Journal of Animal Science and Biotechnology

RESEARCH

Open Access

Combining genetic markers, on-farm information and infrared data for the in-line prediction of blood biomarkers of metabolic disorders in Holstein cattle

Lucio F. M. Mota[1], Diana Giannuzzi[1*], Sara Pegolo[1], Hugo Toledo-Alvarado[2], Stefano Schiavon[1], Luigi Gallo[1], Erminio Trevisi[3], Alon Arazi[4], Gil Katz[4], Guilherme J. M. Rosa[5] and Alessio Cecchinato[1]

Abstract

Background Various blood metabolites are known to be useful indicators of health status in dairy cattle, but their routine assessment is time-consuming, expensive, and stressful for the cows at the herd level. Thus, we evaluated the effectiveness of combining in-line near infrared (NIR) milk spectra with on-farm (days in milk [DIM] and parity) and genetic markers for predicting blood metabolites in Holstein cattle. Data were obtained from 388 Holstein cows from a farm with an AfiLab system. NIR spectra, on-farm information, and single nucleotide polymorphisms (SNP) markers were blended to develop calibration equations for blood metabolites using the elastic net (ENet) approach, considering 3 models: (1) Model 1 (M1) including only NIR information, (2) Model 2 (M2) with both NIR and on-farm information, and (3) Model 3 (M3) combining NIR, on-farm and genomic information. Dimension reduction was considered for M3 by preselecting SNP markers from genome-wide association study (GWAS) results.

Results Results indicate that M2 improved the predictive ability by an average of 19% for energy-related metabolites (glucose, cholesterol, NEFA, BHB, urea, and creatinine), 20% for liver function/hepatic damage, 7% for inflammation/innate immunity, 24% for oxidative stress metabolites, and 23% for minerals compared to M1. Meanwhile, M3 further enhanced the predictive ability by 34% for energy-related metabolites, 32% for liver function/hepatic damage, 22% for inflammation/innate immunity, 42.1% for oxidative stress metabolites, and 41% for minerals, compared to M1. We found improved predictive ability of M3 using selected SNP markers from GWAS results using a threshold of > 2.0 by 5% for energy-related metabolites, 9% for liver function/hepatic damage, 8% for inflammation/innate immunity, 22% for oxidative stress metabolites, and 9% for minerals. Slight reductions were observed for phosphorus (2%), ferric-reducing antioxidant power (1%), and glucose (3%). Furthermore, it was found that prediction accuracies are influenced by using more restrictive thresholds ($-\log_{10}$(*P*-value) > 2.5 and 3.0), with a lower increase in the predictive ability.

Conclusion Our results highlighted the potential of combining several sources of information, such as genetic markers, on-farm information, and in-line NIR infrared data improves the predictive ability of blood metabolites in dairy cattle, representing an effective strategy for large-scale in-line health monitoring in commercial herds.

Keywords Blood metabolites, Dairy cattle, Data integration, Feature selection, Metabolic disorders, NIR, Precision livestock farming

*Correspondence:
Diana Giannuzzi
diana.giannuzzi@unipd.it
Full list of author information is available at the end of the article

Background

The high energy requirements of milk production can lead to metabolic disorders in dairy cows during early lactation and throughout the lactation period [1]. Experiencing stress can trigger lipolysis and proteolysis to support high milk yields [2, 3]. As a result, metabolic stress leads to an increment in blood levels of haptoglobin, non-esterified fatty acids (NEFA), β-hydroxybutyrate (BHB), ceruloplasmin (CuCp), and globulins, followed by a reduction in glucose, paraoxonase (PON) and albumin levels [1]. Variations of blood minerals are also observed, which can be used as biological markers, specifically calcium as a primary indicator of milk fever, and sodium, potassium, and zinc as markers of systemic inflammation and oxidation. These alterations in blood metabolite levels are directly linked with the main metabolic disorders (energy imbalances, ketosis, and milk fever) and negatively impact dairy herd profitability and production by affecting milk production, reproductive efficiency, and overall herd health [4, 5]. Consequently, there has been a growing interest in addressing these metabolic issues during the lactation phase to improve the health and resilience of dairy cows.

Traditionally, the metabolic evaluation of a herd is monitored mainly through blood metabolic profiling [6], allowing the identification and selection of resilient cows less predisposed to developing metabolic disorders. However, its assessment at either an individual or herd level is expensive and time-consuming. Despite this, there has been increased attention on metabolic stress in lactating dairy cows due to its harmful effects on the profitability and sustainability of dairy herds [7]. In this context, noninvasive high-throughput phenotyping technologies based on milk infrared spectroscopy have been applied to predict variation in blood metabolites [8–11]. This is mainly possible because the detailed composition of raw milk reflects the health and nutritional status of dairy cows, and the disruption of metabolic homeostasis is reflected in alterations to these components [12, 13]. Milk spectral analysis is a promising method for assessing the metabolic status of dairy cows on a large scale due to the interaction between metabolic status and milk compounds, mainly fat and protein [14]. In this regard, automated milk quality sensors are used to check collected milk's quality and look for any health biomarkers in real-time at the herd level [10, 15]. These in-line near-infrared (NIR) milk sensors play a vital role in herd management technologies, especially in monitoring cows' nutrition and detecting metabolic alterations [16] by examining changes in milk composition over time. Giannuzzi et al. [10], with a first attempt using various machine learning methods, explored the possibility of predicting blood metabolic profile from the milk of individual cows using an in-line NIR spectroscopy milk analyzer, obtaining low to moderate predictions (from 0.30 to 0.65).

Considering the complex nature of metabolic stress during the lactation period, it is worthwhile to consider multiple sources of information. Integrating different layers of information has already been proposed to enhance the development of more accurate predictive models, increasing the capability to detect metabolic disturbances in dairy herds [8]. Early and accurate detection of cows prone to metabolic disorders is crucial to building strategies to support farm management and breeding decisions to detect metabolic disorders efficiently. Therefore, this study was conducted to assess the potential benefits of integrating in-line NIR milk sensor infrared information with on-farm data (DIM and parity) and genetic markers for predicting the blood metabolic profile in Holstein cattle.

Materials and methods

Field data

A total of 388 Holstein cows from a single herd in northern Italy (Piacenza province) were sampled for blood. These cows received a twice-daily feeding regimen consisting of a diet primarily composed of corn silage and sorghum. Energy-protein supplementation was provided following nutritional guidelines for dairy cattle [17]. The average values (± SD) were 32.3 ± 6.54 for daily milk yield (kg), 4.1 ± 0.36 for fat (%), and 3.7 ± 0.13 for protein (%). The cows had an average for days in milk (DIM) of 127.3 ± 60.22 (varying from 3 to 425) with a percentage of 55%, 42% and 3% at early, mid and late lactation, respectively. The percentages of 86%, 9%, and 5% were observed for 1st, 2nd, and from 3rd to 5th parity, respectively. Prior to collecting samples, a health assessment was performed, and any cows exhibiting clinical diseases or undergoing medical treatment were excluded from the study.

Blood sampling

Blood samples were collected in 7 batches (i.e., sampling date): 3 batches in 2019 (300 cows) and 4 batches in 2020 (88 cows). Each cow was sampled once ($n = 388$). Five milliliters of blood from each cow were collected after the morning milking and before feeding through jugular venipuncture using vacutainer tubes containing 150 USP units of lithium heparin as an anticoagulant (Vacumed; FL Medical, Torreglia, Padua, Italy). All blood samples were maintained on ice until 2 h after blood sampling, followed by centrifugation at 3,500 × g for 1 min at 6 °C (Hettich Universal 16R Centrifuge), and then the plasma samples obtained were collected and stored at −20 °C until the analysis.

Blood metabolic profile

Blood metabolites were analyzed using a clinical autoanalyzer (ILAB 650, Instrumentation Laboratory, Lexington, MA, USA) following methods proposed by Calamari et al. [18] and Hanasand et al. [19]. A complete metabolic profile was assessed covering energy-related metabolites (glucose, cholesterol, NEFA, BHB, urea, and creatinine), liver function/hepatic damage (aspartate aminotransferase [AST], γ-glutamyl transferase [GGT], total bilirubin [BILt], albumin, alkaline phosphatase [ALP], and paraoxonase [PON]), oxidative stress (total reactive oxygen metabolites [ROMt]; advanced oxidation protein products [AOPP]; ferric reducing antioxidant power [FRAP]; total thiol groups [SHp]), inflammation/innate immunity (CuCp, total proteins, globulins, haptoglobin, and myeloperoxidase), and minerals (Ca, P, Mg, Na, K, Cl and Zn). Kits from Instrumentation Laboratory (IL Test) were utilized to measure glucose, total proteins, albumin, haptoglobin, urea, Ca, AST, and GGT levels. Globulin concentration was estimated as the difference between total proteins and albumin, and potassium electrolytes (K^+) were assessed using the potentiometer method (Ion Selective Electrode coupled to ILAB 600). Zn, NEFA, BHB, and CuCp were analyzed using the methods reported by Calamari et al. [18]. The concentrations of AOPP, ROMt, FRAP, and PON were determined according to Premi et al. [20].

AfiLab equipment and near-infrared spectra storage

The AfiLab system is a spectrometer that uses a set of 32 discreet frequencies of light sources in the visible-NIR regimen (400–1,000 nm) based on light-emitting diodes as described by Schmilovitch et al. [21] and gives accurate estimates for fat in the range of 2% to 6% and for protein ranging from 2% to 5% (Afimilk, Israel, internal control) and for cheese-making traits [15]. During milking, the AfiLab system measures milk spectra on every 200 mL of milk flowing through the machine and reports an average of approximately 70 observations per cow in each milking session (~15 kg). Each observation is weighed with respect to its milk quantity (~0.20 to 0.33 mg). In addition, the AfiLab infrared information was zero-set calibrated once a month between the morning and afternoon milking sessions to eliminate possible bias as part of routine maintenance.

The AfiLab milk spectra and on-farm information from Afimilk system were collected concomitantly with the blood sampling. The AfiLab milk spectra were preprocessed considering the first derivative, estimated as the difference between consecutive NIR spectra information (x_i) ($x'_i = x_i - x_{i-1}$). The first derivative was then normalized using a Standard Normal Variate equation $[SVN_i = (x'_i - \overline{x}'_i)/s_{x'_i}]$, where x'_i is the first derivative of spectrum i, $\overline{x}'_i$ represent the mean of the first derivative for spectrum i, and $s_{x'_i} = \sqrt{\frac{\sum(x'_i - \overline{x}'_i)^2}{m}}$ is the is the standard deviation for first derivative for spectrum i and m is the number of cows. The quality control of milk spectra was assessed by principal component analysis combined with Mahalanobis distance at a probability level < 0.05 [22]; after this data processing, four animals were removed from the analysis.

Genomic data

All 388 cows were genotyped with the Geneseek Genomic Profiler Bovine 100k SNP Chip assay. The quality control was performed by removing the non-autosomal regions and autosomal SNP markers with a minor allele frequency of less than 0.05 and a significant deviation from Hardy–Weinberg equilibrium ($P \leq 10^{-5}$). Markers and samples with call rate lower than 0.95 were also removed. After spectra and genomic quality control, 380 cows with information for NIR AfiLab and 61,226 SNP markers remained in the dataset. Principal component analysis was used to assess population substructure based on the SNP markers using the ade4 R package [23] and no evidence of population stratification was found.

Predictive ability

A 5-fold cross-validation (CV) scheme was used for assessing the predictive ability of the elastic-net (ENet) approach, which was chosen as the best-performing machine learning method in the blood metabolites prediction in previous studies of our group [9, 10]. We randomly split the dataset into five independent folds of approximately equal size. Thus, 4-fold were assigned to train the models and 1-fold to validate the model, and this CV procedure was repeated five times, predicting each fold in the validation set once. We used three elastic net (ENet) models to predict the target blood metabolite profile with increased complexity. The baseline model (M1) only considered NIR AfiLab information. In model 2 (M2), we combined NIR AfiLab and on-farm information (DIM and parity), while model 3 (M3) comprised NIR AfiLab, on-farm and genomic information.

Elastic-net (ENet)

The ENet represents a penalized regression that combines LASSO (least absolute shrinkage and selection operator; $l_1 = \sum_{w=1}^{p} |\beta_w|$) and RR (ridge regression; $l_2 = \sum_{w=1}^{p} \beta_w^2$) regularization terms [24]. The ENet alpha parameter (α) controls the balance between the regularization terms l_1 and l_2, providing a balance between selection (LASSO) and shrinkage (RR) of the predictor variables effects. ENet is considered a robust approach when predictor variables have strong collinearity. The optimum weight values for λ and α in the ENet regression model are considered to reduce the loss function as follows:

$$L(\lambda,\alpha,\beta) = min\left[\sum_{i=1}^{N}\{y_i - \sum_{w=1}^{p} x_{iw}\beta_w\}^2 + \lambda\left((1-\alpha)\sum_{w=1}^{p}\beta_w^2 + \alpha\sum_{w=1}^{p}|\beta_w|\right)\right],$$

where N is the number of animals, α is a value between 0 (RR penalty) and 1 (LASSO penalty), and λ is the regularization parameter that controls the amount of variable shrinkage. A random grid search was performed to find optimal values of α and λ ranging from 0.0 to 1.0 with an interval of 0.1 for each parameter. We implemented the ENet model using the glmnet R package [25]. The random search for α and λ was performed using the caret R package [26]. During the learning process of the ENet approach, the training set (4-fold) was split into an 80:20 ratio. The trained model with the highest accuracy and lowest mean square error (MSE) was then applied to a separate validation set (1-fold).

Model performance

The predictive ability of the different models was assessed by Pearson's correlation ($r = \text{cor}(y, \hat{y})$) between observed phenotypes and predicted phenotypes ($\hat{y}$). The predictive root mean squared error (RMSE) was $RMSE = \sqrt{\sum_{i=1}^{N}(y-\hat{y})^2/N}$, where N is the number of animals. The slope of the linear regression of $\hat{y}$ on y was also used to assess prediction bias. The relative difference (RD) in predictive ability was measured as $RD = \frac{(r_{mn}-r_{m1})}{r_{m1}} \times 100$, where r_{m1} is the predictive ability using the M1 approach and r_{mn} is the predictive ability using the other models.

Feature reduction prediction

The GWAS for blood metabolites were obtained with the following single-trait animal model via the genomic BLUP:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{W}\boldsymbol{h} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is a vector of blood metabolite information; $\boldsymbol{b}$ is the vector of fixed effects of days in milk with six classes (1: less than 60 d; 2: 60–120 d; 3: 121–180 d; 4: 181–240 d; 5: 241–300 d and 6: more than 300 d) and parity in 3 classes (1, 2, and ≥ 3 parities). The $\boldsymbol{h}$ and $\boldsymbol{a}$ are the random effects of batch and additive genetic effects, respectively; $\boldsymbol{X}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ are incidence matrices relating $\boldsymbol{y}$ to fixed effects ($\boldsymbol{b}$), batch effects ($\boldsymbol{h}$), and additive genomic breeding value ($\boldsymbol{a}$), respectively; and $\boldsymbol{e}$ is the residual effects.

The model was fitted under the following assumptions: $\boldsymbol{a} \sim \text{N}(0, \boldsymbol{G}\sigma_a^2)$, $\boldsymbol{h} \sim \text{N}(0, \boldsymbol{I}\sigma_{\text{batch}}^2)$ and $\boldsymbol{e} \sim \text{N}(0, \boldsymbol{I}\sigma_e^2)$, where σ_e^2, σ_{batch}^2, and σ_e^2 are variances for additive, batch, and residual effects, respectively; $\boldsymbol{I}$ is an identity matrix; and $\boldsymbol{G}$ is the genomic relationship matrix according to VanRaden [27]: $\boldsymbol{G} = \frac{\boldsymbol{MM'}}{2\sum_j^m=1p_j(1-p_j)}$ where $\boldsymbol{M}$ is the SNP matrix with codes 0, 1, and 2 for genotypes *AA*, *AB*, and *BB*, adjusted for allele frequency, and p_j is the frequency of the second allele of the j-th SNP.

The analyses were performed using the program blupf90+ [28]. The *P*-values were estimated by the SNP effects standardization as follows [29, 30]:

$$P\text{- value} = 2\left(1 - \varphi\left(\frac{|u_i|}{\sigma_{u_i}}\right)\right)$$

where u_i is the vector of the SNP marker effects, σ_{u_i} is the standard deviation of SNP marker effects (u_i) and φ is the cumulative function of the normal distribution for the SNP effects standardization $\left(\frac{|u_i|}{\sigma_{u_i}}\right)$.

In order to evaluate the effectiveness of reducing dimensionality on predictive ability, we selected SNP markers from GWAS results performed in each training fold from 5-fold CV (i.e., 4-fold for training and 1-fold for validation) based on three thresholds of marker significance ($-\log_{10}$(*P*-value)) deemed as higher than 2.0, 2.5, and 3.0. The average number of SNP markers selected in each threshold is shown in Additional file 1: Table S2.

Results

Descriptive statistics of the blood metabolic profile in the investigated population are reported in Additional file 1: Table S1. The cows enrolled in this study showed some relatively large data variability range for blood metabolites, which may indicate a low degree of physiological disturbance. Despite the absence of overt clinical disease, the high variability in certain blood biomarkers suggests the potential presence of subclinical conditions in specific individuals, which is expected in a large population. Specifically, we observed a degree of alteration in globulins (11% of cows > 50 g/L) and albumin (2% of cows < 30 g/L). Regarding urea levels, 43% of cows exceeded the threshold of ≥ 6.78 mmol/L. Less than 1% of the cows showed suspicion

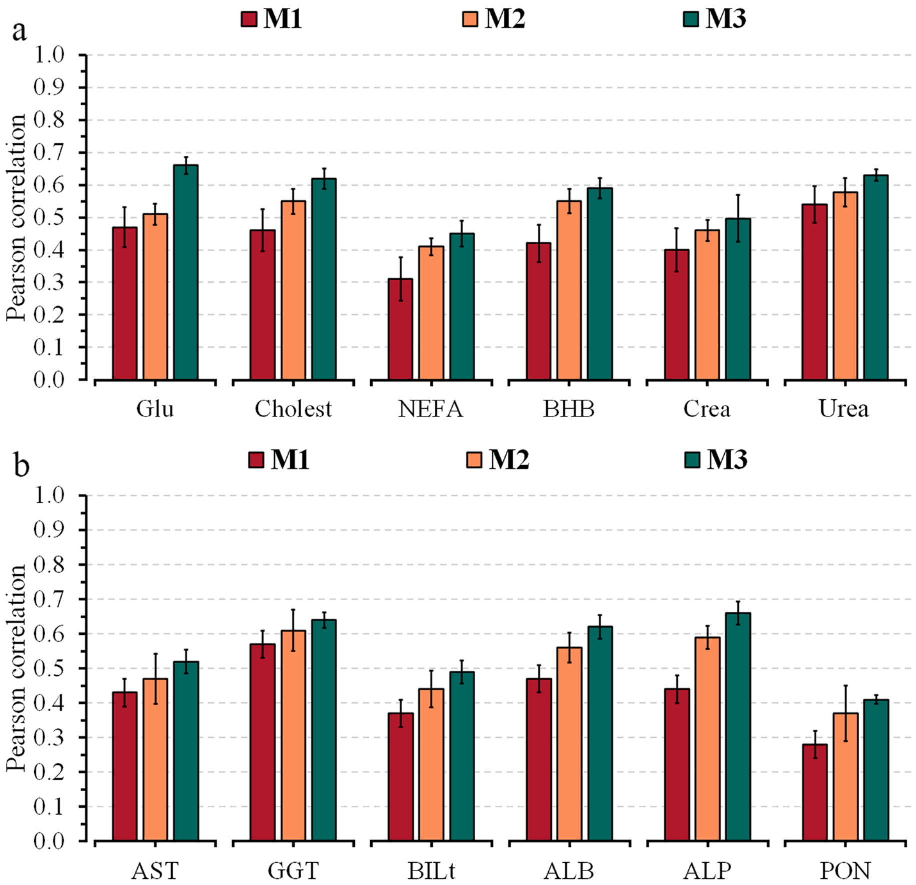

Fig. 1 Predictive ability across 5-fold random cross-validation for Model 1 (NIR AfiLab data), Model 2 (NIR AfiLab data and on-farm), and Model 3 (NIR AfiLab data, on-farm data, and genomic information) considering elastic net (ENet), for energy-related (**a**) and liver function and hepatic damage (**b**) blood metabolites. Data are shown as mean ± SD (black error bar line). Glu, glucose; Cholest, cholesterol; NEFA, non-esterified fatty acids; BHB, β-hydroxybutyrate; Crea, creatinine; AST, aspartate aminotransferase; GGT, γ-glutamyl transferase; BILt, total bilirubin; ALB, albumin; ALP, alkaline phosphatase; PON, paraoxonase

of hypomagnesemia or hypocalcemia, and less than 2% had hyperketonemia associated with high NEFA levels.

Predictive performance of NIR data integrated with on-farm and genomic information

Model M1, which included only the milk NIR information, achieved the lowest predictive ability (r) compared to the models including also on-farm data (M2) and both on-farm and genomic information (M3). For M1, the r-values ranged from 0.31 to 0.54 for energy-related metabolites (Fig. 1a), 0.28 to 0.57 for liver function/hepatic damage (Fig. 1b), 0.38 to 0.59 for inflammation/innate immunity (Fig. 2a), 0.34 to 0.52 for oxidative stress metabolites (Fig. 2b), and from 0.26 to 0.60 for minerals (Fig. 3) (see Additional file 1: Tables S3–S5, respectively). The combination of NIR and information on the farm (M2) achieved an average increase of 19% (3%–59%) in relation to the M1 model, with r-values ranging from 0.41 to 0.58 for energy-related metabolites, 0.37 to 0.61 for liver function/hepatic damage, 0.41 to 0.64 for inflammation/innate immunity, 0.45 to 0.54 for oxidative stress metabolites, and from 0.37 to 0.68 for minerals. Integrating on-farm and genomic information into NIR (M3) resulted in a 39% (12%–85%) average increase of r compared to M1, with r-values varying from 0.45 to 0.66 for energy-related metabolites, 0.41 to 0.66 for liver function/hepatic damage, 0.50 to 0.69 for inflammation/innate immunity, 0.52 to 0.63 for oxidative stress metabolites, and from 0.44 to 0.69 for minerals.

The results obtained from the M2 and M3 with a 5-fold CV show that including on-farm information (DIM and parity) or on-farm and genomic information enhances the predictive ability of NIR infrared prediction (see Additional file 2: Fig. S2). Moreover, the use of on-farm information in NIR infrared predictions improves the predictive ability by 3%–59%, with significant improvements seen for P, ROMt, Ca, ALP, NEFA, PON, and BHB

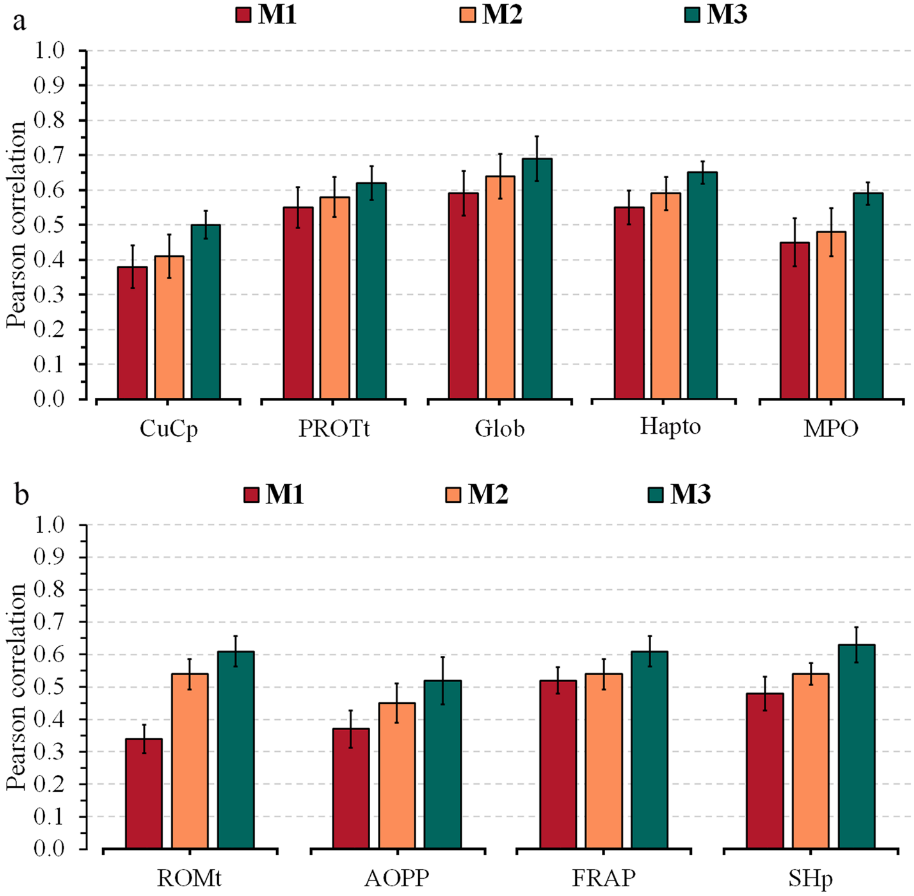

Fig. 2 Predictive ability across 5-fold random cross-validation for Model 1 (NIR AfiLab data), Model 2 (NIR AfiLab data and on-farm), and Model 3 (NIR AfiLab data, on-farm data, and genomic information) using the elastic net (ENet), for blood metabolites related to inflammation/innate immunity response (**a**) and oxidative stress (**b**). Data are shown as mean ± SD (black error bar line). CuCp, ceruloplasmin; PROTt, total proteins; Glob, globulins; Hapto, haptoglobin; MPO, myeloperoxidase; ROMt, total reactive oxygen metabolites; AOPP, advanced oxidation protein products; FRAP, ferric reducing antioxidant power; SHp, total thiol groups

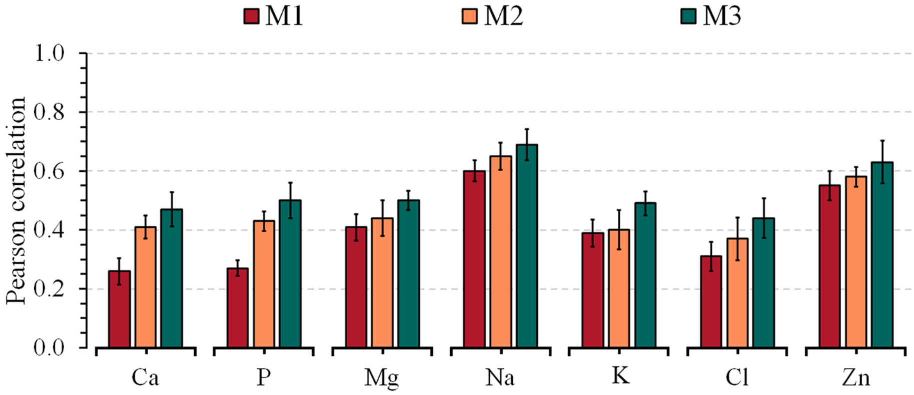

Fig. 3 Predictive ability across 5-fold random cross-validation for Model 1 (NIR AfiLab data), Model 2 (NIR AfiLab data and on-farm), and Model 3 (NIR AfiLab data, on-farm data, and genomic information) using the elastic net (ENet) for blood minerals. Data are shown as mean ± SD (black error bar line). Traits: Ca, calcium; P, phosphorus; Mg, magnesium; Na, sodium; K, potassium; Cl, chlorine and Zn, zinc

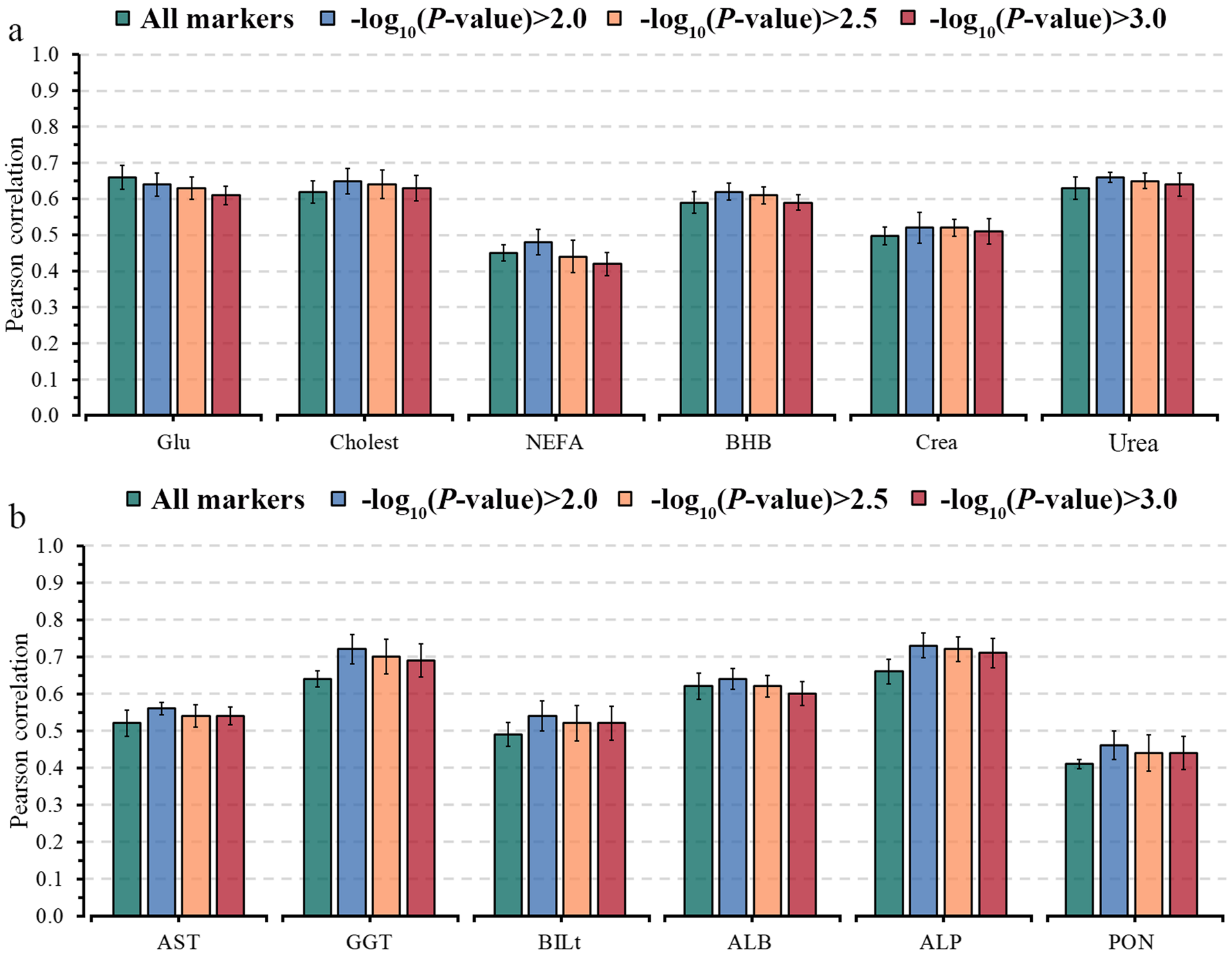

Fig. 4 Predictive ability, including standard errors, for energy-related (**a**) and liver function/hepatic damage (**b**) blood metabolites for ENet fitting all markers and using three thresholds based on marker significance: $-\log_{10}(P\text{-value}) > 2.0$, $-\log_{10}(P\text{-value}) > 2.5$ and $-\log_{10}(P\text{-value}) > 3.0$. Traits: Glu, glucose; Cholest, cholesterol; NEFA, non-esterified fatty acids; BHB, β-hydroxybutyrate; Crea, creatinine; AST, aspartate aminotransferase; GGT, γ-glutamyl transferase; BILt, total bilirubin; ALB, albumin; ALP, alkaline phosphatase; PON, paraoxonase

(see Additional file 2: Fig. S2). When both on-farm and genomic information are combined, the *r*-value increases from 12% to 85%, with an increase of more than 30% in 16 metabolites (P, Ca, ROMt, ALP, PON, NEFA, Cl, AOPP, BHB, glucose, cholesterol, BILt, albumin, CuCp, SHp, and myeloperoxidase) out of the 28 evaluated. The slope coefficients obtained from M2 and M3 indicate that the predictions were slightly underestimated or overestimated. The slope values for M2 ranged from 0.94 to 1.07, while for M3, the values were between 0.95 and 1.05. Model M1 showed more bias, with values varying from 0.85 to 1.29 (Additional file 2: Tables S3–S5).

Impact of feature selection on NIR AfiLab prediction performance

Using selected markers based on GWAS analyses improved the predictive ability when applying a threshold of $-\log_{10}(P\text{-value})$, except for Glu, FRAP, and P. The predictive ability (*r*) varied from 0.48 to 0.66 for energy-related metabolites, 0.46 to 0.73 for liver function/hepatic damage, 0.61 to 0.70 for inflammation/innate immunity, 0.60 to 0.72 for oxidative stress metabolites, and 0.48 to 0.70 (Figs. 4, 5, 6). On average, preselecting markers with a threshold of $-\log_{10}(P\text{-value}) > 2$ predictions achieved higher gains for oxidative stress metabolites (RD = 16%, ranging from −2% to 36%) and for liver function/hepatic damage traits (RD = 9%, ranging from 3% to 12%), while lower gain was observed for energy-related metabolites (RD = 4%, ranging from −3% to 7%).

The predictive ability of the model M3, considering selected markers with a threshold of $-\log_{10}(P\text{-value}) > 2.5$ showed slight improvements in the predictive

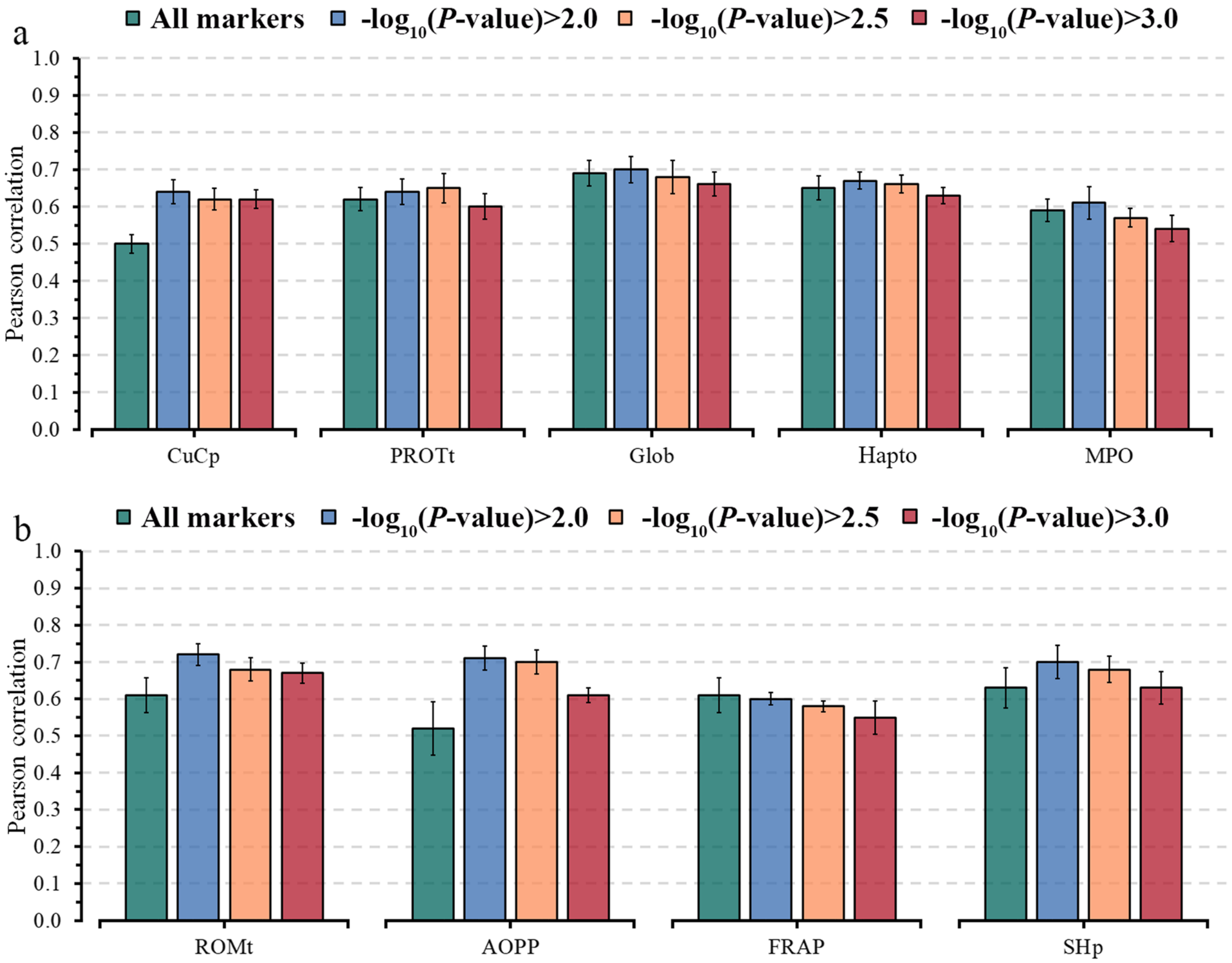

Fig. 5 Predictive ability, including standard errors, for inflammation/innate immunity (**a**) and oxidative stress (**b**) blood metabolites for ENet fitting all markers and using three thresholds based on marker significance: $-\log_{10}(P\text{-value}) > 2.0$, $-\log_{10}(P\text{-value}) > 2.5$ and $-\log_{10}(P\text{-value}) > 3.0$. Traits: CuCp, ceruloplasmin; PROTt, total proteins; Glob, globulins; Hapto, haptoglobin; MPO, myeloperoxidase; ROMt, total reactive oxygen metabolites; AOPP, advanced oxidation protein products; FRAP, ferric reducing antioxidant power; SHp, total thiol groups

ability compared to the threshold of 2. The threshold of 2.5 resulted in an average improvement of 8.8% in the *r*-value for 20 out of 28 evaluated metabolites (Figs. 4, 5, 6). However, using a more restrictive threshold ($-\log_{10}(P\text{-value})> 3$) to preselect markers led to a slight gain or reduction in predictive ability compared to using all markers in M3 (Figs. 4, 5, 6). AOPP and CuCp showed an RD higher than 10% for all evaluated thresholds, indicating that few genetic markers can explain their variability.

Discussion

Predictive performance integrating on-farm and genetic markers in NIR AfiLab

The study's objective was to evaluate the potential of integrating the AfiLab NIR milk analyzer with on-farm data (DIM and parity) and genetic marker information for the prediction of blood metabolites in Holstein cows. The Fourier-transform mid-infrared (FTIR) technique has become a broadly explored tool to predict complex traits, such as the blood metabolic profile in dairy cattle [8, 9, 31, 32]. Although in-line NIR infrared showed low to moderate predictive ability for blood metabolites (Figs. 1, 2, 3), it represents an alternative for daily monitoring at the herd level due to its daily availability. Previous studies have pointed out that using an integration of infrared with on-farm information (e.g., DIM, parity, and behavior parameters) or with on-farm and genetic markers allows improvement in infrared predictive ability for both FTIR [33–35] and NIR [15].

The adoption of the multi-data integration approach for predicting complex phenotypes is on the rise,

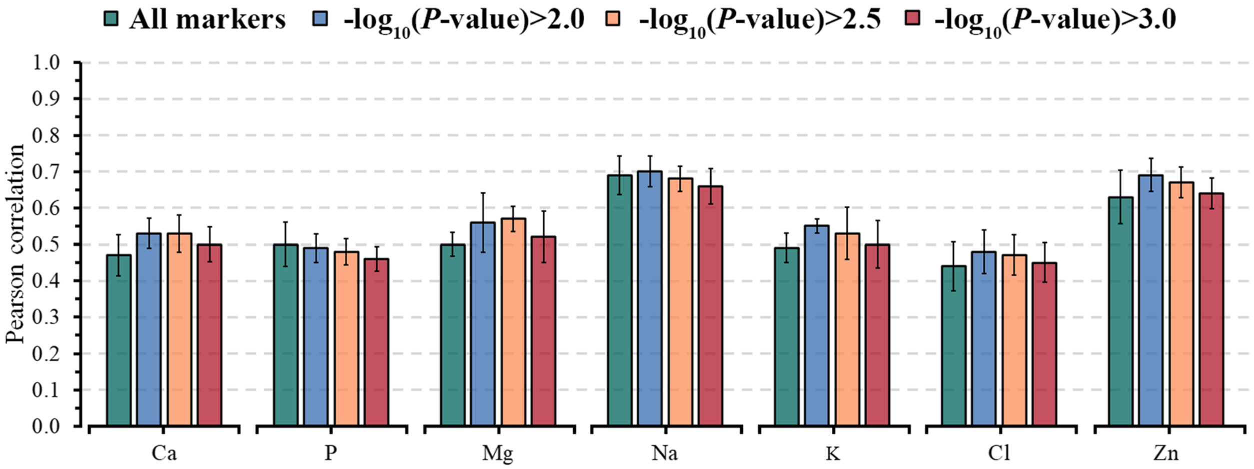

Fig. 6 Predictive ability, including standard errors, for blood minerals for ENet fitting all markers and using three thresholds based on marker significance: $-\log_{10}(P\text{-value}) > 2.0$, $-\log_{10}(P\text{-value}) > 2.5$ and $-\log_{10}(P\text{-value}) > 3.0$. Traits: Ca, calcium; P, phosphorus; Mg, magnesium; Na, sodium; K, potassium; Cl, chlorine and Zn, zinc

primarily because it offers a more precise representation of the intricate biological architecture associated with traits. Creating a training dataset structure from different sources on a massive is an integration and data architecture challenge. These data have different structure, dimensionality, resolution, and integrity. Such data can be used to build predictive models that accurately predict the unknown target traits in different farms and seasons. High-throughput technologies can gather on-farm data using automated in-line sensors installed in milking parlors. These sensors assess milk quality and quantity, recording information related to various aspects of individuals, including DIM, parity, and physiological parameters (e.g., respiratory rate and rumination time). The data obtained by these sensors can be combined with NIR information to predict novel phenotypes, which can then be used for selective breeding and management purposes. In recent years, there has been a growing emphasis on collecting on-farm data and integrating it to predict economically significant phenotypic traits in dairy cattle [8, 15, 33]. In this study, we observed that including on-farm information (DIM and parity) in the NIR predictions (M2) resulted in improvements on *r*-values with an average RD of 19% for energy-related metabolites, 20% for liver function/hepatic damage metabolites, 7% for inflammation/innate immunity metabolites, 24% for oxidative stress metabolites, and 23% for minerals (see Additional file 2: Fig. S1).

The increased *r*-value could be attributed to the direct influence of the lactation period on energy requirements and changes in milk yield, as well as milk fat and protein concentrations [36]. Since metabolic disorders also directly impact milk yield and quality [37], separating the effects of on-farm factors from those caused by metabolic distress can improve the predictive power of statistical algorithms [8]. In this context, Wu et al. [38] found that on-farm information (DIM and parity) has a great influence on variations in serum biochemical parameters and hormones related to protein status, energy supply, liver and kidney function, and oxidative stress of mid-lactation Holstein cows. Thus, by combining DIM and parity with the NIR infrared predictions, we reduced the prediction error in the independent population by accounting for varied physiological conditions along the lactation curve. Furthermore, integrating on-farm explanatory variables allows for capturing the lactation stage that explains the variability of the target phenotype, thus enhancing the accuracy of the NIR predictions.

When predicting blood metabolites using the M2, FRAP, and SHp demonstrated a better predictive ability on average inflammation/innate immunity metabolites, *r*-values ranged from 0.41 for CuCp to 0.64 for globulins and oxidative stress metabolites with *r*-values of 0.45 for AOPP and 0.54 for ROMt. These values were similar to those achieved using NIR spectra [10] but lower when compared to FTIR spectra [8, 9, 39]. A continuous exchange between blood and milk occurs through the blood-milk barrier, leading to a good predictive ability for inflammation and innate immunity-related metabolites because milk contains proteins like CuCp and haptoglobin originating from the bloodstream. In addition, during mammary gland inflammatory processes, the acute-phase proteins are also directly produced by milk leukocytes and mammary epithelial cells [40–42].

Myeloperoxidase, an enzyme released by activated neutrophils during inflammatory responses, is also found in milk and is associated with ongoing infections [40, 43]. On the other hand, oxidative stress has been associated with metabolic disorders in dairy cows, and milk contains measurable plasma thiols, reactive oxygen metabolites, and AOPP, which can indicate the individual's oxidative stress status [44]. In this context, our results demonstrate that using NIR infrared with on-farm data can achieve moderate accuracy in predicting blood oxidant-antioxidant status. This allows the identification of cows with high oxidative stress imbalance and detect energetic and metabolic impairments.

Mota et al. [8] observed improved prediction of blood traits by integrating milk FTIR spectra with on-farm and genetic markers, especially for metabolites under strong genetic control, i.e., higher heritability. Our results suggest that blending on-farm and genomic data in NIR AfiLab prediction can help predict variations in blood metabolites. In particular, we observed an average increase of 34% in the *r*-value for the investigated metabolites compared with M1 and 13% with M2 (see Additional file 2: Fig. S1). Increasing the power of prediction methods is an area of active research that aims to enable more efficient identification and allocation of cows less prone to be affected by metabolic disorders. Data integration could help predict the evolution of the metabolic response in the medium and long term and understand whether the animals are in a phase of adaptation or chronic stress [20].

Phenotype prediction using marker selection on NIR AfiLab prediction performance

The standard procedure for using genetic markers for infrared predictions is to use all information obtained from the SNP array as predictors. However, several complex biological downstream processes affect the phenotypic variability, and using SNP predictors more closely linked to the true quantitative trait loci (QTL) affecting the target phenotypes may increase the NIR AfiLab prediction performance. Selecting optimal markers based on their significance for the target trait is a crucial step in reducing the dimensionality of information in predictive models when multiple sources of information are combined. This helps to minimize the number of parameters in the model, avoiding overfitting and potentially improving the accuracy of predictions. The improvements in the predictive ability of a selected subset of markers depend on how well it matches the genetics underlying the phenotypic trait(s), and with a sufficient number of markers able to capture the trait variability (see Additional file 1: Table S2 and Additional file 2: Fig. S2–S4).

Previous studies adopted different strategies to preselect predictors by directly excluding uninformative markers via machine learning [45–47] or assigning weights to markers according to their contributions to trait variability [48]. Piles et al. [47] and Li et al. [49] showed that feature selection strategies improved the predictive ability of complex traits. We observed that preselected markers using a less conservative threshold ($-\log_{10}(P\text{-value}) > 2.0$) led to improvements in the *r*-value from 10% to 36%, even if a reduction was observed for glucose (3%) and FRAP (2%). The notable improvement in predictive capability seen with CuCp and AOPP can be attributed to the advantage gained from utilizing the most influential SNPs that bear biological relevance to the target trait (Additional file 2: Fig. S6 and S8). This is especially pronounced in traits influenced by QTL, which has a relatively significant effect (Additional file 2: Fig. S5–S9). Fragomeni et al. [50] and Mancin et al. [45] highlighted the advantages of removing non-informative SNP, where better accuracy was achieved by constructing the $\boldsymbol{G}$ matrix by considering the window region where the QTL was identified or by using only QTL information.

Selected SNPs have also been observed to capture significant within-family variation and Mendelian segregation effects [51]. Our findings emphasized that combining NIR infrared and on-farm data with selected markers significantly associated ($-\log_{10}(P\text{-value}) > 2.0$) with the target trait increased the predictive ability for predicting blood metabolites in dairy cattle (see Additional file 2: Fig. S2–S4). On the other hand, when dealing with more complex traits (i.e., polygenic traits), combining NIR infrared and on-farm information with approximately 5k selected markers (see Additional file 1: Table S2) resulted in a decrease in predictive ability compared to using all markers for glucose, NEFA, albumin, myeloperoxidase, FRAP, P and Na (see Additional file 2: Fig. S2–S4). These reductions were more remarkable as the selection criteria were more restrictive, i.e., $-\log_{10}(P\text{-value}) > 2.5$ and 3.0 (see Additional file 2: Fig. S2–S4), and this could be due to reduced linkage disequilibrium between the SNP and the true QTL [52].

Given this, comparing less restrictive threshold (> 2) to more restrictive (> 2.5 and > 3) showed predictive abilities that were lower by about 2% and 5% for energy-related metabolites, 3% and 4% for liver function/hepatic damage metabolites, 2% and 6% for inflammation/innate immunity metabolites, 3% and 10% for oxidative stress metabolites, and 2% and 7% for minerals. This result highlights the importance of preselecting markers for predicting complex phenotypes depending on how much this dimension reduction accurately selects predictor variables related to the target trait. Hence, our findings indicate that combining NIR infrared, on-farm and genomic information, or selected markers from GWAS, considering a threshold of $-\log_{10}(P\text{-value}) > 2.0$ can enhance the predictive ability of metabolic imbalances in dairy cattle.

As such, using multi-layer information to predict blood metabolites at the herd level in a rapid, affordable, and real-time manner unveils the promising potential of milk NIR spectra predictions in the early detection of metabolic disorders. Additionally, the outcomes of our study reveal moderate to high predictive abilities, making the prediction equations potentially useful in guiding herd management especially for its ability to capture day-by-day fluctuations, and formulating breeding recommendations for more resilient cows.

Conclusions

Integrating NIR spectra with on-farm and genomic information yielded a better predictive ability for blood metabolites than the model that relied solely on AfiLab milk NIR spectra in Holstein cattle. Indeed, the combination of NIR spectral data with on-farm and genomic information consistently outperformed prediction based on NIR spectra by an average of 34%. Preselecting genetic markers from GWAS has been shown to be an efficient strategy for dimensionality reduction by selecting trait-relevant markers, improving predictive ability because it extracts a smaller number of informative markers. We showed that preselecting genetic markers with a less restrictive threshold ($-\log_{10}(P\text{-value}) > 2.0$) resulted in better performance than considering all markers. Additionally, we found that using more restrictive thresholds ($-\log_{10}(P\text{-value}) > 2.5$ and 3.0) led to a negligible improvement in the predictive ability of blood metabolites.

Abbreviations

ALP	Alkaline phosphatase
AOPP	Advanced oxidation protein products
AST	Aspartate aminotransferase
BHB	β-Hydroxybutyrate
BILt	Total bilirubin
CuCp	Ceruloplasmin
CV	Cross–validation
DIM	Days in milk
ENet	Elastic net
FRAP	Ferric reducing antioxidant power
FTIR	Fourier transform mid infrared
GGT	γ-Glutamyl transferase
GWAS	Genome wide association study
LASSO	Least absolute shrinkage and selection operator
M1	Base model
M2	Model 2
M3	Model 3
NEFA	Non–esterified fatty acids
NIR	Near infrared
PON	Paraoxonase
QTL	Quantitative trait loci
r	Coefficient of correlation
RD	Relative difference
RMSE	Root mean squared error
ROMt	Total reactive oxygen metabolites
RR	Ridge regression
SHp	Total thiol group
SNP	Single nucleotide polymorphism

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40104-024-01042-3.

Additional file 1: Table S1. Descriptive statistics for blood metabolites in Holstein cows. **Table S2.** Average number of SNP markers selected for each training fold used during the cross-validation performance considering 5-fold. **Table S3.** Average prediction performance (± SD) of milk AfiLab NIR alone (model 1, M1), considering the systematic effect of days in milk and parity (model 2, M2) and considering the systematic effect of days in milk, parity, and genomic information (model 3, M3) for the 5-fold random cross-validation scenario using the elastic net method for energy-related and liver function/hepatic damage blood metabolites. **Table S4.** Average prediction performance (± SD) of milk AfiLab NIR alone (model 1, M1), considering the systematic effect of days in milk and parity (model 2, M2) and considering the systematic effect of days in milk, parity, and genomic information (model 3, M3) for the 5-fold random cross-validation scenario using the elastic net method for inflammation/innate immunity response and oxidative stress metabolites. **Table S5.** Average prediction performance (± SD) of milk AfiLab NIR alone (model 1, M1), considering the systematic effect of days in milk and parity (model 2, M2) and considering the systematic effect of days in milk, parity, and genomic information (model 3, M3) for the 5-fold random cross-validation scenario using the elastic net method for blood minerals.

Additional file 2: Fig. S1. Relative difference (%) in predictive ability for 5-fold random cross-validation scenarios using Elastic-net for Model 2 (M2; milk NIR data and on-farm data) and Model 3 (M3; milk NIR data, on-farm and genomic information) against Model 1, which considers only the NIR infrared data. Data are shown as mean ± SD (red error bar line). Glu – glucose; Cholest – cholesterol; NEFA – non-esterified fatty acids; BHB – β-hydroxybutyrate; Crea – creatinine; AST – aspartate aminotransferase; GGT – γ-glutamyl transferase; BILt – total bilirubin; ALB – albumin; ALP – alkaline phosphatase; PON – paraoxonase; CuCp – ceruloplasmin; Glob – globulins; PROTt – total proteins; Hapto – haptoglobin; MPO – myeloperoxidase; ROMt – total reactive oxygen metabolites; AOPP – advanced oxidation protein products; FRAP – ferric reducing antioxidant power; SHp – total thiol groups; Ca – calcium; P – phosphorus; Mg – magnesium; K – potassium; Na – sodium; Cl – chlorine; Zn – zinc. **Fig. S2.** Relative gain in predictive ability Pearson correlation, considering three thresholds based on marker significance (-log$_{10}$(*P*-value)) higher than 2.0, 2.5, and 3.0 against fitting all 61k SNPs, including standard errors, assessed for energy-related (**a**) and liver function and hepatic damage (**b**) blood metabolites. Data are shown as mean ± SD (black error bar line). Glu – glucose; Cholest – cholesterol; NEFA – non-esterified fatty acids; BHB – β-hydroxybutyrate; Crea – creatinine; AST – aspartate aminotransferase; GGT – γ-glutamyl transferase; BILt – total bilirubin; ALB – albumin; ALP – alkaline phosphatase; PON – paraoxonase. **Fig. S3.** Relative gain in predictive ability Pearson correlation, considering three thresholds based on marker significance (-log$_{10}$(*P*-value)) higher than 2.0, 2.5 and 3.0 against fitting all 61k SNPs, including standard errors, assessed for inflammation/innate immunity response (**a**) and oxidative stress blood metabolites (**b**). Data are shown as mean ± SD (black error bar line). CuCp – ceruloplasmin; PROTt – total proteins; Glob – globulins; Hapto – haptoglobin; MPO – myeloperoxidase; ROMt – total reactive oxygen metabolites; AOPP – advanced oxidation protein products; FRAP – ferric reducing antioxidant power; SHp – total thiol groups. **Fig. S4.** Relative gain in predictive ability Pearson correlation, considering three thresholds based on marker significance (higher than 2.0, 2.5 and 3.0 against fitting all 61k SNPs, including standard errors, assessed for blood minerals. Data are shown as mean ± SD (black error bar line). Ca – calcium; P – phosphorus; Mg – magnesium; Na – sodium; K – potassium; Cl – chlorine; Zn – zinc. **Fig. S5.** Manhattan plot for the average value of markers significance (-log$_{10}$(*P*-value)) across the 5-fold cross-validation for energy-related metabolites. Glu – glucose; Cholest – cholesterol; NEFA – non-esterified fatty acids; BHB – β-hydroxybutyrate; Crea – creatinine. **Fig. S6.** Manhattan plot for the average value of markers significance (-log$_{10}$(*P*-value)) across the 5-fold cross-validation for blood metabolites related to inflammation/innate immunity response. CuCp – ceruloplasmin; PROTt – total proteins; Glob – globulins; Hapto – haptoglobin; MPO – myeloperoxidase.

Fig. S7. Manhattan plot for the average value of markers significance (-$\log_{10}$(P-value)) across the 5-fold cross-validation for blood metabolites related to liver function and hepatic damage. AST – aspartate aminotransferase; GGT – γ-glutamyl transferase; BILt – total bilirubin; ALB – albumin; ALP – alkaline phosphatase; PON – paraoxonase. **Fig. S8.** Manhattan plot for the average value of markers significance (-$\log_{10}$(*P*-value)) across the 5-fold cross-validation for oxidative stress blood metabolites. ROMt – total reactive oxygen metabolites; AOPP – advanced oxidation protein products; FRAP – ferric reducing antioxidant power; SHp – total thiol groups. **Fig. S9.** Manhattan plot for the average value of markers significance (-$\log_{10}$(*P*-value)) across the 5-fold cross-validation for blood minerals. Ca – calcium; P – phosphorus; Mg – magnesium; K – potassium; Na – sodium; Cl – chlorine; Zn – zinc.

Acknowledgements
Not applicable.

Authors' contributions
AC conceived and led the coordination of the study. LFMM, AC, and GJMR performed the study design. LFMM, DG, HTA, AA, and GK work on data mining. LFMM and HTA conducted formal analysis. LFMM led the manuscript preparation and DG, and AC contributed to the manuscript preparation. GJMR, AC, LFMM, DG, SP, SS, LG, and ET participated in interpreting and discussing results. All authors read, revised, and approved the final manuscript.

Funding
Open access funding provided by Università degli Studi di Padova. The research was part of the project PROH-DAIRY project (Development of precision livestock breeding tools toward One Health in Italian and Israeli dairy chains) funded by the Ministry of Foreign Affairs and International Cooperation (MAECI) within the Italy-Israel R&D Cooperation Program (Roma, Italy). Moreover, the study was carried out within the Agritech National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4-D.D. 1032 17/06/2022, CN00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Availability of data and materials
The phenotypic and genotypic information are available for academic use from the authors upon reasonable request. The spectral data that support the findings of this study are deposited with Afimilk Ltd., and access is restricted as they were used under license for the current study and are therefore not publicly available. However, they can be obtained from the authors upon reasonable request and with the permission of Afimilk Ltd.

Declarations

Ethics approval and consent to participate
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 followed ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Agronomy, Food, Natural resources, Animals and Environment (DAFNAE), University of Padova, Legnaro, Padova 35020, Italy. [2]Department of Genetics and Biostatistics, School of Veterinary Medicine and Zootechnics, National Autonomous University of Mexico, Ciudad Universitaria, Mexico City 04510, Mexico. [3]Department of Animal Science, Food and Nutrition (DIANA) and the Romeo and Enrica Invernizzi Research Center for Sustainable Dairy Production (CREI), Faculty of Agricultural, Food, and Environmental Sciences, Università Cattolica del Sacro Cuore, Piacenza 29122, Italy. [4]Afimilk LTD, Afikim 15148, Israel. [5]Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI 53706, USA.

Received: 19 February 2024 Accepted: 28 April 2024
Published online: 09 June 2024

References

1. Lopreiato V, Mezzetti M, Cattaneo L, Ferronato G, Minuti A, Trevisi E. Role of nutraceuticals during the transition period of dairy cows: a review. J Anim Sci Biotechnol. 2020;11:96.
2. Zwald NR, Weigel KA, Chang YM, Welper RD, Clay JS. Genetic selection for health traits using producer-recorded data. II. Genetic correlations, disease probabilities, and relationships with existing traits. J Dairy Sci. 2004;87:4295–302.
3. Koeck A, Miglior F, Jamrozik J, Kelton DF, Schenkel FS. Genetic associations of ketosis and displaced abomasum with milk production traits in early first lactation of Canadian Holsteins. J Dairy Sci. 2013;96:4688–96.
4. Reksen O, Havrevoll Ø, Gröhn YT, Bolstad T, Waldmann A, Ropstad E. Relationships among body condition score, milk constituents, and postpartum luteal function in Norwegian dairy cows. J Dairy Sci. 2002;85:1406–15.
5. Giannuzzi D, Piccioli-Cappelli F, Pegolo S, Bisutti V, Schiavon S, Gallo L, et al. Observational study on the associations between milk yield, composition and coagulation properties with blood biomarkers of health in Holstein cows. J Dairy Sci. 2023;107:1397–412.
6. Mezzetti M, Cattaneo L, Passamonti MM, Lopreiato V, Minuti A, Trevisi E. The transition period updated: a review of the new insights into the adaptation of dairy cows to the new lactation. Dairy. 2021;2:617–36.
7. McArt JAA, Nydam DV, Oetzel GR, Overton TR, Ospina PA. Elevated non-esterified fatty acids and β-hydroxybutyrate and their association with transition dairy cow performance. Vet J. 2013;198:560–70.
8. Mota LFM, Giannuzzi D, Pegolo S, Trevisi E, Ajmone-Marsan P, Cecchinato A. Integrating on-farm and genomic information improves the predictive ability of milk infrared prediction of blood indicators of metabolic disorders in dairy cows. Genet Sel Evol. 2023;55:23.
9. Giannuzzi D, Mota LFM, Pegolo S, Tagliapietra F, Schiavon S, Gallo L, et al. Prediction of detailed blood metabolic profile using milk infrared spectra and machine learning methods in dairy cattle. J Dairy Sci. 2023;106:3321–44.
10. Giannuzzi D, Mota LFM, Pegolo S, Gallo L, Schiavon S, Tagliapietra F, et al. In-line near-infrared analysis of milk coupled with machine learning methods for the daily prediction of blood metabolic profile in dairy cattle. Sci Rep. 2022;12:8058.
11. de Roos APW, van den Bijgaart HJCM, Hørlyk J, de Jong G. Screening for subclinical ketosis in dairy cattle by fourier transform infrared spectrometry. J Dairy Sci. 2007;90:1761–6.
12. Gross JJ, Bruckmaier RM. Invited review: metabolic challenges and adaptation during different functional stages of the mammary gland in dairy cows: perspectives for sustainable milk production. J Dairy Sci. 2019;102:2828–43.
13. Giannuzzi D, Toscano A, Pegolo S, Gallo L, Tagliapietra F, Mele M, et al. Associations between milk fatty acid profile and body condition score, ultrasound hepatic measurements and blood metabolites in Holstein cows. Animals (Basel). 2022;12:1202.
14. Etzion Y, Linker R, Cogan U, Shmulevich I. Determination of protein concentration in raw milk by mid-infrared fourier transform infrared/attenuated total reflectance spectroscopy. J Dairy Sci. 2004;87:2779–88.
15. Mota LFM, Giannuzzi D, Bisutti V, Pegolo S, Trevisi E, Schiavon S, et al. 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. 2022;105:4237–55.
16. Melfsen A, Hartung E, Haeussermann A. Robustness of near-infrared calibration models for the prediction of milk constituents during the milking process. J Dairy Res. 2013;80:103–12.
17. National Research Council. Nutrient requirements of dairy cattle. 7th ed. Washington: The National Academies; 2021.

18. Calamari L, Ferrari A, Minuti A, Trevisi E. Assessment of the main plasma parameters included in a metabolic profile of dairy cow based on fourier transform mid-infrared spectroscopy: preliminary results. BMC Vet Res. 2016;12:4.
19. Hanasand M, Omdal R, Norheim KB, Gøransson LG, Brede C, Jonsson G. Improved detection of advanced oxidation protein products in plasma. Clin Chim Acta. 2012;413:901–6.
20. Premi M, Mezzetti M, Ferronato G, Barbato M, Piccioli Cappelli F, Minuti A, et al. Changes of plasma analytes reflecting metabolic adaptation to the different stages of the lactation cycle in healthy multiparous Holstein dairy cows raised in high-welfare conditions. Animals (Basel). 2021;11:1714.
21. Schmilovitch Z, Shmulevich I, Notea A, Maltz E. Near infrared spectrometry of milk in its heterogeneous state. Comput Electron Agric. 2000;29:195–207.
22. Shah NK, Gemperline PJ. A program for calculating Mahalanobis distances using principal component analysis. Trends Anal Chem. 1989;8:357–61.
23. Dray S, Dufour AB. The ade4 package: implementing the duality diagram for ecologists. J Stat Softw. 2007;22:1–20.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67:301–20.
25. Tay JK, Narasimhan B, Hastie T. Elastic net regularization paths for all generalized linear models. J Stat Softw. 2023;106:1.
26. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28:1–26.
27. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
28. Misztal I, Tsuruta S, Lourenco D, Aguilar I, Legarra A, Vitezica Z. Manual for BLUPF90 family of programs. Athens: University of Georgia; 2018.
29. Mota LFM, Santos SWB, Júnior GAF, Bresolin T, Mercadante MEZ, Silva JAV, et al. Meta-analysis across Nellore cattle populations identifies common metabolic mechanisms that regulate feed efficiency-related traits. BMC Genomics. 2022;23:424.
30. Mota LFM, Lopes FB, Fernandes Júnior GA, Rosa GJM, Magalhães AFB, Carvalheiro R, et al. Genome-wide scan highlights the role of candidate genes on phenotypic plasticity for age at first calving in Nellore heifers. Sci Rep. 2020;10:6481.
31. Grelet C, Bastin C, Gelé M, Davière JB, Johan M, Werner A, et al. 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. 2016;99:4816–25.
32. Aernouts B, Adriaens I, Diaz-Olivares J, Saeys W, Mäntysaari P, Kokkonen T, et al. Mid-infrared spectroscopic analysis of raw milk to predict the blood nonesterified fatty acid concentrations in dairy cows. J Dairy Sci. 2020;103:6422–38.
33. Baba T, Pegolo S, Mota LFM, Peñagaricano F, Bittante G, Cecchinato A, et al. Integrating genomic and infrared spectral data improves the prediction of milk protein composition in dairy cattle. Genet Sel Evol. 2021;53:29.
34. Dórea JRR, Rosa GJM, Weld KA, Armentano LE. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. J Dairy Sci. 2018;101:5878–89.
35. Lahart B, McParland S, Kennedy E, Boland TM, Condon T, Williams M, et al. Predicting the dry matter intake of grazing dairy cows using infrared reflectance spectroscopy analysis. J Dairy Sci. 2019;102:8907–18.
36. Van QCD, Knapp E, Hornick JL, Dufrasne I. Influence of days in milk and parity on milk and blood fatty acid concentrations, blood metabolites and hormones in early lactation Holstein cows. Animals (Basel). 2020;10:2081.
37. Bruckmaier RM, Gross JJ. Lactational challenges in transition dairy cows. Anim Prod Sci. 2017;57:1471–81.
38. Wu X, Sun HZZ, Xue M, Wang D, Guan L, Liu J. Days-in-milk and parity affected serum biochemical parameters and hormone profiles in mid-lactation Holstein cows. Animals (Basel). 2019;9:230.
39. Luke TDW, Rochfort S, Wales WJ, Bonfatti V, Marett L, Pryce JE. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra. J Dairy Sci. 2019;102:1747–60.
40. Cooray R. Use of bovine myeloperoxidase as an indicator of mastitis in dairy cattle. Vet Microbiol. 1994;42:317–26.
41. Upadhyaya I, Thanislass J, Veerapandyan A, Badami S, Antony P. Characterization of haptoglobin isotype in milk of mastitis-affected cows. Vet Sci. 2016;3:29.
42. Nielsen BH, Jacobsen S, Andersen PH, Niewold TA, Heegaard PMH. Acute phase protein concentrations in serum and milk from healthy cows, cows with clinical mastitis and cows with extramammary inflammatory conditions. Vet Rec. 2004;154:361–5.
43. Kimura K, Goff JP, Kehrli ME. Effects of the presence of the mammary gland on expression of neutrophil adhesion molecules and myeloperoxidase activity in periparturient dairy cows. J Dairy Sci. 1999;82:2385–92.
44. Gabai G, De Luca E, Miotto G, Zin G, Stefani A, Da Dalt L, et al. Relationship between protein oxidation biomarkers and uterine health in dairy cows during the postpartum period. Antioxidants. 2019;8:21.
45. Mancin E, Mota LFM, Tuliozi B, Verdiglione R, Mantovani R, Sartori C. Improvement of genomic predictions in small breeds by construction of genomic relationship matrix through variable selection. Front Genet. 2022;13:814264.
46. Lopes FB, Baldi F, Brunes LC, Oliveira e Costa MF, da Costa Eifert E, Rosa GJM, et al. Genomic prediction for meat and carcass traits in Nellore cattle using a markov blanket algorithm. J Anim Breed Genet. 2023;140:1–12.
47. Piles M, Bergsma R, Gianola D, Gilbert H, Tusell L. Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. Front Genet. 2021;12:611506.
48. da Silva Neto JB, Peripoli E, Pereira ASC, Stafuzza NB, Lôbo RB, Fukumasu H, et al. Weighted genomic prediction for growth and carcass-related traits in Nelore cattle. Anim Genet. 2023;54:271–83.
49. Li B, Zhang N, Wang Y-G, George AW, Reverter A, Li Y. Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. Front Genet. 2018;9:237.
50. Fragomeni BO, Lourenco DAL, Masuda Y, Legarra A, Misztal I. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet Sel Evol. 2017;49:65.
51. Chen ZQ, Klingberg A, Hallingbäck HR, Wu HX. Preselection of QTL markers enhances accuracy of genomic selection in Norway spruce. BMC Genomics. 2023;24:147.
52. Ling AS, Hay EH, Aggrey SE, Rekaya R. Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection. BMC Genomic Data. 2021;22:26.