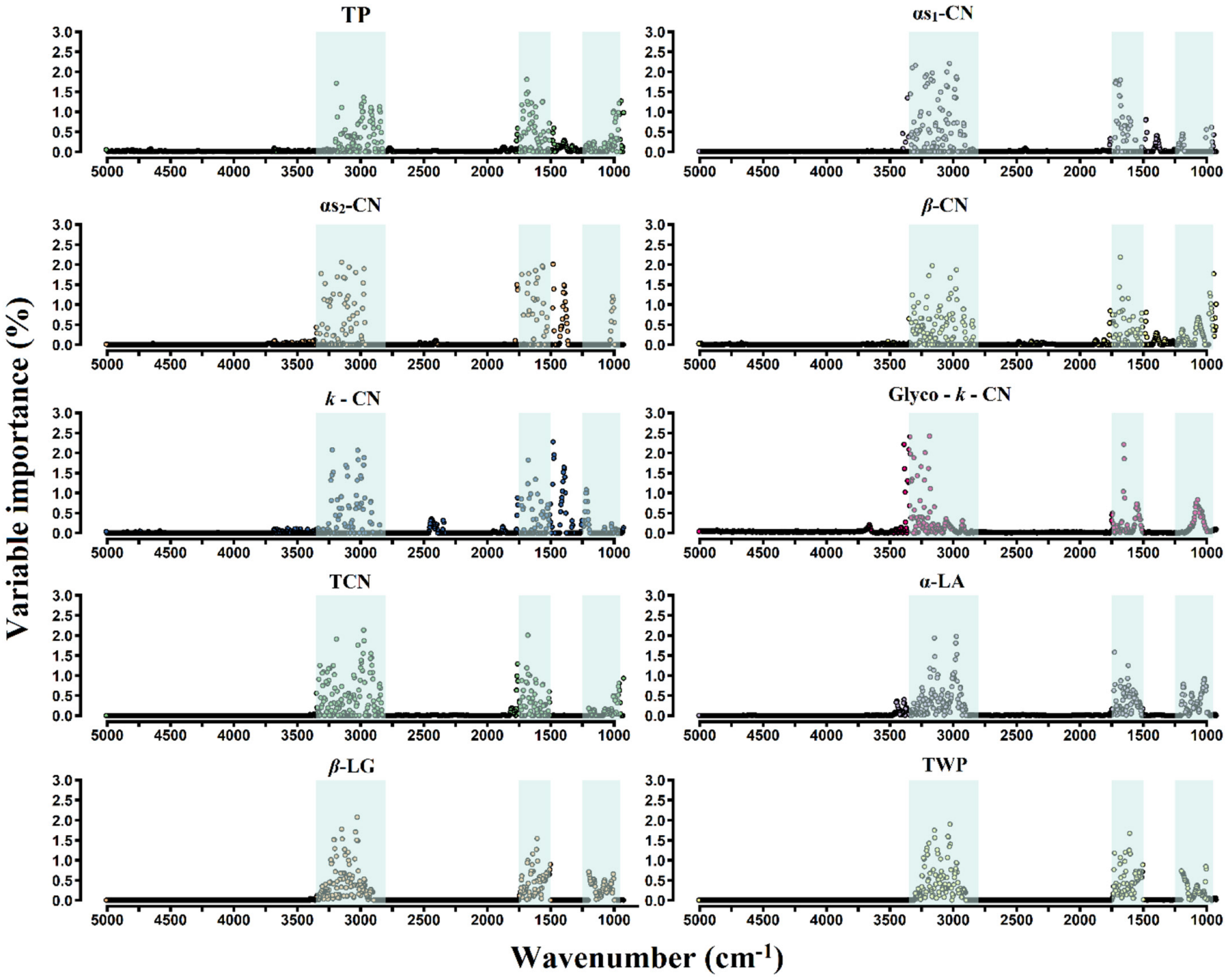

Figure 2. Variable importance for single-wavelength absorbance associations across the entire Fourier-transform infrared spectrum (1,060 wavelengths) for milk protein fractions [true protein (TP); major casein fractions: α_{S1}-CN, α_{S2}-CN, κ-CN, glycosylated-κ-CN (glyco-κ-CN), β-CN; total casein (TCN); major whey proteins: β-LG and α-LA; and total whey protein (TWP)], expressed in grams per liter of milk.

of importance in the GBM approach (Figure 3). The number of significant individual FTIR wavelengths ranged from 44 for β-CN to 63 for α_{S1}-CN, κ-CN, and TCN (Supplemental Figure S5), which covered the following regions: 4,900 to 4,650 cm^{-1}, 3,600 to 3,350 cm^{-1}, 3,200 to 2,900 cm^{-1}, 2,550 to 2,400 cm^{-1}, 1,750 to 1,500 cm^{-1}, and 1,250 to 950 cm^{-1}. These regions are related to overtones and combinations of the vibrations of some chemical bonds, such as C–O symmetric stretching, C=O stretching, C–H, N–H, O–H, and S–H. Some peaks exhibited moderate to strong associations with milk protein fractions expressed as % N in these regions (Figure 3). The major wavelength shared by at least 6 milk proteins was 1,603 cm^{-1} [variable importance (**VI**) > 0.90%], which was shared by TP, α_{S1}-CN, α_{S2}-CN, TCN, β-LG, and TWP. The wavelength 3,245 cm^{-1} (VI 0.71% to 1.66%) and 1,688 cm^{-1} (VI 0.60% to 3.03%) were each shared by the following 6 milk protein fractions: TP, α_{S1}-CN, α_{S2}-CN, β-CN, TCN, and β-LG in the former case, and β-CN, κ-CN, glyco-κ-CN, α-LA, β-LG, and TWP in the latter case. Fourteen wavelengths (i.e., 3,041, 1,611, 971, 3,091, 3,234, 1,607, 3,207, 3,049, 1,646, 1,665, 1,580, 3,026, 3,211, 3,029) were shared by a group of 5 milk proteins (Supplemental Figure S5B) and explained 0.61 to 2.70% of the predictive ability of the GBM approach. The Pearson correlations among the major milk FTIR wavelength regions with the target milk protein fractions expressed

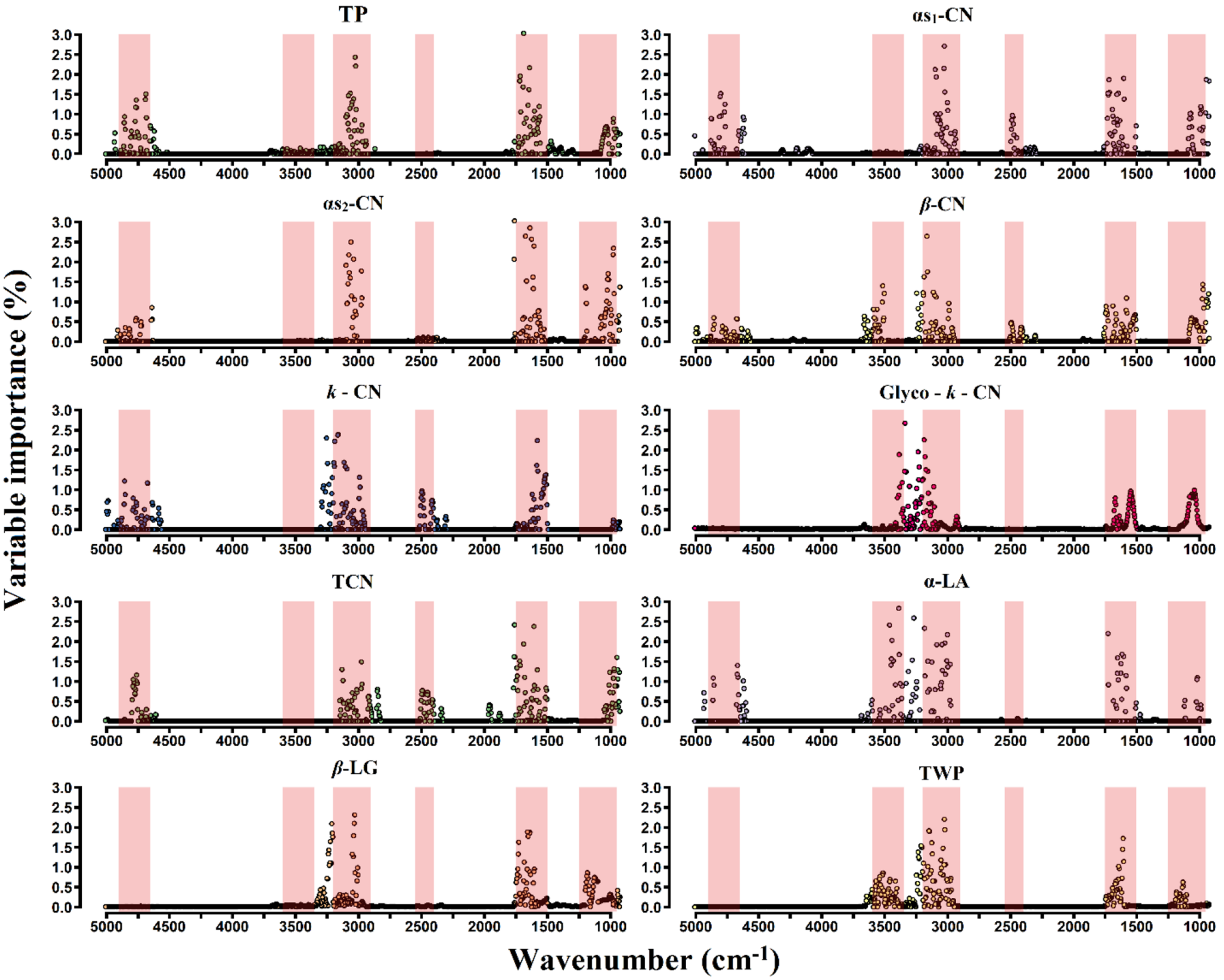

Figure 3. Variable importance for single-wavelength absorbance associations across the entire Fourier-transform infrared spectrum (1,060 wavelengths) for milk protein fractions [true protein (TP); major casein fractions: αs_1-CN, αs_2-CN, κ-CN, glycosylated-κ-CN (glyco-κ-CN), β-CN; total casein (TCN); major whey proteins: β-LG and α-LA; and total whey protein (TWP)], expressed as the percentage of the total milk nitrogen content (% N).

as % N were highly correlated, with ranges of −0.18 to −0.98 and 0.17 to 0.99 (Supplemental Figure S7, https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023).

Genetic Parameters of Laboratory-Measured and FTIR-Predicted Milk Protein Fractions

Table 5 reports the genetic parameter estimates for laboratory-measured and FTIR-predicted milk protein fractions expressed in g/L, which yielded heritability estimates that were either moderate (TCN, TP, glyco-κ-CN, α_{S2}-CN, and α-LA) or high (β-CN, α_{S1}-CN, β-LG, κ-CN, and TWP). Heritability estimates for the FTIR-based predictions were slightly lower than those obtained for the laboratory measurements (Table 5). However, these reductions were slight, ranging from −1.93% for β-CN to −7.25% for α-LA, meaning that FTIR-based predictions effectively capture the variability in milk protein fractions (Figure 4A). On the other hand, the additive genetic, residual, and phenotypic variances for the FTIR-based predictions were considerably lower than for the laboratory measurements; specifically, between −6.62% (α_{S1}-CN) and 33.33% (α_{S2}-CN) for genetic variance, between −1.25% (α_{S1}-CN) and −29.17% (α_{S2}-CN) for residual variance, and between −3.01% (α_{S1}-CN) and −30.47% (α_{S2}-CN) for phenotypic variance (Figure 4A).

Table 5. The average and, in parentheses, the range of SD of genetic parameters estimates across the 5-folds from cross-validation for genetic parameters for milk protein fractions expressed as grams per liter for the gold-standard measurement (laboratory) and Fourier-transform infrared predicted[1]

		Genetic parameter			
Trait[2]		σ_a^2	σ_e^2	σ_p^2	h^2
TP	Laboratory	2.984 (0.4692–0.5044)	4.611 (0.6125–0.8598)	7.595 (0.3496–0.7235)	0.395 (0.0334–0.0409)
	Predicted	2.764 (0.4189–0.4552)	4.476 (0.5039–0.6473)	7.244 (0.2497–0.4597)	0.383 (0.0286–0.0421)
Major casein					
α_{S1}-CN	Laboratory	0.529 (0.0067–0.0851)	0.401 (0.0115–0.0858)	0.931 (0.0091–0.0179)	0.569 (0.0093–0.0292)
	Predicted	0.494 (0.0045–0.0408)	0.409 (0.0089–0.0384)	0.903 (0.0143–0.0561)	0.547 (0.0083–0.0324)
α_{S2}-CN	Laboratory	0.174 (0.025 0–0.0349)	0.384 (0.0409–0.0449)	0.558 (0.0283–0.0351)	0.313 (0.0496–0.0731)
	Predicted	0.116 (0.0105–0.0160)	0.272 (0.0221–0.0289)	0.388 (0.0162–0.0260)	0.302 (0.0292–0.0480)
β-CN	Laboratory	0.626 (0.0801–0.0962)	0.388 (0.0953–0.1356)	1.014 (0.0542–0.0962)	0.621 (0.0358–0.0707)
	Predicted	0.510 (0.0575–0.0735)	0.333 (0.0673–0.1084)	0.843 (0.0351–0.0796)	0.609 (0.0348–0.0597)
κ-CN	Laboratory	0.251 (0.0384–0.0438)	0.226 (0.0433–0.0605)	0.477 (0.0202–0.0431)	0.529 (0.0427–0.0655)
	Predicted	0.192 (0.0259–0.0321)	0.188 (0.0319–0.0451)	0.380 (0.0101–0.0319)	0.517 (0.0288–0.0426)
Glyco-κ-CN	Laboratory	0.125 (0.0192–0.0219)	0.251 (0.0311–0.0455)	0.376 (0.0219–0.0401)	0.337 (0.0332–0.0522)
	Predicted	0.096 (0.0129–0.0163)	0.202 (0.0201–0.0341)	0.298 (0.0152–0.0301)	0.327 (0.0479–0.0657)
TCN	Laboratory	1.961 (0.2852–0.3421)	2.965 (0.3311–0.4566)	4.926 (0.0251–0.3212)	0.399 (0.0355–0.0524)
	Predicted	1.699 (0.2429–0.2916)	2.755 (0.2862–0.4095)	4.453 (0.1950–0.2911)	0.382 (0.0306–0.0409)
Whey protein					
α-LA	Laboratory	0.011 (0.0008–0.0015)	0.031 (0.0013–0.0019)	0.042 (0.0010–0.0011)	0.262 (0.0268–0.0365)
	Predicted	0.008 (0.0004–0.0008)	0.025 (0.0009–0.0044)	0.033 (0.0013–0.0044)	0.243 (0.0215–0.0338)
β-LG	Laboratory	0.198 (0.0259–0.0298)	0.168 (0.0363–0.0487)	0.366 (0.0255–0.0401)	0.546 (0.0402–0.0454)
	Predicted	0.149 (0.0182–0.0205)	0.135 (0.0271–0.0364)	0.284 (0.0190–0.0296)	0.530 (0.0289–0.0578)
TWP	Laboratory	0.182 (0.0274–0.0332)	0.208 (0.0372–0.0475)	0.390 (0.0251–0.0351)	0.470 (0.0359–0.0493)
	Predicted	0.147 (0.0176–0.0259)	0.179 (0.0277–0.0361)	0.326 (0.0189–0.0252)	0.452 (0.0284–0.0489)

[1] For more details, see Supplemental Table S2 (https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023).

[2] TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein. σ_a^2 = genetic variance; σ_e^2 = residual variance; σ_p^2 = phenotypic variance.

For the milk protein fractions expressed as % N, the heritability estimates observed for α-LA (h^2 = 0.266), α_{S2}-CN (h^2 = 0.293), and TCN (h^2 = 0.375) were moderate, whereas those observed for TP, α_{S1}-CN, β-CN, κ-CN, glyco-κ-CN, β-LG, and TWP were high, with values ranging from 0.434 to 0.798 (Table 6). The heritability estimates for the FTIR predictions displayed the same trend as the laboratory measurements, although they were slightly lower (Figure 4B). The differences were smaller for α_{S2}-CN (−1.68%), glyco-κ-CN (−1.77%) and β-LG (−1.87%), and larger for α_{S1}-CN (−6.63%) and TWP (−786%; Figure 4B). However, we observed considerably smaller additive genetic, residual, and phenotypic variances in the FTIR-based predictions compared with the laboratory measurements (Figure 4B). The differences ranged from −7.72% (TP) to −41.83% (α_{S2}-CN) for genetic variance, from −1.75% (TWP) to −40.48% (α_{S2}-CN) for residual variance, and from −5.07% (TP) to −40.88% (α_{S2}-CN) for phenotypic variance (Figure 4B).

Genetic and Phenotypic Correlations Between Laboratory Measurements and FTIR Predictions

The estimated posterior densities of the genetic and phenotypic correlations between the laboratory measurements and FTIR-based predictions of milk proteins are reported in Figure 5 in g/L and Figure 6 as % N. For the protein fractions expressed in g/L, the averages of the posterior genetic correlations were high, ranging from 0.88 ± 0.033 for α-LA to 0.98 ± 0.005 for TP. The phenotypic correlations were lower, with values ranging between 0.64 ± 0.034 for α-LA and 0.86 ± 0.012 for TP (Figure 5). The posterior densities were skewed and their shape was similar across subsets of the data for genetic correlations, whereas slightly different densities were observed for the phenotypic correlations (Figure 5). The genetic correlations were high for the protein fractions expressed as % N, ranging from 0.87 ± 0.017 for α-LA to 0.97 ± 0.013 for TP (Figure 6). However, the phenotypic correlations were lower than 0.80, except for TP (0.81 ± 0.0212) and TCN (0.87 ± 0.0153; Figure 6). The posterior distributions of the genetic correlations were skewed with small differences across the subsets, whereas, in contrast, large differences were observed in the posterior distributions of the phenotypic correlations (Figure 6).

DISCUSSION

FTIR Predictive Ability

The magnitude of the predictive ability of FTIR determines its effectiveness for farm management and

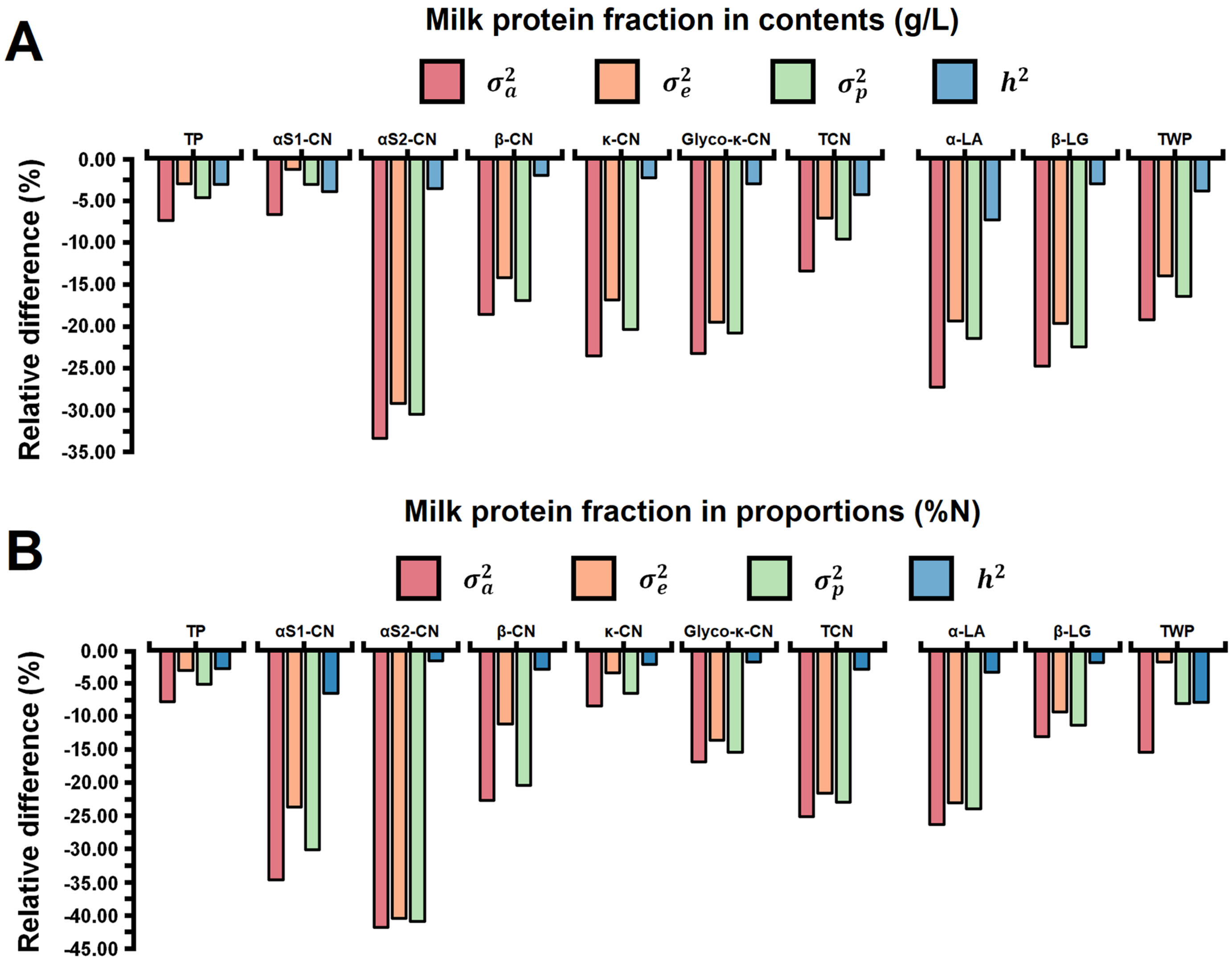

Figure 4. Relative difference for genetic $\left(\sigma_a^2\right)$, residual $\left(\sigma_r^2\right)$, and phenotypic $\left(\sigma_p^2\right)$ variance estimates and heritability (h^2) for Fourier-transform infrared (FTIR) prediction and gold standard measurement of milk proteins expressed as g/L (A) and % N (B). The relative difference (%) was calculated as $[(genpar_{pred} - genpar_{meas})/genpar_{meas}] \times 100$, where $genpar_{pred}$ and $genpar_{meas}$ are the genetic parameters (σ_a^2, σ_r^2, σ_p^2, and h^2) for FTIR predicted and measured milk protein fractions, respectively. TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein.

breeding purposes. Traditionally, FTIR predictive ability is assessed by phenotyping a small number of animals using gold-standard measurements, which results in under-optimistic evaluations of complex phenotypes. Mota et al. (2021a) evaluated CRV performed on a training set comprising specialized and dual-purpose breeds to ensure a sufficiently large population size and observed improvements in predictive ability over a single breed population. In the present study, we evaluated different CRV scenarios and a multibreed population (specialized and dual-purpose breeds) to evaluate the prediction performance of the model for milk protein fractions. Comparing the performances of the different CRV scenarios, smaller reductions in the predictive ability were observed as the independence between the training and validation sets increased. These reductions in the coefficient of determination ($\mathbf{R^2}$) values were around −9.01% (−4.76 to −19.75%) and −10.97% (−6.25 to −15.79%) for F/B, −4.92% (−1.39 to −12.66%) and −6.55% (−2.94 to −10.42%) for herd/date-out, and −2.80% (−1.14 to −6.17%) and −1.54 (−8.33 to 4.71%) for CRV-gen, for milk protein fractions expressed in g/L (Table 3) and % N (Table 4), respectively.

Overall, milk protein fractions expressed in g/L had higher predictive ability across the CRV scenarios.

Table 6. The average and, in parentheses, the range of SD of genetic parameter estimates across the 5-folds from cross-validation for genetic parameters for milk protein fractions expressed as a percentage of nitrogen for the gold-standard measurement (laboratory) and Fourier-transform infrared predicted[1]

Trait[2]		Genetic parameter[3]			
		σ_a^2	σ_e^2	σ_p^2	h^2
TP	Laboratory	1.295 (0.2045–0.2246)	1.701 (0.2996–0.3357)	2.996 (0.2103–0.2496)	0.434 (0.0759–0.0861)
	Predicted	1.195 (0.1803–0.1979)	1.649 (0.2355–0.2738)	2.845 (0.1517–0.1906)	0.420 (0.0699–0.0753)
Major casein					
α_{S1}-CN	Laboratory	0.447 (0.0795–0.0867)	0.313 (0.0639–0.1217)	0.760 (0.0329–0.0869)	0.588 (0.0411–0.0543)
	Predicted	0.292 (0.0333–0.0374)	0.239 (0.0493–0.0829)	0.531 (0.0349–0.0762)	0.549 (0.0690–0.0929)
α_{S2}-CN	Laboratory	0.619 (0.0699–0.0814)	1.459 (0.0986–0.1139)	2.078 (0.0701–0.0796)	0.298 (0.0317–0.0386)
	Predicted	0.360 (0.0164–0.0199)	0.868 (0.0229–0.0408)	1.229 (0.0166–0.0356)	0.293 (0.0285–0.0357)
β-CN	Laboratory	3.428 (0.4608–0.4943)	0.866 (0.5641–0.6252)	4.294 (0.4602–0.5012)	0.798 (0.0297–0.0558)
	Predicted	2.651 (0.3022–0.3761)	0.770 (0.3924–0.4678)	3.421 (0.2987–0.3605)	0.775 (0.0225–0.0445)
κ-CN	Laboratory	1.229 (0.2039–0.2374)	0.773 (0.2809–0.3336)	2.002 (0.2004–0.2395)	0.614 (0.0221–0.0395)
	Predicted	1.126 (0.1678–0.2121)	0.747 (0.2369–0.2968)	1.873 (0.1704–0.2108)	0.601 (0.0149–0.0293)
Glyco-κ-CN	Laboratory	0.956 (0.2057–0.2604)	0.739 (0.1851–0.2941)	1.695 (0.0830–0.1091)	0.564 (0.0531–0.0740)
	Predicted	0.795 (0.1397–0.1693)	0.639 (0.1587–0.2329)	1.434 (0.0637–0.1603)	0.554 (0.0355–0.0781)
TCN	Laboratory	0.418 (0.0811–0.0976)	0.664 (0.0907–0.1072)	1.082 (0.0321–0.0509)	0.386 (0.0798–0.0872)
	Predicted	0.313 (0.0292–0.0349)	0.521 (0.0353–0.0487)	0.834 (0.0199–0.0351)	0.375 (0.0591–0.0662)
Whey protein					
α-LA	Laboratory	0.160 (0.0095–0.0129)	0.422 (0.0287–0.0699)	0.582 (0.0013–0.007)	0.275 (0.0221–0.0897)
	Predicted	0.118 (0.0123–0.0145)	0.325 (0.0140–0.0209)	0.443 (0.0075–0.0145)	0.266 (0.0324–0.0372)
β-LG	Laboratory	0.546 (0.0948–0.1068)	0.472 (0.1029–0.1329)	1.018 (0.0220–0.0930)	0.536 (0.0458–0.0810)
	Predicted	0.475 (0.0506–0.0655)	0.428 (0.0614–0.1285)	0.903 (0.0191–0.1101)	0.526 (0.0619–0.0871)
TWP	Laboratory	1.403 (0.2366–0.2824)	1.658 (0.2743–0.3399)	3.061 (0.0941–0.1898)	0.458 (0.0762–0.0836)
	Predicted	1.187 (0.1727–0.2150)	1.629 (0.1875–0.2621)	2.816 (0.0733–0.1604)	0.422 (0.0703–0.0799)

[1] For more details, see Supplemental Table S3 (https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023).

[2] TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein.

[3] σ_a^2 = genetic variance; σ_e^2 = residual variance; σ_p^2 = phenotypic variance.

However, we observed a greater reduction in R^2 with the F/B scenario compared with a random 10-fold CRV, from −19.8 to −4.8% in g/L and from −6.3 to 15.8% as % N, which could be ascribed to differences in FTIR acquisition over time (i.e., oldest, 2013 vs. newest, 2019–2020), this being an extreme case of independence between the training and validation sets. In addition, FTIR measurements over time can present variations in the interferometer signal leading to changes in the vibrational bands caused by altered shapes, intensities, and relative intensities (Pelletier, 2003), which reduce the prediction accuracy, mainly for more complex milk components, such as fatty acids (Bonfatti et al., 2017a). Nieuwoudt et al. (2021) evaluated the day-to-day variation in FTIR spectra and observed a significant effect on accuracy; they used variance-simultaneous component analysis to monitor spectral variation, which allowed them to correct shifts in peak intensity or band shape, which would reduce predictive ability. Pretreatments for spectral noise reduction are very common and often important for obtaining robust predictive models, mainly the Savitzky–Golay smoothing algorithm used to attenuate high-frequency signals coming from noise and tends to retain important chemical signals (Savitzky and Golay, 1964). The principal component analysis of the milk FTIR spectra is useful for detecting possible differences in spectra values over time and using a noise reduction strategy to remove these differences across files. When the principal component analysis indicates a dissimilarity across FTIR information, increase the distance between structural relationships between variables and find potential clusters affecting the predictive model ability biases due to differences in baseline absorbance. However, we found no significant differences between the old and new FTIR data sets ($P > 0.05$), which did not contribute to biased or lower FTIR predictions.

The FTIR-based predictive abilities for milk proteins expressed as % N ranged from 0.38 to 0.86 for random 10-fold CRV, 0.35 to 0.79 for herd/date-out, 0.32 to 0.73 for F/B, and 0.38 to 0.81 for CRV-gen, and R^2 were higher than those previously obtained using different statistical approaches, which were in the ranges 0.14 to 0.82 (Baba et al., 2021), 0.18 to 0.28 (Rutten et al., 2011), and 0.13 to 0.36 (Bonfatti et al., 2011). Higher predictive abilities were obtained in the case of TP, TCN, and TWP in g/L and % N compared with the other traits, which might be due to their higher concentrations in milk (Table 3). The lower predictive ability observed for milk protein fractions expressed as % N compared with g/L agrees with previous results (Bonfatti et al., 2011). This suggests that FTIR infor-

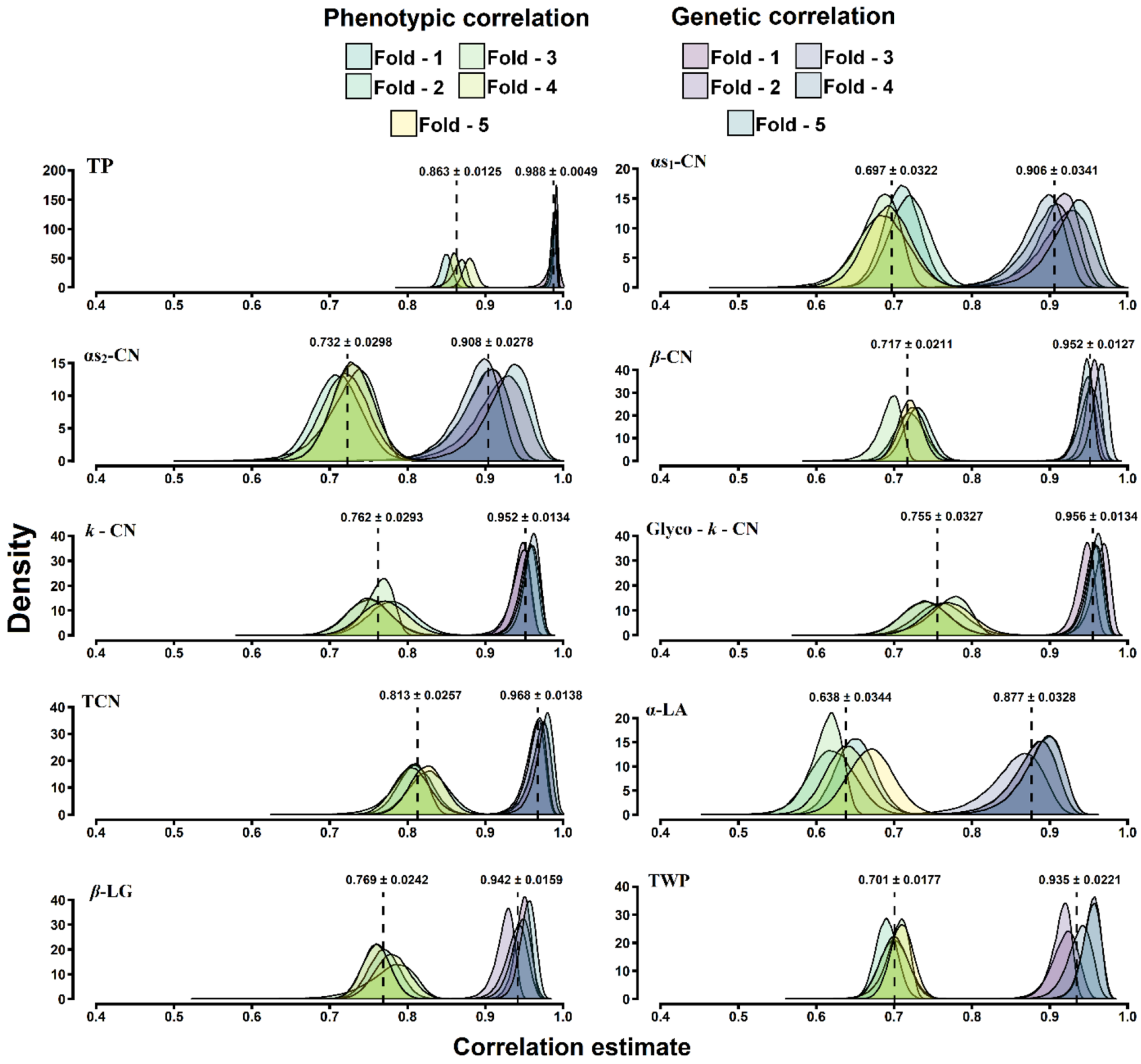

Figure 5. Posterior distribution of genetic and phenotypic correlation between gold standard and Fourier-transform infrared (FTIR) prediction using the cross-validation for genetic parameter scenarios for milk protein fraction expressed in grams per liter of milk. TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein.

mation can capture different biological data related to variations in milk protein fractions according to how these traits are expressed.

For practical application in breeding, the usefulness of FTIR predictions as potential indicator traits for genetic evaluation rests on obtaining FTIR-predicted values and gold standard measurements from a large number of samples. The CRV-gen scenario we devised to make FTIR predictions for genetic analyses was performed on large training (n = 1,253 cows) and validation sets (n = 1,184 cows) and gave predictive ability values ranging from moderate to high (Tables 3 and 4). Similarly, Rutten et al. (2010) demonstrated that assembling large reference populations makes it possible to improve the accuracy of FTIR-based predictions intended for estimating genetic parameters. We obtained moderate to high FTIR-based predictive abilities (R^2) for milk protein fractions expressed in both g/L and %

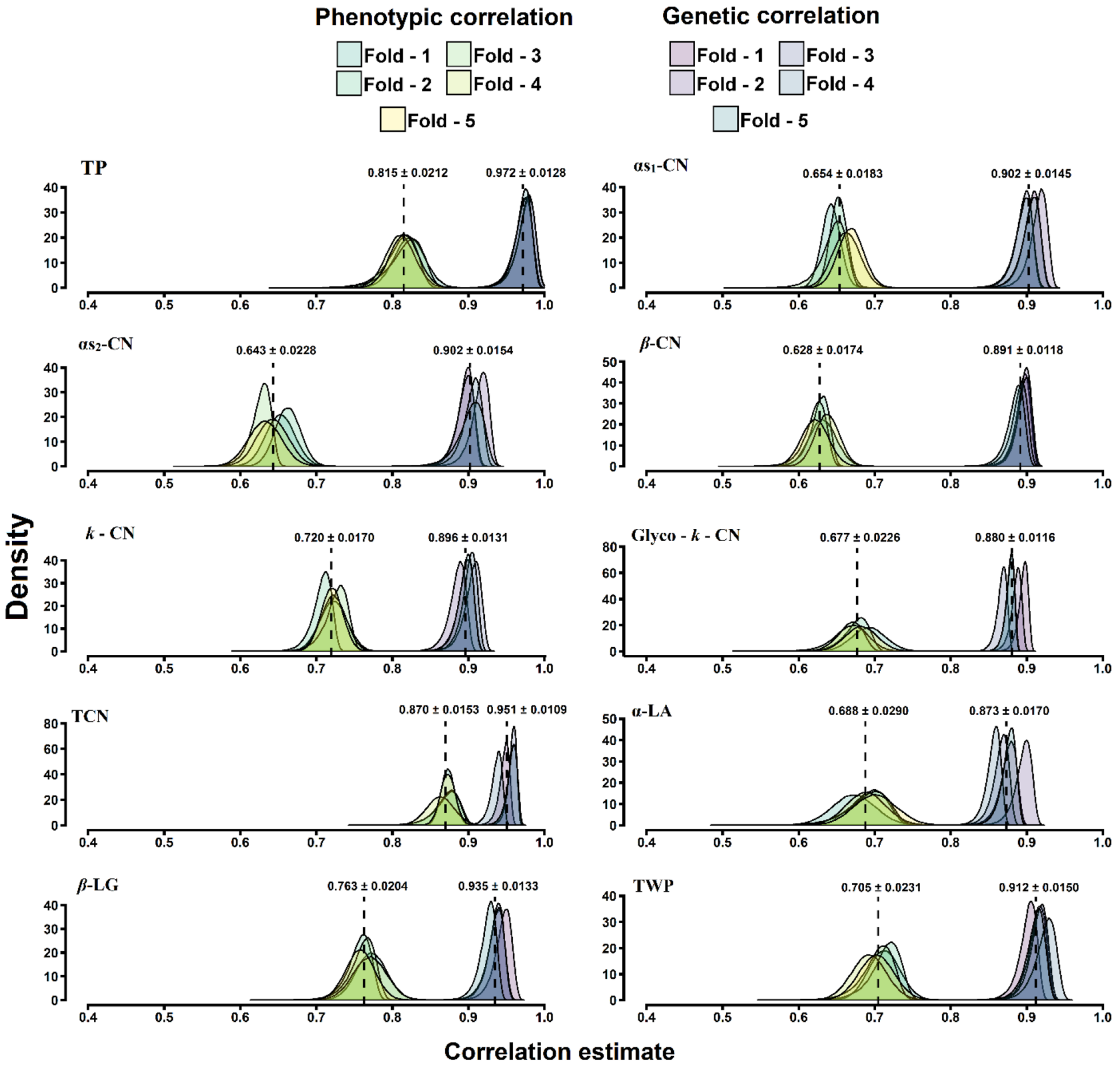

Figure 6. Posterior distribution of genetic and phenotypic correlation between gold standard and Fourier-transform infrared (FTIR) prediction for milk protein fraction expressed as the percentage of the total milk nitrogen content (% N). TP = true protein; glyco-κ-CN = glycosylated-κ-CN; TCN = total casein; TWP = total of whey protein.

N, indicating their potential use for breeding purposes. Although Soyeurt et al. (2011) suggested that an R^2 higher than 0.75 is required for use in animal breeding programs, Poulsen et al. (2014) observed that moderate predictive ability also provides valuable information for breeding programs. In this case, when FTIR predictive ability is moderate, the breeding value of a bull based on information from many progenies allows noise prediction correction. The greater predictive ability obtained may be due to the GBM selecting the milk spectra that can capture greater variability in milk chemical composition (Figures 2 and 3) and by their flexibility in mapping the complex associations between predictors and target phenotypes (Friedman, 2002; Azodi et al., 2020). Mota et al. (2021b), comparing machine learning and penalized regression against PLS regression, observed a superior ability of GBM to predict difficult-to-measure milk traits. A similar pattern was also found in this study in which GBM showed superiority against PLS, with R^2 increasing from 2 to 49% for protein fractions (Supplemental Table S1, https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023).

Associations Between Milk FTIR Wavelength Absorbances and Milk Protein Fractions

The FTIR wavelength absorbance is characterized by the effect of electromagnetic radiation, which is correlated with the stretching and bending vibrations of specific chemical bonds within a molecule (Karoui et al., 2010). The number of spectral regions associated with milk protein fractions varies according to how the proteins are expressed: 3 main regions for g/L and 4 or 5 for % N. Consequently, the biological background of milk protein fractions expressed as % N is more complex and requires more wavelength regions for its prediction. In addition, milk protein fractions expressed in g/L and % N shared 3 wavelength regions (3,200–2,900 cm^{-1}, 1,750–1,500 cm^{-1}, and 1,250–950 cm^{-1}), which are related to the fingerprint region (C–O, C–C, C=C, C–H, N–O, C–N, $C{=}CH_2$, O–H, amide II, and amide III bands), corresponding to common chemical bonds present in milk components such as fat, protein, lactose, carbohydrates, and organic acids (Soyeurt et al., 2010; Bittante and Cecchinato, 2013; Wang et al., 2016; Zaalberg et al., 2019). In particular, casein profiles are expected to be associated with absorption peaks related to the wavenumbers 1,250 cm^{-1} (amide III), 1,550 cm^{-1} (amide II), and 1,650 cm^{-1} (amide I; Osborne, 2000). However, Wang et al. (2016) found a significant association between *CSN3* and the wavenumbers around 1,269 and 1,550 cm^{-1}.

The infrared band amides I, II, and III are frequently used to assess milk protein contents (Etzion et al., 2004). However, vibrations on the water wavelengths related to the O–H groups are sensitive to interactions between water and the polar lipids and proteins present in milk, affecting the contribution of water to spectrum variability (Dousseau and Pézolet, 1990). These regions, mapping on 4,900 to 4,650 cm^{-1}, 3,600 to 3,350 cm^{-1}, and 2,550 to 2,400 cm^{-1}, explained the significant effect on milk protein fractions when expressed as % N. Wavelength regions 4,500 to 5,000 cm^{-1} contribute to vibrations of the N–H and C = O groups in the proteins (Subramanian and Rodriguez-Saona, 2009), and the genes *DGAT1* and *CSN3* are significantly associated with this region (Wang et al., 2016). The wavelength region 3,600 to 3,350 cm^{-1} consists of absorbance from stretching vibrations of hydroxyl groups (O–H) and amide A of proteins (N–H). Overall, the wavenumbers are known to contain information on milk components, and statistical approaches that can perform variable selection (GBM) have the advantage of being able to map the complex associations (e.g., nonlinear and interactions) between the FTIR wavelengths and the target trait (Natekin and Knoll, 2013).

Genetic Parameters for Laboratory Measurements and FTIR-Based Predictions of Milk Proteins

Phenotyping milk protein fraction is still a bottleneck, so techniques for precisely and reliably recording them are required to improve breeding program selection efficiency. Increasing the genetic gain rate using FTIR technologies can reduce the cost of measuring complex phenotypes on a large scale during different stages of lactation (Seidel et al., 2020). However, it is important to identify their genetic variations. The heritability estimates for milk protein fractions expressed in g/L and % N, assessed by gold standard laboratory measurements and FTIR-based predictions, show that the additive genetic effect influences them. However, notable reductions in the genetic parameters were observed for α_{S2}-CN and β-LG expressed in g/L and for α_{S1}-CN, and α_{S2}-CN expressed as % N. In contrast, the heritability estimates for α-LA and TCN in g/L, and TWP and α_{S1}-CN as % N were large. Our findings show that robust predictive models that include a larger number of samples in the training data set and a more complex algorithm may be able to capture the relationships between milk FTIR and milk protein fraction more accurately, corroborating the suitability of FTIR prediction of milk protein fractions for genetic evaluation purposes.

The observed reductions in the genetic parameters between FTIR-based predictions and gold standard measurements are consistent with results from previous studies (Cecchinato et al., 2009, 2020; Rutten et al., 2010). However, reductions in the heritability estimates observed in our study are smaller than those found by Cecchinato et al. (2020), ranging from −32 to −81%. This difference may be explained by the different abilities of the statistical models used to deal with complex associations between infrared spectra and the target phenotype in the calibration equations. Bonfatti et al. (2017b) estimated the genetic parameters for FTIR prediction of different milk-related traits and observed a significant association between predictive ability and reductions in the genetic parameters for FTIR-predicted traits. These reductions were smaller for traits predicted with an R^2 higher than 0.90 than those with an R^2 lower than 0.80.

Correlations Between Laboratory Measurements and FTIR Predictions of Milk Protein Fractions

The magnitude of the genetic correlations between FTIR-based predictions and gold-standard laboratory measures is the main parameter for determining the feasibility of including such indicator traits in animal selection for breeding purposes (Cecchinato et al., 2009, 2020; Rutten et al., 2011). Successful incorporation into a breeding program depends on the degree of genetic gain attained through indirect selection, which is directly related to the strength of the genetic correlation between the target and FTIR-predicted trait. Milk protein fractions are important for the dairy industry because they influence milk's technological properties, mainly during the coagulation process, whereby the milk protein fractions α_{S1}-CN and κ-CN lead to reductions in coagulation time (Amalfitano et al., 2019). Our estimates of the genetic correlations between the gold standard measurements and FTIR predictions were high and ranged from 0.87 to 0.99 (Figures 5 and 6). The strength of the genetic correlations varied as a function of predictive ability: as FTIR predictive ability increased, the genetic correlation between the predicted values and the gold standard measures also increased, as shown in Supplemental Figure S8A (https://doi.org/10.6084/m9.figshare.21864596.v1; Mota et al., 2023). In this regard, we obtained higher genetic correlations than in previous studies (Bonfatti et al., 2017b), especially for β-CN (0.95 vs. 0.63), α-LA (0.88 vs. 0.57), and β-LG (0.94 vs. 0.77). Slight differences in the genetic correlations were observed for TP (0.99 vs. 0.98), α_{S1}-CN (0.91 vs. 0.94), α_{S2}-CN (0.91 vs. 0.87), κ-CN (0.95 vs. 0.90), whereas no difference was observed for TCN. On the other hand, the genetic correlations for milk protein fractions expressed as % N ranged from 0.88 to 0.97, strikingly different from the results of previous studies, ranging from 0.23 to 0.90 (Bonfatti et al., 2017b; Cecchinato et al., 2020). Furthermore, although we observed moderate predictive ability for α_{S1}-CN, α_{S2}-CN, β-CN, κ-CN, Glyco-κ-CN, α-LA, β-LG, and TWP expressed as % N, the genetic correlation between the gold standard measures and the FTIR predictions was greater than 0.80. This high correlation indicates that there is little or no reranking of the animals concerning their expected breeding value according to gold-standard measurements. Rutten et al. (2011) assessed the genetic parameters of FTIR predictions and milk proteins expressed as % N and found predictive ability to vary from 0.18 (α_{S1}-CN) to 0.56 (β-LG), resulting in genetic correlations ranging from 0.62 (β-CN) to 0.97 (TWP), good enough for exploitation in breeding programs. Concerning milk technological traits, Cecchinato et al. (2009) found R^2 values from 0.46 to 0.52 for infrared predictions of curd firming, with genetic correlations between the measures and the predictions ranging from 0.71 to 0.87, and R^2 values from 0.61 to 0.69 for infrared predictions of coagulation time, with genetic correlations ranging from 0.91 to 0.96.

The phenotypic correlations between the gold standard measurements and FTIR predictions of milk protein fractions expressed in g/L and % N ranged from 0.63 to 0.87 and were dependent on FTIR predictive ability (Supplemental Figure S8B). Differences in the association between the phenotypic correlations and the model's predictive ability according to whether milk proteins were expressed in g/L or % N can be explained by differences in the contributions of the genetic and environmental effects. The same trend has been observed in dairy cattle (Rutten et al., 2010; Bonfatti et al., 2017b) and beef cattle (Cecchinato et al., 2011; Savoia et al., 2021), where the genetic correlations between the infrared predictions and measured traits were less dependent on predictive ability than the phenotypic correlations. Milk protein fractions with the highest predictive abilities in the calibration equation exhibited the highest genetic and phenotypic correlations with the relative gold standard measurement. However, a low to moderate R^2 can also give rise to acceptable genetic and phenotypic correlations. Therefore, our results support for the potential application of the developed prediction equation for breeding purposes to enhance milk quality and cheesemaking aptitude.

CONCLUSIONS

This study showed that FTIR spectra can be successfully exploited for the prediction of milk protein fractions expressed both in g/L and % N, although the predictions were in general more reliable when proteins were expressed in g/L, as in % N requires more FTIR wavelengths to capture the phenotypic variability. Similar regions of the FTIR spectra were found to explain the variability of traits expressed in g/L and in % N, confirming that they share the same biological background. The heritability estimates for milk protein fractions assessed by laboratory measurements and FTIR predictions followed the same trend with slight differences among them. The high genetic correlations between the FTIR predictions and the laboratory measurements found in our study provide evidence for their potential use as indicator traits in breeding programs aimed at altering protein fractions and improving milk quality and cheesemaking ability. Further studies could be conducted applying the FTIR calibrations on a population database, provided that FTIR spectra are available, and estimating genetic parameters and genomic breeding values exploiting longitudinal data and random regression models.

ACKNOWLEDGMENTS

The research was part of the projects (1) LATSAN funded by the Ministero delle politiche agricole alimentari, forestali e del turismo (MIPAAF, Rome, Italy), (2) BENELAT, Interventi a breve e lungo termine per il miglioramento del benessere, dell'efficienza e della qualità delle produzioni dei bovini da latte della Lombardia, Bando per il finanziamento di progetti di ricerca in campo agricolo e forestale 2018 (d.d.s. 28 marzo 2018, n. 4403), and (3) AGER 2 Project under Grant number 2017-1130 funded by the Fondazione CARIPLO. The authors are also grateful to the Italian Holstein-Friesian and Jersey Cattle Breeders Association (ANAFIJ, Cremona, Italy) for collaborating in the research activities and for providing pedigree information. The authors have not stated any conflicts of interest.

REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743–752. https://doi.org/10.3168/jds.2009-2730.

Amalfitano, N., C. Cipolat-Gotet, A. Cecchinato, M. Malacarne, A. Summer, and G. Bittante. 2019. Milk protein fractions strongly affect the patterns of coagulation, curd firming, and syneresis. J. Dairy Sci. 102:2903–2917. https://doi.org/10.3168/jds.2018-15524.

Azodi, C. B., J. Tang, and S. H. Shiu. 2020. Opening the black box: Interpretable machine learning for geneticists. Trends Genet. 36:442–455. https://doi.org/10.1016/j.tig.2020.03.005.

Baba, T., S. Pegolo, L. F. M. Mota, F. Peñagaricano, G. Bittante, A. Cecchinato, and G. Morota. 2021. Integrating genomic and infrared spectral data improves the prediction of milk protein composition in dairy cattle. Genet. Sel. Evol. 53:29. https://doi.org/10.1186/s12711-021-00620-7.

Bisutti, V., S. Pegolo, D. Giannuzzi, L. F. M. Mota, A. Vanzin, A. Toscano, E. Trevisi, P. A. Marsan, M. Brasca, and A. Cecchinato. 2022. The β-casein (*CSN2*) A2 allelic variant alters milk protein profile and slightly worsens coagulation properties in Holstein cows. J. Dairy Sci. 105:3794–3809. https://doi.org/10.3168/JDS.2021-21537.

Bittante, G., and A. Cecchinato. 2013. Genetic analysis of the Fourier-transform infrared spectra of bovine milk with emphasis on individual wavelengths related to specific chemical bonds. J. Dairy Sci. 96:5991–6006. https://doi.org/10.3168/jds.2013-6583.

Bittante, G., B. Contiero, and A. Cecchinato. 2013. Prolonged observation and modelling of milk coagulation, curd firming, and syneresis. Int. Dairy J. 29:115–123. https://doi.org/10.1016/j.idairyj.2012.10.007.

Bonfatti, V., G. Di Martino, and P. Carnier. 2011. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows. J. Dairy Sci. 94:5776–5785. https://doi.org/10.3168/jds.2011-4401.

Bonfatti, V., A. Fleming, A. Koeck, and F. Miglior. 2017a. Standardization of milk infrared spectra for the retroactive application of calibration models. J. Dairy Sci. 100:2032–2041. https://doi.org/10.3168/jds.2016-11837.

Bonfatti, V., D. Vicario, A. Lugo, and P. Carnier. 2017b. Genetic parameters of measures and population-wide infrared predictions of 92 traits describing the fine composition and technological properties of milk in Italian Simmental cattle. J. Dairy Sci. 100:5526–5540. https://doi.org/10.3168/jds.2016-11667.

Cecchinato, A., A. Albera, C. Cipolat-Gotet, A. Ferragina, and G. Bittante. 2015. Genetic parameters of cheese yield and curd nutrient recovery or whey loss traits predicted using Fourier-transform infrared spectroscopy of samples collected during milk recording on Holstein, Brown Swiss, and Simmental dairy cows. J. Dairy Sci. 98:4914–4927. https://doi.org/10.3168/jds.2014-8599.

Cecchinato, A., C. Cipolat-Gotet, J. Casellas, M. Penasa, A. Rossoni, and G. Bittante. 2013. Genetic analysis of rennet coagulation time, curd-firming rate, and curd firmness assessed over an extended testing period using mechanical and near-infrared instruments. J. Dairy Sci. 96:50–62. https://doi.org/10.3168/jds.2012-5784.

Cecchinato, A., M. De Marchi, L. Gallo, G. Bittante, and P. Carnier. 2009. Mid-infrared spectroscopy predictions as indicator traits in breeding programs for enhanced coagulation properties of milk. J. Dairy Sci. 92:5304–5313. https://doi.org/10.3168/jds.2009-2246.

Cecchinato, A., M. De Marchi, M. Penasa, A. Albera, and G. Bittante. 2011. Near-infrared reflectance spectroscopy predictions as indicator traits in breeding programs for enhanced beef quality. J. Anim. Sci. 89:2687–2695. https://doi.org/10.2527/jas.2010-3740.

Cecchinato, A., H. Toledo-Alvarado, S. Pegolo, A. Rossoni, E. Santus, C. Maltecca, G. Bittante, and F. Tiezzi. 2020. Integration of wet-lab measures, milk infrared spectra, and genomics to improve difficult-to-measure traits in dairy cattle populations. Front. Genet. 11:563393. https://doi.org/10.3389/fgene.2020.563393.

Dousseau, F., and M. Pézolet. 1990. Determination of the secondary structure content of proteins in aqueous solutions from their amide I and amide II infrared bands. Comparison between classical and partial least-squares methods. Biochemistry 29:8771–8779. https://doi.org/10.1021/bi00489a038.

Etzion, Y., R. Linker, U. Cogan, and I. Shmulevich. 2004. Determination of protein concentration in raw milk by mid-infrared Fourier transform infrared/attenuated total reflectance spectroscopy. J. Dairy Sci. 87:2779–2788. https://doi.org/10.3168/jds.S0022-0302(04)73405-0.

Friedman, J. H. 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38:367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. 4th ed. J. M. Bernado, J. O. Berger, A. P. Smith, and A. F. M. Dawid, ed. Clarendon Press.

Grelet, C., C. Bastin, M. Gelé, J. B. Davière, M. Johan, A. Werner, R. Reding, J. A. Fernandez Pierna, F. G. Colinet, P. Dardenne, N. Gengler, H. Soyeurt, and F. Dehareng. 2016. Development of Fourier transform mid-infrared calibrations to predict acetone, β-hydroxybutyrate, and citrate contents in bovine milk through a European dairy network. J. Dairy Sci. 99:4816–4825. https://doi.org/10.3168/jds.2015-10477.

Grelet, C., J. A. Fernández Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. J. Dairy Sci. 98:2150–2160. https://doi.org/10.3168/jds.2014-8764.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning. 2nd ed. Springer Series in Statistics. Springer.

Karoui, R., G. Downey, and C. Blecker. 2010. Mid-Infrared spectroscopy coupled with chemometrics: A tool for the analysis of intact food systems and the exploration of their molecular structure−Quality relationships—A review. Chem. Rev. 110:6144–6168. https://doi.org/10.1021/cr100090k.

Maurmayr, A., A. Cecchinato, L. Grigoletto, and G. Bittante. 2013. Detection and quantification of αS1-, αS2-, β-, κ-casein, α-lactalbumin, β-lactoglobulin and lactoferrin in bovine milk by reverse-phase high- performance liquid chromatography. ACS Agric. Conspec. Sci. 78:201–205.

Mevik, B.-H., and R. Wehrens. 2007. The pls package: Principal component and partial least squares regression in R. J. Stat. Softw. 18. https://doi.org/10.18637/jss.v018.i02.

Mota, L., V. Bisutti, A. Vanzin, S. Pegolo, A. Toscano, S. Schiavon, F. Tagliapietra, L. Gallo, P. Ajmone Marsan, and A. Cecchinato. 2023. Predicting milk protein fraction using infrared spectroscopy and a gradient boosting machine for breeding purposes in Holstein cattle. Figshare. Figure. https://doi.org/10.6084/m9.figshare.21864596.v1.

Mota, L. F. M., D. Giannuzzi, V. Bisutti, S. Pegolo, E. Trevisi, S. Schiavon, L. Gallo, D. Fineboym, G. Katz, and A. Cecchinato. 2022. Real-time milk analysis integrated with stacking ensemble learning as a tool for the daily prediction of cheese-making traits in Holstein cattle. J. Dairy Sci. 105:4237–4255. https://doi.org/10.3168/jds.2021-21426.

Mota, L. F. M., S. Pegolo, T. Baba, G. Morota, F. Peñagaricano, G. Bittante, and A. Cecchinato. 2021a. Comparison of single-breed and multi-breed training populations for infrared predictions of novel phenotypes in Holstein cows. Animals (Basel) 11:1993. https://doi.org/10.3390/ani11071993.

Mota, L. F. M., S. Pegolo, T. Baba, F. Peñagaricano, G. Morota, G. Bittante, and A. Cecchinato. 2021b. Evaluating the performance of machine learning methods and variable selection methods for predicting difficult-to-measure traits in Holstein dairy cattle using milk infrared spectral data. J. Dairy Sci. 104:8107–8121. https://doi.org/10.3168/jds.2020-19861.

Natekin, A., and A. Knoll. 2013. Gradient boosting machines, a tutorial. Front. Neurorobot. 7:21. https://doi.org/10.3389/fnbot.2013.00021.

Nieuwoudt, M. K., C. Giglio, F. Marini, G. Scott, and S. E. Holroyd. 2021. Routine monitoring of instrument stability in a milk testing laboratory with ASCA: A pilot study. Front Chem. 9:733331. https://doi.org/10.3389/fchem.2021.733331.

Osborne, B. G. 2000. Near-Infrared Spectroscopy in Food Analysis. John Wiley & Sons Ltd.

Pegolo, S., L. F. M. Mota, V. Bisutti, M. Martinez-Castillero, D. Giannuzzi, L. Gallo, S. Schiavon, F. Tagliapietra, A. Revello Chion, E. Trevisi, R. Negrini, P. Ajmone Marsan, and A. Cecchinato. 2021. Genetic parameters of differential somatic cell count, milk composition, and cheese-making traits measured and predicted using spectral data in Holstein cows. J. Dairy Sci. 104:10934–10949. https://doi.org/10.3168/jds.2021-20395.

Pelletier, M. J. 2003. Quantitative analysis using Raman spectrometry. Appl. Spectrosc. 57:20A–42A. https://doi.org/10.1366/000370203321165133.

Pérez, P., and G. de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–495. https://doi.org/10.1534/genetics.114.164442.

Poulsen, N. A., C. E. A. Eskildsen, T. Skov, L. B. Larsen, and A. J. Buitenhuis. 2014. Production Comparison of Genetic Parameters Estimation of Fatty Acids from Gas Chromatography and FT-IR in Holsteins. Proc. 10th World Congress of Genetics Applied to Livestock. American Society of Animal Science.

Rutten, M. J., H. Bovenhuis, J. M. L. M. L. Heck, and J. A. M. A. M. van Arendonk. 2011. Predicting bovine milk protein composition based on Fourier transform infrared spectra. J. Dairy Sci. 94:5683–5690. https://doi.org/10.3168/jds.2011-4520.

Rutten, M. J. M., H. Bovenhuis, and J. A. M. van Arendonk. 2010. The effect of the number of observations used for Fourier transform infrared model calibration for bovine milk fat composition on the estimated genetic parameters of the predicted data. J. Dairy Sci. 93:4872–4882. https://doi.org/10.3168/jds.2010-3157.

Sanchez, M. P., M. Ferrand, M. Gelé, D. Pourchet, G. Miranda, P. Martin, M. Brochard, and D. Boichard. 2017. Genetic parameters for milk protein composition predicted using mid-infrared spectroscopy in the French Montbéliarde, Normande, and Holstein dairy cattle breeds. J. Dairy Sci. 100:6371–6375. https://doi.org/10.3168/jds.2017-12663.

Savitzky, A., and M. J. E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36:1627–1639. https://doi.org/10.1021/ac60214a047.

Savoia, S., A. Albera, A. Brugiapaglia, L. Di Stasio, A. Cecchinato, and G. Bittante. 2021. Prediction of meat quality traits in the abattoir using portable near-infrared spectrometers: heritability of predicted traits and genetic correlations with laboratory-measured traits. J. Anim. Sci. Biotechnol. 12:29. https://doi.org/10.1186/s40104-021-00555-5.

Seidel, A., N. Krattenmacher, and G. Thaller. 2020. Dealing with complexity of new phenotypes in modern dairy cattle breeding. Anim. Front. 10:23–28. https://doi.org/10.1093/af/vfaa005.

Silva, S. V., and F. X. Malcata. 2005. Caseins as source of bioactive peptides. Int. Dairy J. 15:1–15. https://doi.org/10.1016/j.idairyj.2004.04.009.

Smith, B. J. 2007. boa: An R package for MCMC output convergence assessment and posterior inference. J. Stat. Softw. 21:1–37. https://doi.org/10.18637/jss.v021.i11.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 94:1657–1667. https://doi.org/10.3168/jds.2010-3408.

Soyeurt, H., C. Grelet, S. McParland, M. Calmels, M. Coffey, A. Tedde, P. Delhez, F. Dehareng, and N. Gengler. 2020. A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra. J. Dairy Sci. 103:11585–11596. https://doi.org/10.3168/jds.2020-18870.

Soyeurt, H., I. Misztal, and N. Gengler. 2010. Genetic variability of milk components based on mid-infrared spectral data. J. Dairy Sci. 93:1722–1728. https://doi.org/10.3168/jds.2009-2614.

Stocco, G., C. Cipolat-Gotet, T. Bobbo, A. Cecchinato, and G. Bittante. 2017. Breed of cow and herd productivity affect milk composition and modeling of coagulation, curd firming, and syneresis. J. Dairy Sci. 100:129–145. https://doi.org/10.3168/jds.2016-11662.

Subramanian, A., and L. Rodriguez-Saona. 2009. Fourier Transform Infrared (FTIR) Spectroscopy. Elsevier.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980.

Walstra, P. 1999. Casein sub-micelles: Do they exist? Int. Dairy J. 9:189–192. https://doi.org/10.1016/S0958-6946(99)00059-X.

Wang, Q., A. Hulzebosch, and H. Bovenhuis. 2016. Genetic and environmental variation in bovine milk infrared spectra. J. Dairy Sci. 99:6793–6803. https://doi.org/10.3168/jds.2015-10488.

Young, R. S. 1978. Calibration and standardization of the infrared milk analyzer. The California experience. J. Dairy Sci. 61:1279–1283. https://doi.org/10.3168/jds.S0022-0302(78)83718-7.

Zaalberg, R. M., N. A. Poulsen, H. Bovenhuis, J. Sehested, L. B. Larsen, and A. J. Buitenhuis. 2021. Genetic analysis on infrared-predicted milk minerals for Danish dairy cattle. J. Dairy Sci. 104:8947–8958. https://doi.org/10.3168/jds.2020-19638.

Zaalberg, R. M. M., N. Shetty, L. Janss, and A. J. J. Buitenhuis. 2019. Genetic analysis of Fourier transform infrared milk spectra in Danish Holstein and Danish Jersey. J. Dairy Sci. 102:503–510. https://doi.org/10.3168/jds.2018-14464.