Article

# Selection of the Most Informative Wavenumbers to Improve Prediction Accuracy of Milk Fatty Acid Profile Based on Milk Mid-Infrared Spectra Data

Wenqi Lou , Luiz Fernando Brito , Xiuxin Zhao , Jianbin Li * , Yachun Wang *

Posted Date: 20 October 2023

doi: 10.20944/preprints202310.1324.v1

Keywords: feature selection; milk mid-infrared spectra; fatty acids concentration; regression

*Article*

# Selection of the Most Informative Wavenumbers to Improve Prediction Accuracy of Milk Fatty Acid Profile Based on Milk Mid-Infrared Spectra Data

**Wenqi Lou [1], Luiz F. Brito [2], Xiuxin Zhao [3], Jianbin Li [3,*] and Yachun Wang [1,*]**

[1] Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture of China, State Key Laboratory of Animal Biotech Breeding, National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, 100193, China
[2] Department of Animal Sciences, Purdue University, West Lafayette, IN, 47907, USA
[3] Institute of Animal Science and Veterinary Medicine, Shandong Academy of Agricultural Sciences, Jinan, 250100, China
* J. Li: msdljb@163.com, +(86) 186 7865 9769; Y. Wang: wangyachun@cau.edu.cn, (+86) 158 0159 5851.

**Simple Summary:** Noisy and redundant wavenumbers in mid-infrared (MIR) spectra usually lead to low robustness and interpretability of the prediction models. Utilizing MIR spectra with feature selection (i.e., Uninformative Variable Elimination and Competitive Adaptive Reweighted Sampling) is more appropriate than simple screening, e.g., removing water-related regions. Feature selection improved prediction accuracy of milk FA concentration and identified the relevant wavenumbers of FAs. It can improve prediction models' accuracy in FAs based on small datasets and thereby providing more available FAs phenotype in dairy cows' breeding.

**Abstract:** Milk MIR spectra have been shown to provide valuable information on a wide range of traits to be used in dairy cattle breeding programs. Selecting the most informative variables from complex data can improve prediction accuracy and model robustness and, consequently, the interpretability of MIR spectra. Thus, we aimed to investigate the prediction performance of feature selection methods based on MIR spectra data, using the milk fatty acid (FA) profile as an example to illustrate the evaluated procedure. Data of MIR spectra, milk test-day records, and reference FA concentrations of 155 first-parity Holstein cows were used in the analyses. Four models comprising different explanatory variables and five feature selection methods were evaluated. The results indicated that the Competitive Adaptive Reweighted Sampling (CARS) method can effectively select the most informative variables from the MIR spectra, resulting in higher prediction accuracies than other variable selection approaches. The model including selected MIR spectra and cow information variables [days in milk at the test day, age at the test day, pregnancy stage (in days), number of days open, number of inseminations, and somatic cell count] yielded the best FA profile predictions based on Partial Least Square regression. In particular, ten FAs (C8:0, C10:0, C14:1, C17:0 isomers, C18:1, C18:1 isomer, medium-chain FA, unsaturation FA, monounsaturated FA, and polyunsaturated FA) presented accuracies based on the determination coefficient ($R^2cv$) ranging from 0.66 to 0.85 in internal validation and from 0.65 to 0.84 in external validation. By running CARS 1,000 times in internal validations, we obtained the frequency of selected milk MIR wavenumber for 35 FAs. The most related wavenumbers to FAs were found within 1,003 to 1,145 $cm^{-1}$, while other discrete areas were between 1,651 to 1,797 and 2,834 to 2,954 $cm^{-1}$. These biomarkers may give insights into the relationship between MIR spectra and FA phenotypes. In conclusion, using CARS and cow information improved predictions of FAs based on MIR spectra in Chinese Holstein dairy cows. Additional validation studies should be conducted as larger datasets become available.

**Keywords:** feature selection; milk mid-infrared spectra; fatty acids concentration; regression

## 1. Introduction

Mid-infrared (MIR) spectroscopy is a rapid, cost-effective, and classical tool widely used in official milk recording schemes and dairy cattle breeding programs for phenotyping milk composition traits (e.g., fat and protein) [1,2,3,4]. Many milk MIR-based prediction models have been reported in dairy cattle for deriving indicator traits of mineral content [5], methane emission [6], energy metabolites [7], fat globule size [8], fatty acids [9,10], ketosis [11], feed intake [12], lameness [13], nitrogen use efficiency [14,15], pregnancy status [16], cow diet [17], body condition score [18], and others. Hence, milk MIR spectra provide insight into milk composition and the physiological status of cows, which are helpful for farm management and breeding purposes. However, accurate estimation of detailed milk composition is challenging as it requires enough representative milk samples and collection of accurate information to get more robust predictions. This is even more important for detecting reference concentration of the target trait with low content in milk through some traditional and expensive methods (e.g., high-performance liquid chromatography). In the case of low research budgets and small-scale applications, many studies have used rather small datasets for research purposes. Alternatively, combining small datasets across laboratories and time is a feasible strategy for obtaining higher model accuracies, but standardization is still required to combine the datasets [19].

High-dimensionality of MIR spectra (e.g., 899 wavenumbers in Bentley instrument) tends to result in model overfit due to the impacts of collinearity, band overlaps, and redundant noise, especially for small datasets. The Partial Least Squares (PLS) approach can reduce these impacts, as one of powerful methods for quantitative analysis of milk MIR spectra [20]. Another practical approach to improve model robustness is to select optimal variables before modeling MIR spectra data, such as removing the regions associated with water or less informative MIR spectra regions (i.e., 1,600 to 1,700 and 3,000 to 3,500 cm$^{-1}$) to reduce the data noise [21]. Some feature selection algorithms have been shown to accurately screen for informative variables and generate significant improvements in the prediction accuracy for milk titratable acidity and calcium content [22] protein fractions [23], A1 and A2 milk [24], and cow's live weight [25] by combining the PLS method with the Uninformative Variable Elimination (UVE) or Competitive Adaptive Reweighted Sampling (CARS) methods based on MIR spectra [26,27]. Therefore, a feature selection algorithm is recommended to build simpler but more robust models to avoid overfitting or deleting valid variables, improve the prediction ability of MIR spectra, and identify the wavenumbers or their combinations that are more related to the target traits.

Milk fatty acid (FA) profile is crucial because it influences the characteristics of milk products [28] and is associated with human health [29-31]. However, the number of milk samples used to build prediction models for single and groups of FA in most published papers are small (less than 1,000) [9,32-35], and water-related wavenumbers were usually excluded from the analyses. Previous studies have genetically assessed these related water regions and concluded that useful physical-chemical information in prediction may exist [20,36-37]. In this context, the impact of directly removing water-related regions is unknown. Therefore, the primary objectives of this study were to 1) investigate the effects of UVE and CARS procedures on the accuracy of milk FA prediction based on PLS and MIR spectra data; 2) assess the prediction accuracy of milk FA profile from the models comprising different explanatory variables; and, 3) identify the most relevant wavenumbers from MIR spectra that contribute to better prediction accuracies of milk FA in Holstein cattle.

## 2. Materials and Methods

### *2.1. Sampling*

The milk samples used to develop prediction models were collected from 155 first-parity Holstein cows in a commercial dairy farm in Eastern China. Each cow was sampled once, and all samples were collected throughout lactation (6 to 305 days after calving) to obtain a good representation of the milk FA profile in the studied population. All the cows were housed in tie-stall barns and fed a total mixed ration four times daily formulated based on the National Research Council nutritional requirements [38], with ad libitum access to water.

Milking was done four times a day (6:00, 12:00, 18:00, and 24:00), and samples were collected in the morning milking sessions. The Afimilk's milking equipment was used to collect 160 mL milk (four repeated samples, 40 mL per sample) per cow, and then the samples were numbered sequentially and placed into clean and specific sampling bottles obtained from a Dairy Herd Improvement (DHI) test laboratory. To prevent the milk from spoilage, we added two tablets of preservative (bronopol, 2-bromo-2-nitropropan-1, 3-diol) in each sample container, which was gently shaken for uniform mixing. Three out of four samples per cow were transported and stored at 4 ℃ at the Dairy Cattle Center of Shandong Academy of Agricultural Science (Shandong, China) [23]. The fourth sample was sent to a commercial laboratory (www.iphenome.com.cn/index.jsp) for detection of milk FA based on ultra-high performance liquid chromatography – high resolution mass spectrometry (UPLC-HRMS).

### *2.2. Collection of MIR Spectra, Fatty Acids Profile, and Other Information*

Three repeated samples were promptly analyzed using one FTS machine on the same day (Bentley Instruments, Chaska, Minnesota, USA) to verify the reproducibility of MIR spectra produced by the FTS machine. One of the MIR spectra was used in later analyses. For each milk sample, the spectra contained 899 wavenumbers represented with absorbance value, indicating the absorption of infrared light through the milk sample at a particular wavelength in the range from 649.03 to 3,999.59 $cm^{-1}$. As shown in Supplementary Table 1, the cow information variables included days in milk (DIM, in days) at the herd test day, pregnancy length (PL, in days), days open (DO, in days), age (in days) at the herd test day, and number of inseminations (NI). Milk information for individual cows included fat percentage (FP, %), protein percentage (PP, %), lactose percentage (LP, %), somatic cell count (SCC, *1,000/mL), total milk solids (%), milk solids of non-fat (SNF, %), milk urea nitrogen (MUN, mg/dL), freezing point depression (FPD, m℃), casein (%), and β-hydroxybutyrate (BHB, mmol/L). All values of these milk components were predicted from milk MIR spectra via internal inbuilt modules in FTS machines.

With the benchmarked standard reference method of UPLC-HRMS [26,39], the concentrations (unit, μg/mL) of 25 individual milk FAs were measured from the fourth milk sample, including the caprylic acid (C8:0), capric acid (C10:0), undecanoic acid (C11:0), dodecanoic acid (C12:0), tridecanoic acid (C13:0), myristic acid (C14:0), myristoleic acid (C14:1), pentadecanoic acid (C15:0), palmitic acid (C16:0), heptadecanoic acid (C17:0), heptadecanoic acid isomer (C17:0 isomers), stearic acid (C18:0), oleic acid (C18:1), oleic acid isomer (C18:1 isomers), linoleic acid (C18:2), linoleic acid isomer (C18:2 isomers), linolenic acid ($\alpha$-C18:3), $\gamma$-linolenic acid ($\gamma$-C18:3), arachidic acid (C20:0), cis-8,11,14-eicosatrienoic acid (C20:3), arachidonic acid (C20:4), cis-5,8,11,14,17-eicosapentaenoic acid (C20:5), behenic acid (C22:0), tricosanoic acid (C23:0), and lignoceric acid (C24:0). Furthermore, other groups of milk FA were considered, including groups of middle-chain fatty acids (MCFA) containing C8:0 to C15:0; long-chain fatty acids (LCFA) containing C16:0 to C24:0; SFA grouping all the saturated FAs; UFA grouping all the unsaturated FAs; monounsaturated FA (MUFA) grouping C14:1, C18:1, and C18:1 isomers; polyunsaturated FA (PUFA) containing C18:2, C18:2 isomer, $\alpha$-C18:3, $\gamma$-C18:3, C20:3, C20:4, and C20:5; the ratio (U/S) of UFA to SFA; the unsaturated percentage (C14 index, %) for the sum of C14:1 and C14:0; the unsaturated percentage (C18 index, %) for the sum of C18:0, C18:1, C18:1 isomers, C18:2, C18:2, $\alpha$-C18:3, and $\gamma$-C18:3; and the unsaturated percentage (C20 index, %) for the sum of C20:0, C20:3, C20:4, and C20:5. Altogether, 35 FAs were evaluated (25 individual FAs, six milk FA groups, and four FA indicators).

### *2.3. Data Merging, Preprocessing, and Feature Variable Extraction*

All cows with 35 measured FAs were merged with MIR spectra, cow information (DIM, PL, DO, Age, and NI), and milk variables (FP, PP, LP, SCC, Solids, SNF, MUN, FPD, Casein, BHB, milk Density). The centeralization approach avoided magnitude effects from different explanatory variables in non-MIR data [40]. Based on the different preprocessing methods' coefficient of determination ($R^2$cv) from the model that only used MIR spectra and PLS to predict FAs (Supplementary Figure S1), MIR spectra's absorbance values were smoothed by the moving-average method (the size of the window was set as 11) with the 'prospectr' R package (https://github.com/l-ramirez-lopez/prospectr). This method defines an average of a fixed number of variables in the time series which move through the series by dropping and adding the next in each successive average [41]. This preprocessing method is a general method to decrease the perturbations from water and other molecules in milk and amplify the most informative signals.

Three feature selection methods were investigated and compared based on $R^2$cv of the model that only used MIR spectra and PLS to predict FAs, and the non-selection of MIR spectra variables was used as a control group. The first method was developed by deleting water-related regions from all MIR spectra variables (designated as non-water regions), as water-related regions in MIR spectra tend to be uninformative [21]. A total of 538 spectral wavenumbers (928 to 1,596 $cm^{-1}$, 1,693 to 3,025 $cm^{-1}$) out of the 899 were kept based on a previous study [3]. The second method was UVE, and the third was CARS, which were applied to all MIR spectra variables and selected variables based on PLS regression coefficients. The difference between UVE and CARS is that the ratio of the mean and standard deviation of the regression coefficient vectors are used to measure the significance of the variables in UVE, whereas CARS is used to retain variables with large absolute values of the regression coefficients and exclude those with small absolute values. To optimize the $R^2$cv, the threshold parameter of UVE was set to 0.9, while the length of Monte Carlo sampling was set to 50 and the resampling rate was set to 0.9 in CARS. Finally, the best feature selection method was used in the followed modeling with MIR spectra.

### *2.3. Construction of Models and Performance Evaluation*

Four models including different explanatory variables were tested for prediction ability of 35 FA traits (Table 1). The optimal principal components in PLS identified by $R^2$cv of each model. To simulate a scenario without MIR spectra, model 1 included cow information (DIM, Age, PL, DO, and NI) and model 2 incorporated all milk component profiles into the model 1. Then creating model 3 with the selected wavenumbers based on the best feature selection method. To evaluate the contribution of adding cow information to model performance, model 4 additionally used DIM, Age, PL, DO, NI, and SCC in model 3. As mentioned above, the concentrations of milk components in test-day records were tested by MIR spectroscopy, and only SCC was considered in model 4 to avoid double counting from milk information.

**Table 1.** Variables used in four models to predict milk fatty acid profile in Holstein dairy cows.

| Model | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| MIR spectra [1] | | | √ | √ |
| Cow information [2] | | | | |
| DIM | √ | √ | | √ |
| PL | √ | √ | | √ |
| DO | √ | √ | | √ |
| Age | √ | √ | | √ |
| NI | √ | √ | | √ |
| Milk information [3] | | | | |
| FP | | √ | | |
| PP | | √ | | |
| LP | | √ | | |
| SCC[1] | | √ | | √ |
| Solids | | √ | | |
| SNF | | √ | | |
| MUN | | √ | | |
| FPD | | √ | | |
| Casein | | √ | | |
| BHB | | √ | | |
| Density | | √ | | |

[1] MIR spectra: mid-infrared spectra; [2] DIM: days in milk; PL: pregnancy length; DO: days open; Age: age at test day; NI: number of inseminations; [3] FP: fat percentage; PP: protein percentage; LP: lactose percentage; SCC:

somatic cell count; SNF: milk solids of non-fat; MUN: milk urea nitrogen; FPD: freezing point depression; BHB: blood β-hydroxybutyrate.

Internal and external ten-fold cross-validations evaluated model performance. The data in internal validation was randomly divided into ten groups of equal size, and one group at a time was reserved as holdout data for testing, and all other samples were used to train the model. This process was repeated ten times until each group had been predicted, and the predicted value was saved to create the prediction accuracy of cross-validation. $R^2$cv was calculated based on all the ten groups' predictions and true values. Due to the random assignment of samples in the cross-validation procedure, the internal validation was repeated 1,000 times. The results were averaged to account for instability in the prediction model performance. To simulate the external validation of the prediction model on the relatively unfamiliar dataset, the internal validation process was repeated, but the difference in group division was divided data into ten specific groups (6 to 35 d, 36 to 65 d, …, 276 to 305 d) according to the DIM, i.e., the model developed with early and mid-lactation data were used to predict FA in late lactation. All calculations were carried out using scripts developed using MATLAB (2016a) and R software (v 4.10).

## 3. Results

### *3.1. Descriptive Statistics*

Paired t-tests of MIR spectra from three duplicated milk samples were performed and there were no significant differences. The mean, standard deviation (SD), and coefficient of variation (CV) of the UPLC-HRMS measurements of individual and grouped FA expressed on a milk basis (μg/mL) are summarized in Table 2. The individual FA with the highest concentrations were C16:0, C18:0, C18:1, and C18:1 isomers with mean values ranging from 963.96 to 7,950.18 μg/mL, while those with the lowest concentrations were C13:0, α-C18:3, and C20:5 with mean values ranging from 4.47 to 8.20 μg/mL. SFA was the most frequent FA, followed by MUFA and PUFA, which ratio was around 75%, 23%, and 2%. There was considerable variation in milk FA among cows, and the CV for milk samples in the entire dataset ranged from 4.89 to 73.81%.

**Table 2.** Description of 35 milk fatty acids (unit, μg/mL) in Chinese Holstein cows.

| Fatty acids | No. records | Minimum | Maximum | Mean | SD | CV |
|---|---|---|---|---|---|---|
| Individual fatty [1] acids | | | | | | |
| C8:0 | 155 | 11.67 | 76.35 | 33.28 | 11.88 | 35.69 |
| C10:0 | 155 | 9.82 | 199.35 | 55.66 | 32.49 | 58.36 |
| C11:0 | 155 | 13.50 | 37.78 | 23.26 | 4.60 | 19.79 |
| C12:0 | 155 | 97.41 | 571.31 | 236.41 | 86.64 | 36.65 |
| C13:0 | 155 | 2.46 | 22.27 | 8.20 | 3.60 | 43.84 |
| C14:0 | 155 | 95.00 | 1,538.24 | 481.71 | 266.91 | 55.41 |
| C14:1 | 154 | 1.37 | 350.07 | 84.18 | 62.14 | 73.81 |
| C15:0 | 151 | 1.49 | 187.70 | 54.87 | 34.66 | 63.17 |
| C16:0 | 155 | 4,805.08 | 11,397.96 | 7,950.18 | 1,448.56 | 18.22 |
| C17:0 | 155 | 19.56 | 79.44 | 38.61 | 10.79 | 27.96 |
| C17:0 isomers | 148 | 0.72 | 89.39 | 25.58 | 18.17 | 71.02 |
| C18:0 | 155 | 2,742.35 | 10,155.76 | 5,351.18 | 1,102.20 | 20.60 |
| C18:1 isomers | 155 | 285.01 | 2,261.21 | 963.96 | 411.81 | 42.72 |
| C18:1 | 154 | 5.10 | 11,681.66 | 3,445.55 | 1,986.80 | 57.66 |
| C18:2 isomers | 152 | 0.50 | 159.84 | 45.70 | 30.02 | 65.70 |
| C18:2 | 155 | 3.79 | 370.68 | 114.82 | 76.45 | 66.58 |
| α-C18:3 | 155 | 9.59 | 526.38 | 163.57 | 100.02 | 61.14 |
| γ-C18:3 | 153 | 0.19 | 12.26 | 4.47 | 19.20 | 61.86 |
| C20:0 | 155 | 23.58 | 238.78 | 45.28 | 19.20 | 42.41 |

| | | | | | | |
|---|---|---|---|---|---|---|
| C20:3 | 155 | 0.22 | 79.92 | 22.02 | 15.41 | 69.97 |
| C20:4 | 154 | 0.15 | 66.84 | 21.73 | 13.32 | 61.28 |
| C20:5 | 151 | 0.02 | 18.15 | 5.08 | 2.93 | 57.68 |
| C22:0 | 155 | 140.30 | 184.42 | 166.97 | 8.16 | 4.89 |
| C23:0 | 155 | 110.74 | 185.62 | 146.29 | 11.99 | 8.19 |
| C24:0 | 155 | 151.50 | 270.49 | 213.76 | 21.98 | 10.28 |
| Fatty acids group [2] | | | | | | |
| MCFA | 150 | 254.96 | 2,966.70 | 998.20 | 482.41 | 48.33 |
| LCFA | 145 | 12,150.86 | 32,291.46 | 19,000.58 | 4,092.60 | 21.54 |
| SFA | 146 | 9,793.13 | 23,204.17 | 14,891.33 | 2,553.94 | 17.15 |
| UFA | 148 | 974.11 | 15,506.05 | 5,048.91 | 2,575.58 | 51.01 |
| MUFA | 154 | 863.24 | 14,292.94 | 4,516.03 | 2,414.44 | 53.46 |
| PUFA | 148 | 21.98 | 1,213.11 | 392.08 | 229.52 | 58.54 |
| U/S | 145 | 0.09 | 0.78 | 0.34 | 0.14 | 42.55 |
| C14 index | 154 | 1.24 | 26.77 | 13.05 | 4.72 | 36.17 |
| C18 index | 150 | 14.73 | 73.38 | 45.15 | 14.56 | 32.24 |
| C20 index | 150 | 4.85 | 78.81 | 48.88 | 17.53 | 35.86 |

[1] C8:0: caprylic acid; C10:0: capric acid; C11:0: undecanoic acid; C12:0: dodecanoic acid; C13:0: tridecanoic acid; C14:0: myristic acid; C14:1: myristoleic acid; C15:0: pentadecanoic acid; C16:0: palmitic acid; C17:0: heptadecanoic acid; C17:0 isomers: heptadecanoic acid isomer; C18:0: stearic acid; C18:1: oleic acid; C18:1 isomers: oleic acid isomer; C18:2: linoleic acid; C18:2 isomers: linoleic acid isomer; $\alpha$-C18:3: linolenic acid; $\gamma$-C18:3: $\gamma$-linolenic acid; C20:0: arachidic acid; C20:3: cis-8,11,14-eicosatrienoic acid; C20:4: arachidonic acid; C20:5: cis-5,8,11,14,17-eicosapentaenoic acid; C22:0: behenic acid; C23:0: tricosanoic acid; C24:0: lignoceric acid; [2] MCFA: middle-chain fatty acids that contained C8:0 to C15:0; LCFA: long-chain fatty acids that contained C16:0 to C24:0; SFA: all the saturated FAs; UFA: all the unsaturated FAs; MUFA: the sum of C14:1, C18:1, and C18:1 isomers; PUFA: the sum of C18:2, C18:2 isomer, $\alpha$-C18:3, $\gamma$-C18:3, C20:3, C20:4, and C20:5; U/S: the ratio of UFA to SFA; C14 index: the unsaturated percentage for the sum of C14:1 and C14:0; C18 index: the unsaturated percentage for the sum of C18:0, C18:1, C18:1 isomers, C18:2, C18:2, $\alpha$-C18:3, and $\gamma$-C18:3; C20 index: the unsaturated percentage for the sum of C20:0, C20:3, C20:4, and C20:5; SD: standard deviation; C.V: coefficient of variation (unit, %).

### *3.2. Variable Optimization by UVE and CARS*

Figure 1 illustrates the process of UVE extracting variables. The curve on the left is the matrix of real variables (preprocessed MIR spectra), and the right is the matrix of artificially added random noise variables, whose dimension was equal to the spectral dimensions. The two horizontal dotted lines represent the noise thresholds, and most of the added noise variables' reliability values lie within the two thresholds. MIR spectral variables with reliability values between these two lines were recognized as noise variables and removed as non-informative variables, while out of the threshold lines' wavenumbers were identified as feature variables. In CARS, the number of feature variables decreased exponentially, then slowly decreased, and finally tended to stabilize as the number of iterations increased (Figure 2a). Figure 2b shows the variation in the root mean square error of cross-validation as a function of the number of iterations, and Figure 2c depicts the corresponding trend of R$^2$cv, the optimal combination of variables was obtained at the 21st iteration (the blue vertical line) and the minimum root mean square error of cross-validation value was 5.48.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 October 2023 doi:10.20944/preprints202310.1324.v1

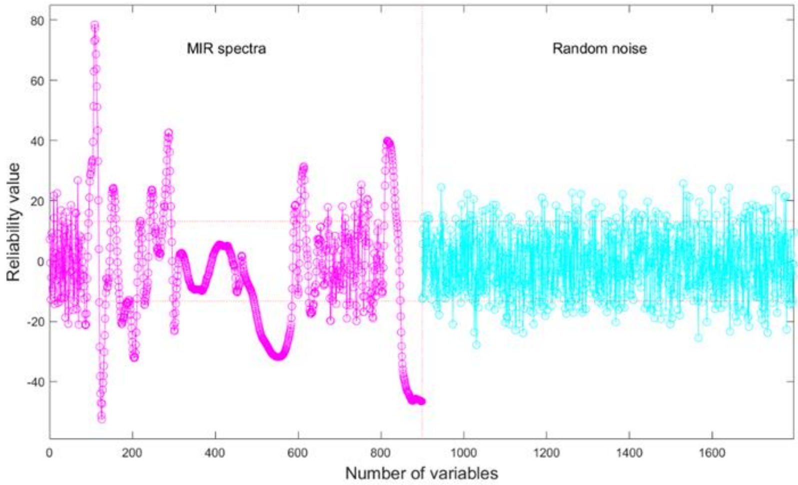

**Figure 1.** Features selected based on the Uninformative Variable Elimination method in milk mid-infrared spectra.

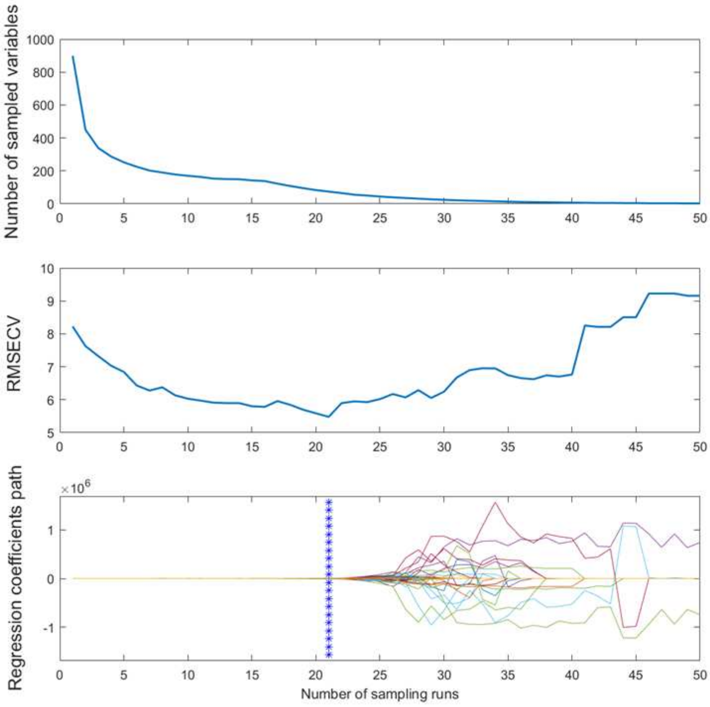

**Figure 2.** The changing trend of the number of sampled variables (plot a), 10-fold root mean square error of cross-validation (RMSECV) values (plot b), and regression coefficients of each variable (plot c) with the increasing of sampling runs based on Competitive Adaptive Reweighted Sampling method in milk mid-infrared spectra. The line (marked by asterisk) denotes the optimal point where root mean square error of cross-validation of 10-fold values achieve the lowest.

### *3.3. Comparisons of Feature Selections*

The mean values of 1,000 times internal validation results for the control group and three feature selection methods based on model 3 (only used MIR spectra) are shown in Table 3. The average prediction accuracies with the three methods evaluated (non-water regions, UVE, and CARS) were higher than the control group's results (no-selection of MIR spectra variables). The PLS regression had the best performance when the feature variables were extracted by CARS with the $R^2$cv for the 35 FA ranging from 0.01±0.01 to 0.76±0.03 and the mean was 0.46±0.20, which is higher than UVE (range = 0.02±0.01 to 0.67±0.03; mean = 0.42±0.17), non-water regions (range = 0.11±0.02 to 0.62±0.01; mean = 0.41±0.14), and non-selection (range = 0.03±0.01 to 0.51±0.02; mean = 0.32±0.14). The average $R^2$cv improved by 41% (CARS), 31% (UVE), and 28% (non-water regions) in comparison to the control group.

**Table 3.** The determination coefficient (mean ± standard deviation) of three feature selection methods and control group (none-selection in mid-infrared spectra) of 35 fatty acids using model 3 (only included mid-infrared spectra) and 1,000 times of internal validation.

| Fatty acids | None-selection [3] | Non-water regions [4] | UVE [5] | CARS [6] |
|---|---|---|---|---|
| Individual fatty acids [1] | | | | |
| C8:0 | 0.51± 0.02 | 0.62± 0.01 | 0.67± 0.03 | 0.76± 0.03 |
| C10:0 | 0.45± 0.02 | 0.53± 0.01 | 0.54± 0.01 | 0.60± 0.02 |
| C11:0 | 0.11± 0.02 | 0.27± 0.02 | 0.21± 0.03 | 0.19± 0.06 |
| C12:0 | 0.38± 0.02 | 0.53± 0.01 | 0.48± 0.02 | 0.56± 0.01 |
| C13:0 | 0.39± 0.02 | 0.52± 0.01 | 0.49± 0.04 | 0.55± 0.03 |
| C14:0 | 0.45± 0.02 | 0.52± 0.01 | 0.55± 0.03 | 0.61± 0.02 |
| C14:1 | 0.47± 0.02 | 0.58± 0.01 | 0.55± 0.02 | 0.62± 0.02 |
| C15:0 | 0.39± 0.02 | 0.49± 0.01 | 0.51± 0.02 | 0.57± 0.02 |
| C16:0 | 0.16± 0.02 | 0.22± 0.02 | 0.22± 0.02 | 0.25± 0.04 |
| C17:0 | 0.30± 0.02 | 0.33± 0.02 | 0.40± 0.02 | 0.44± 0.02 |
| C17:0 isomers | 0.43± 0.02 | 0.47± 0.01 | 0.54± 0.03 | 0.61± 0.03 |
| C18:0 | 0.06± 0.02 | 0.11± 0.02 | 0.06± 0.02 | 0.05± 0.03 |
| C18:1 isomers | 0.40± 0.03 | 0.52± 0.01 | 0.55± 0.04 | 0.60± 0.04 |
| C18:1 | 0.42± 0.02 | 0.44± 0.01 | 0.53± 0.03 | 0.62± 0.02 |
| C18:2 isomers | 0.41± 0.02 | 0.48± 0.01 | 0.52± 0.02 | 0.58± 0.03 |
| C18:2 | 0.39± 0.02 | 0.47± 0.01 | 0.50± 0.02 | 0.59± 0.2 |
| $\alpha$-C18:3 | 0.36± 0.02 | 0.45± 0.01 | 0.49± 0.03 | 0.56± 0.02 |
| $\gamma$-C18:3 | 0.31± 0.02 | 0.43± 0.02 | 0.39± 0.05 | 0.47± 0.07 |
| C20:0 | 0.03± 0.01 | 0.13± 0.04 | 0.02± 0.01 | 0.01± 0.01 |
| C20:3 | 0.42± 0.02 | 0.46± 0.01 | 0.52± 0.04 | 0.58± 0.04 |
| C20:4 | 0.41± 0.02 | 0.50± 0.01 | 0.51± 0.02 | 0.59± 0.01 |
| C20:5 | 0.23± 0.02 | 0.37± 0.02 | 0.33± 0.02 | 0.32± 0.08 |
| C22:0 | 0.13± 0.02 | 0.21± 0.02 | 0.15± 0.04 | 0.13± 0.07 |
| C23:0 | 0.09± 0.02 | 0.15± 0.01 | 0.11± 0.02 | 0.06± 0.04 |
| C24:0 | 0.09± 0.02 | 0.27± 0.01 | 0.16± 0.03 | 0.18± 0.08 |
| Fatty acids group [2] | | | | |
| MCFA | 0.46± 0.02 | 0.58± 0.01 | 0.56± 0.03 | 0.64± 0.02 |
| LCFA | 0.31± 0.02 | 0.35± 0.02 | 0.42± 0.03 | 0.45± 0.05 |
| SFA | 0.13± 0.02 | 0.18± 0.02 | 0.20± 0.03 | 0.20± 0.03 |
| UFA | 0.42± 0.02 | 0.48± 0.01 | 0.56± 0.03 | 0.62± 0.03 |
| MUFA | 0.38± 0.02 | 0.50± 0.01 | 0.53± 0.03 | 0.59± 0.04 |
| PUFA | 0.47± 0.02 | 0.51± 0.01 | 0.60± 0.03 | 0.66± 0.02 |
| U/S | 0.28± 0.03 | 0.38± 0.01 | 0.37± 0.04 | 0.35± 0.12 |
| C14 index | 0.34± 0.02 | 0.55± 0.01 | 0.49± 0.03 | 0.55± 0.03 |
| C18 index | 0.35± 0.03 | 0.47± 0.01 | 0.52± 0.03 | 0.57± 0.03 |
| C20 index | 0.34± 0.02 | 0.45± 0.01 | 0.47± 0.02 | 0.53± 0.02 |

[1] C8:0: caprylic acid; C10:0: capric acid; C11:0: undecanoic acid; C12:0: dodecanoic acid; C13:0: tridecanoic acid; C14:0: myristic acid; C14:1: myristoleic acid; C15:0: pentadecanoic acid; C16:0: palmitic acid; C17:0: heptadecanoic acid; C17:0 isomers: heptadecanoic acid isomer; C18:0: stearic acid; C18:1: oleic acid; C18:1 isomers: oleic acid isomer; C18:2: linoleic acid; C18:2 isomers: linoleic acid isomer; $\alpha$-C18:3: linolenic acid; $\gamma$-C18:3: $\gamma$-linolenic acid; C20:0: arachidic acid; C20:3: cis-8,11,14-eicosatrienoic acid; C20:4: arachidonic acid; C20:5: cis-5,8,11,14,17-eicosapentaenoic acid; C22:0: behenic acid; C23:0: tricosanoic acid; C24:0: lignoceric acid; MCFA:

middle-chain fatty acids that contained C8:0 to C15:0; LCFA: long-chain fatty acids that contained C16:0 to C24:0; SFA: all the saturated FAs; UFA: all the unsaturated FAs; [2] MUFA: the sum of C14:1, C18:1, and C18:1 isomers; PUFA: the sum of C18:2, C18:2 isomer, $\alpha$-C18:3, $\gamma$-C18:3, C20:3, C20:4, and C20:5; U/S: the ratio of UFA to SFA; C14 index: the unsaturated percentage for the sum of C14:1 and C14:0; C18 index: the unsaturated percentage for the sum of C18:0, C18:1, C18:1 isomers, C18:2, C18:2, $\alpha$-C18:3, and $\gamma$-C18:3; C20 index: the unsaturated percentage for the sum of C20:0, C20:3, C20:4, and C20:5; SD: standard deviation; C.V: coefficient of variation (unit, %); [3] None-selection: non-selection to MIR spectra variables; [4] Non-water regions: removing the water-related regions from MIR spectra variables; [5] UVE: Uninformative Variable Elimination method; [6] CARS: Competitive Adaptive Reweighted Sampling method.

### 3.4. Prediction Accuracy of the Models

The $R^2$cv of the 35 FAs for models 1 to 4 was obtained through internal validations (10-folds, random cross-validation) based on PLS regression (Figure 3). Only using cow demographic information (model 1, DIM, PL, DO, age, and IS) resulted in a range of the 35 FA from 0.04±0.02 to 0.83±0.01, with a mean of 0.40. The accuracy slightly improved by incorporating all milk information in model 2, in which the $R^2$cv ranged from 0.04±0.02 to 0.84±0.01 (mean $R^2$cv = 0.45). Based on the CARS filtered variables from MIR spectra, adding MIR spectra in models 3 and 4 further improved the model performance in comparison to model 2, and the average $R^2$cv across 35 FAs was the highest in model 4. For example, the improvement in $R^2$cv was 0.09 for C18:1 from model 2 ($R^2$cv = 0.53±0.01) to 3 ($R^2$cv = 0.62±0.02) based on the internal validation. In model 4, adding cow information and SCC increased the $R^2$cv of C18:1 to 0.68±0.02 in the internal validation. In addition, the prediction accuracy of internal validation with random grouping was consistently higher than that of external validation with DIM grouping. The $R^2$cv in the best model (model 4) of ten FAs (C8:0, C10:0, C14:1, C17:0 isomers, C18:1, C18:1 isomer, MCFA, UFA, MUFA, and PUFA) remained consistently high even in the external validation with $R^2$cv of 0.65±0.01 to 0.84±0.01 (Supplementary Figure 2).

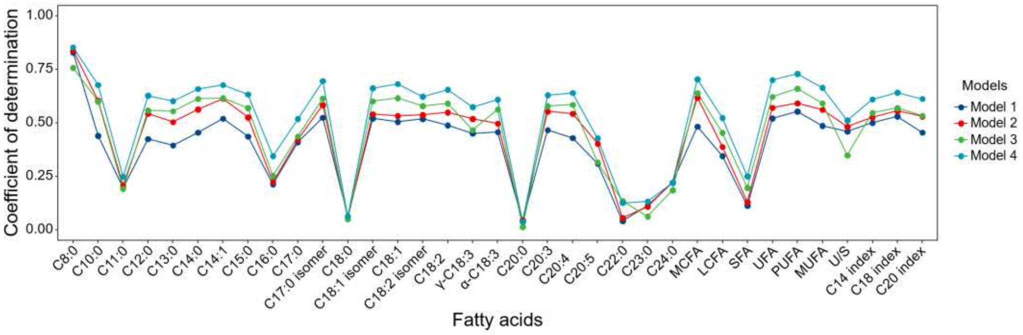

**Figure 3.** Coefficient of determination of 35 milk fatty acids based on the four models, competitive adaptive reweighted sampling, and partial least square regression in internal validation of Chinese Holstein dairy cows.

### 3.5. Identifying the Relevant Regions of FA in MIR Spectra

The results are summarized as a heat map (Figure 4) to illustrate the distribution of related regions in MIR spectra for 35 FA. Individual and grouped milk FA have similar trends in related wavenumbers from 640 to 4,000 $cm^{-1}$. As shown in Figure 4, the most related to FAs and common spectral region was found in 1,003 to 1,145 $cm^{-1}$ because its color was darker than the other regions. For the C18:0, C22:0, C23:0, and C24:0, the most related region was located in 2,834 to 2,954 $cm^{-1}$. Also, there were some small and separated related areas from 649 to 970, 1,651 to 1,797, 2,984 to 3,051, and 3,077 to 3,767 $cm^{-1}$.

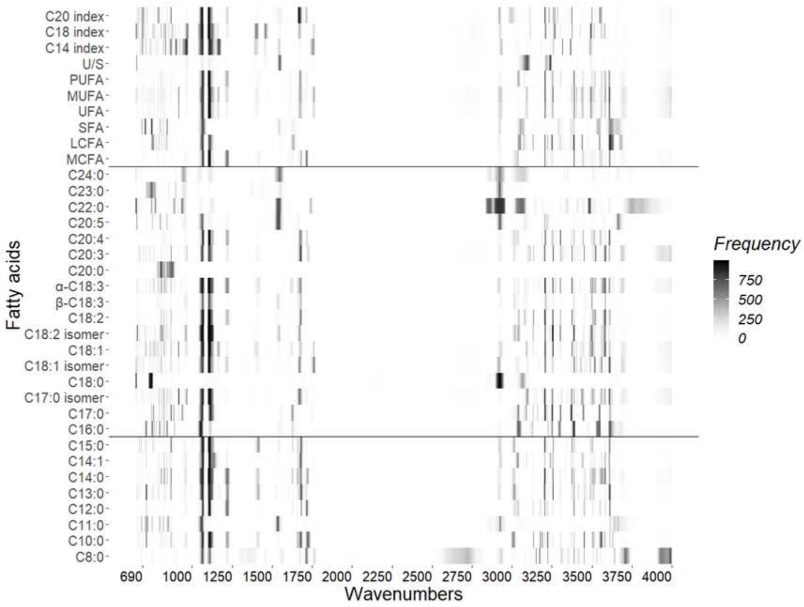

**Figure 4.** The selected frequency of each milk mid-infrared wavenumber (cm$^{-1}$) by running competitive adaptive reweighted sampling method with 1,000 times in internal validation of 35 milk fatty acids.

## 4. Discussion

We investigated the improvement of UVE and CARS procedures on the accuracy of milk FA prediction based on PLS and MIR spectra data in Holstein cattle, and found that CARS had the best performance in FA prediction. The model, included DIM, PL, DO, age at test day, NI, SCC, and selected MIR spectra, have the highest prediction accuracy than others. The most relevant wavenumber of the FA profile was from 1,003 to 1,145 cm$^{-1}$ in MIR spectra. This lays a foundation for the further research.

### *4.1. Descriptive Statistics*

There were no significant differences in MIR spectra of duplicated milk samples, which indicates high consistency among them and that the outputs of the FTS machine are reliable. Also, an accurate benchmark standard reference method is critical for the analyses. Compared with the method commonly used for FA detection (Gas Chromatography), UHPLC coupled with HRMS brings further advantages in terms of selectivity, sensitivity, and high throughput for the resolution analysis of complex samples and smaller particles [42], thereby allowing for the acquisition of more accurate reference FAs values. However, the high cost of testing resulted in a small dataset (n = 155) used in this study, which is lower than published studies such as Ferrand-Calmels et al. (2014) (n = 349) [33], Fleming et al. (2017b) (n = 2,023) [10], and Rovere et al. (2021) (n = 777) [35].

Milk FA profile has a dynamic pattern and the concentration of milk fat relies on factors such as measuring unit, breed, nutrition, individual FA, and period of lactation [43-45]. Not all 25 individual FA were detected in each milk sample, possibly due to low to zero concentrations in the samples. The relative magnitude among the 25 FA are consistent with previous findings, where C16:0, C18:0, and C18:1 were reported to be the main FA in Brown Swiss, Jersey, and Holstein cattle milk [32,46,47]. Regarding the saturation of milk fat, literature reports indicate SFA, MUFA, and PUFA concentration in dairy cows' milk, respectively, of 70%, 25%, and 5% [48]; 66%, 29%, and 5% [49], which close to the ratio in this study. There was considerable variation in milk FA among cows. C14:1, C17:0 isomer, C18:1, C18:2, C18:3, C20:3, and C20:4 resulted in the greatest values of CV (around and higher than 60%), which is similar to the results reported by Mele et al. (2009) [50]. Traits with larger phenotypic variation are preferred when genetically selecting for FAs that are beneficial to human health such as C18:1 (oleic acid), C18:2 (linoleic acid), C18:3 (α-linolenic acid and β-linolenic acid), and C20:4 (arachidonic acid).

### 4.2. Comparisons of Feature Selections

In this study, detecting noisy regions in MIR spectra can effectively improve the FA predictions. In particular, the improvements from CARS and UVE in MIR spectra were higher than that of non-water regions, indicating that the use of methods such as CARS and UVE could accurately select variables and obtain higher prediction accuracies instead of simple removing like removing water-regions from all MIR variables. In addition, CARS and UVE would be particularly useful when data at the population level are available because they can exponentially reduce computation time by using fewer but effective wavenumbers in FAs' modeling.

To our knowledge, previous studies have tested genetic algorithms for predicting FAs with MIR spectra in dairy cows. Ferrand et al. (2011) used genetic algorithm combined with PLS regression and reported improvements in accuracy of milk FA prediction by 15% on average based on milk MIR spectra (n = 153; 446 wavenumbers without water regions; Foss instrument) from Holstein X Normande dairy cows [51]. Caredda et al. (2016) obtained lower biases by testing genetic algorithm for FA prediction in sheep milk MIR spectra (n = 250; 550 wavenumbers without water regions; Foss instrument) [52]. Related studies based on CARS and UVE have focused on other traits in dairy cows. For instance, Gottardo et al. (2015) was the first to use UVE to enhance the prediction of milk titratable acidity (increase 10%) and Ca (increase 20%) content based on MIR spectra (n = 208; 520 wavenumbers without water regions; Foss instrument) from Holsten-Friesian cows [22]. Niero et al. (2016) reported that UVE improved the accuracy of prediction by 6.0 to 66.7% for milk protein fractions based on MIR spectra (n = 114; 520 wavenumbers without water regions; Foss instrument) in Holstein-Friesian, Brown Swiss, and Jersey cows [23]. CARS and UVE had similar performance in the classification of A1 and A2 milk based on MIR spectra (n = 838; 1,060 wavenumbers; Foss instrument) in Chinese Holstein cows [24]. Additionally, Zhang et al. (2021) compared three selection algorithms (sum of ranking difference algorithm, UVE, and Elastic Net regression) and found that they could improve the model's robustness in the prediction of dairy cow liveweight from MIR spectra and transferability to other brand of spectrometers [25]. Regardless of the selection algorithm and trait, removal of noisy variables often leads to better accuracy and performance of the models.

### 4.2. Prediction Accuracy of the Models

The increase $R^2$cv of the 35 FAs for models 1 to 4 indicates that incorporating more information into the model could improve the predictive accuracy of milk FAs. Milk MIR spectra could provide and capture more information about changes in milk composition and other biological differences in metabolic status in cows, because the syntheses of different milk FA profile are derived from adipose tissues mobilization and the volatile fatty acids produced by microbial fermentation [53].

The best prediction (model 4) of individual FA in this study was slightly lower than reports from Soyeurt et al. (2006b, 2011) ($R^2$cv = 0.01 to 0.98; unit = g/dL of milk) [9,49] and Rutten et al. (2009) ($R^2$cv = 0.14 to 0.96; unit = g/dL of milk) in accuracy [54], but higher than some results from De Marchi et al. (2011) ($R^2$cv = 0.51 to 0.77; unit = g/dL of milk) [32]. Notably, the prediction accuracies of UFA, such as C14:1, C18:1, C18:2, C18:3, C20:3, and C20:4, were similar or better than those of the same FA in the published results cited above. However, worse results were found for C11:0, C18:0, C20:0, C22:0, and C23:0 in this study, which also happened in other studies [32], and may be related to the pre-treatment as part of the reference method – UHPLC-HRMS. The above-mentioned saturated FAs also leads to lower accuracies of prediction models for SFA and U/S. The accuracies of PUFA, MUFA, and UFA in this study are within the ranges of the published reports cited above. This study attempted to predict MCFA, LCFA, U/S, C14, C18, and C20 index, which facilitate the direct determination of the chain length and degree of unsaturation of milk fats. The differences in the accuracy of FA predictions from published results are related to the variability of reference data, data size, reference method to determine FA composition, spectra pre-processing, and modeling method.

In addition, several authors who reported a reduction in prediction accuracy in external validation compared with internal validation based on herd by herd or herd-year by herd-year of data division [55, 56]. This implies that DIM-based division may be used as a supplemental way for external validation (see Supplementary Figure S2). Although the prediction accuracy of some FAs in this study did not reach the accuracy ($R^2cv$ ~ 0.89 to 1.00) of any application or quality control, it could be used for a rough screening of high or low FA values [20], or even for breeding as Tiplady et al. 2022 indicated that predicted FAs from MIR spectra with moderate accuracy ($R^2cv$ = 0.18 to 0.65) have solid genetic correlations (0.72 to 1.00) with directly measured FA [57].

### *4.3. Identifying the Relevant Regions of FA in MIR Spectra*

The wavenumber-reduced model provided more interpretable insights into the relationship between MIR and FAs. The CARS method is as stable as other methods in terms of feature selection [26]. Therefore, we statistically computed the selected frequency of each wavenumber (899 wavenumbers in MIR spectra) by running CARS 1,000 times in internal validation based on the best model (model 4). The most related to FAs and common spectral region was in 1,003 to 1,145 $cm^{-1}$. This is not surprising as the region located from 926 to 1,616 $cm^{-1}$ is called fingerprint region [58], containing much information on chemical bonds like –CH3, =CH2, -OH, C-O, and C-C in alcohols, ester, and carboxylic acids. Furthermore, this region had the highest heritability with estimates around 0.4 in Belgium and Canadian Holstein cows [59,60]. These estimates are similar to the MIR-predicted FAs in Brown Swiss, dual-purpose Belgian Blue, Holstein-Friesian, Jersey, Canadian Holstein cows, and other populations [61,62].

Other regions were located in 649 to 970, 1,651 to 1,797, 2,984 to 3,051, 2,834 to 2,954, and 3,077 to 3,767 $cm^{-1}$, which potentially reflect the homology and uniqueness in chemical structure among the 35 FAs evaluated in this study. The fact that selected regions were distributed in a wide range of the spectra is in agreement with the complex structure characteristics of FA, such as different vibration modes (stretching or bending), the complicated microenvironment in milk, and the interaction of C–H, C=O, and O–H bonds. For example, the regions of 1,651 to 1,797, 2,834 to 2,954 $cm^{-1}$ included fat A (~1,747 $cm^{-1}$; carbonyl stretch in fat) and fat B (~2,873 $cm^{-1}$; carbon-hydrogen stretch in fat) regions that were known to be associated with C=O (carbonyl) and C–H stretching, respectively [63]. However, the region of 1,800 to 2,900 $cm^{-1}$ did not contribute to most of the FA predictions, as these are typically considered to be meaningless regions in MIR spectra [64]. The frequency signals tended to be weaker and sparser within some discrete areas (i.e., 649 to 970, 2,984 to 3,051, and 3,077 to 3,767 $cm^{-1}$), as also reported by Rovere et al. (2021) [35]. Generally, the above regions provide a reference for building prediction models for FA.

## 5. Conclusions

This study combined feature selection methods to predict milk FA concentrations (unit, μg/mL) based on milk MIR spectra in Chinese Holstein cattle. The results revealed that CARS and UVE had the best performance compared to the traditional approaches, such as filtering the related water regions from MIR spectra, whereas selecting important wavenumbers from high dimensional data significantly improved the model fitting. The best model included DIM, PL, DO, age at test day, NI, SCC, and selected MIR spectra. Ten FAs (C8:0, C10:0, C14:1, C17:0 isomers, C18:1, C18:1 isomer, MCFA, UFA, MUFA, and PUFA) presented $R^2cv$ ranging from 0.65 to 0.84 in external validation. The regions of 1,003 to 1,145, 1,651 to 1,797, 2,834 to 2,954 $cm^{-1}$, and others were strongly associated with milk FAs. Further studies using larger datasets should be done in the future to validate these results.

**Supplementary Materials:** The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

**Author Contributions:** Resources, Xiuxin Zhao and Jianbin Li; data curation, Xiuxin Zhao and Jianbin Li; Formal analysis, Wenqi Lou; writing—original draft preparation, Wenqi Lou; writing—review and editing, Wenqi Lou and Luiz F. Brito; Supervision, Yachun Wang; Project administration, Jianbin Li and Yachun Wang. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the China Agriculture Research System, grant number CARS-36; the National Natural Science Foundation of China, grant number 32150610478; the Program for Changjiang Scholar and Innovation Research Team in University, grant number IRT_15R62; the National Agriculture Key Science & Technology Project, grant number NK20221201; the Key Research Project of Henan Province, grant number

221111111100; the National Key R&D Program of China, grant number 2021YFF1000701-06; the Shandong Provincial Natural Science Foundation, grant number ZR2021MC070; and China Scholarship Council, grant number 201913043.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Lynch, J.; Barbano, D.; Schweisthal, M.; Fleming, J. Precalibration Evaluation Procedures for Mid-Infrared Milk Analyzers. J. Dairy Sci. 2006, 89, 2761-2774.
2. De Marchi, M.; Toffanin, V.; Cassandro, M.; Penasa, M. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. J. Dairy Sci. 2014, 97, 1171-1186.
3. De Marchi, M.; Penasa, M.; Zidi, A.; Manuelian, C. Invited review: Use of infrared technologies for the assessment of dairy products—Applications and perspectives. J. Dairy Sci. 2018, 101, 10589-10604.
4. Tiplady, K.M.; Lopdell, T.J.; Littlejohn, M.D.; Garrick, D.J. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. J. Anim. Sci. Biotechnol. 2020, 11, 1-13.
5. Toffanin, V.; De Marchi, M.; Lopez-Villalobos, N.; Cassandro, M. Effectiveness of mid-infrared spectroscopy for prediction of the contents of calcium and phosphorus, and titratable acidity of milk and their relationship with milk quality and coagulation properties. Int. Dairy J. 2015, 41, 68-73.
6. Vanrobays, M.; Bastin, C.; Vandenplas, J.; Hammami, H.; Soyeurt, H.; Vanlierde, A.; Dehareng, F.; Froidmont, E.; Gengler, N. Changes throughout lactation in phenotypic and genetic correlations between methane emissions and milk fatty acid contents predicted from milk mid-infrared spectra. J. Dairy Sci. 2016, 99, 7247-7260.
7. Grelet, C.; Bastin, C.; Gelé, M.; Davière, J.; Johan, M.; Werner, A.; Reding, R.; Fernandez Pierna, J.; Colinet, F.; Dardenne, P.; Gengler, N.; Soyeurt, H.; Dehareng, F. 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, 4816-4825.
8. Fleming, A.; Schenkel, F.; Chen, J.; Malchiodi, F.; Ali, R.; Mallard, B.; Sargolzaei, M.; Corredig, M.; Miglior, F. Variation in fat globule size in bovine milk and its prediction using mid-infrared spectroscopy. J. Dairy Sci. 2017, 100, 1640-1649.
9. Soyeurt, H.; Dardenne, P.; Dehareng, F.; Lognay, G.; Veselko, D.; Marlier, M.; Bertozzi, C.; Mayeres, P.; Gengler, N. Estimating Fatty Acid Content in Cow Milk Using Mid-Infrared Spectrometry. J. Dairy Sci. 2006, 89, 3690-3695.
10. Fleming, A.; Schenkel, F.; Chen, J.; Malchiodi, F.; Bonfatti, V.; Ali, R.; Mallard, B.; Corredig, M.; Miglior, F. Prediction of milk fatty acid content with mid-infrared spectroscopy in Canadian dairy cattle using differently distributed model development sets. J. Dairy Sci. 2017, 100, 5073-5081.
11. Pralle, R.; Weigel, K.; White, H. Predicting blood β-hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network. J. Dairy Sci. 2018, 101, 4378-4387.
12. Wallén, S.; Prestløkken, E.; Meuwissen, T.; McParland, S.; Berry, D. Milk mid-infrared spectral data as a tool to predict feed intake in lactating Norwegian Red dairy cows. J. Dairy Sci. 2018, 101, 6232-6243.
13. Bonfatti, V.; Ho, P.N.; Pryce, J.E. Usefulness of milk mid-infrared spectroscopy for predicting lameness score in dairy cows. J. Dairy Sci. 2020, 103, 2534-2544
14. Grelet, C.; Froidmont, E.; Foldager, L.; Salavati, M.; Hostens, M.; Ferris, C.; Ingvartsen, K.; Crowe, M.; Sorensen, M.; Fernandez Pierna, J.; Vanlierde, A.; Gengler, N.; Dehareng, F. Potential of milk mid-infrared spectra to predict nitrogen use efficiency of individual dairy cows in early lactation. J. Dairy Sci. 2020, 103, 4435-4445.
15. Shi, R.; Lou, W.; Ducro, B.; van der Linden, A.; Mulder, H.A.; Oosting, S.J.; Li, S.; Wang, Y. Predicting nitrogen use efficiency, nitrogen loss and dry matter intake of individual dairy cows in late lactation by including mid-infrared spectra of milk samples. J. Anim. Sci. Biotechnol. 2023, 14, 8.
16. Brand, W.; Wells, A.; Smith, S.; Denholm, S.; Wall, E.; Coffey, M. Predicting pregnancy status from mid-infrared spectroscopy in dairy cow milk using deep learning. J. Dairy Sci. 2021, 104, 4980-4990.
17. Frizzarin, M.; Gormley, I.C.; Berry, D.P.; McParland, S. Estimation of body condition score change in dairy cows in a seasonal calving pasture-based system using routinely available milk mid-infrared spectra and machine learning techniques. J. Dairy Sci. 2023, 106, 4232-4244.
18. Frizzarin, M.; Visentin, G.; Ferragina, A.; Hayes, E.; Bevilacqua, A.; Dhariyal, B.; Domijan, K.; Khan, H.; Ifrim, G.; Nguyen, T. L.; Meagher, J.; Menchetti, L.; Singh, A.; Whoriskey, S.; Williamson, R.; Zappaterra, M.; Casa, A. Classification of cow diet based on milk Mid Infrared Spectra: A data analysis competition at

the "International Workshop on Spectroscopy and Chemometrics 2022". Chemometr Intell Lab Syst. 2023, 234, 104755.

19. Grelet, C.; Pierna, J.A.F.; Dardenne, P.; Soyeurt, H.; Vanlierde, A.; Colinet, F.; Bastin, C.; Gengler, N.; Baeten, V.; Dehareng, F. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. J. Dairy Sci. 2017, 100, 7910-7921.
20. Grelet, C.; Dardenne, P.; Soyeurt, H.; Fernandez, J.; Vanlierde, A.; Stevens, F.; Gengler, N.; Dehareng, F. Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions. Methods. 2021, 186, 97-111.
21. Li-Chan, E.; Chalmers, J.M.; Griffiths, P.R. Applications of vibrational spectroscopy in food science. John Wiley & Sons, Chichester, United Kindom, 2010.
22. Gottardo, P.; De Marchi, M.; Cassandro, M.; Penasa, M. Technical note: Improving the accuracy of mid-infrared prediction models by selecting the most informative wavelengths. J. Dairy Sci. 2015, 98, 4168-4173.
23. Niero, G.; Penasa, M.; Gottardo, P.; Cassandro, M.; De Marchi, M. Short communication: Selecting the most informative mid-infrared spectra wavenumbers to improve the accuracy of prediction models for detailed milk protein content. J. Dairy Sci. 2016, 99, 1853-1858.
24. Xiao, S.; Wang, Q.; Li, C.; Liu, W.; Zhang, J.; Fan, Y.; Su, J.; Wang, H.; Luo, X.; Zhang, S. Rapid identification of A1 and A2 milk based on the combination of mid-infrared spectroscopy and chemometrics. Food Control. 2022, 134, 108659.
25. Zhang, L.; Tedde, A.; Ho, P.; Grelet, C.; Dehareng, F.; Froidmont, E.; Gengler, N.; Brostaux, Y.; Hailemariam, D.; Pryce, J.; Soyeurt, H. Mining data from milk mid-infrared spectroscopy and animal characteristics to improve the prediction of dairy cow's liveweight using feature selection algorithms based on partial least squares and Elastic Net regressions. Computers and Electronics in Agriculture. 2021, 184, 106106.
26. Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Analytica Chimica Acta. 2009, 648, 77-84.
27. Ji, H.; Wang, W.; Chong, D.; Zhang, B. CARS Algorithm-Based Detection of Wheat Moisture Content before Harvest. Symmetry. 2019, 12, 115.
28. Huppertz, T.; Kelly, A.L. Properties and Constituents of Cow's Milk. 2008, 23-47.
29. Stewart, J.E.; Feinle-Bisset, C.; Keast, R.S. Fatty acid detection during food consumption and digestion: Associations with ingestive behavior and obesity. Prog. Lipid Res. 2011, 50, 225-233.
30. Ruiz-Núñez, B.; Dijck-Brouwer, D.J.; Muskiet, F.A. The relation of saturated fatty acids with low-grade inflammation and cardiovascular disease. J. of Nutr. Biochem. 2016, 36, 1-20.
31. Dehghan, M.; Mente, A.; Rangarajan, S.; Sheridan, P.; Mohan, V.; Iqbal, R.; Gupta, R.; Lear, S.; Wentzel-Viljoen, E.; Avezum, A.; Lopez-Jaramillo, P.; Mony, P.; Varma, R. P.; Kumar, R.; Chifamba, J.; Alhabib, K. F.; Mohammadifard, N.; Oguz, A.; Lanas, F.; . . . Yusuf, S. Association of dairy intake with cardiovascular disease and mortality in 21 countries from five continents (PURE): A prospective cohort study. The Lancet. 2018, 392, 2288-2297.
32. De Marchi, M.; Penasa, M.; Cecchinato, A.; Mele, M.; Secchiari, P.; Bittante, G. Effectiveness of mid-infrared spectroscopy to predict fatty acid composition of Brown Swiss bovine milk. Animal. 2010, 5, 1653-1658.
33. Ferrand-Calmels, M.; Palhière, I.; Brochard, M.; Leray, O.; Astruc, J.; Aurel, M.; Barbey, S.; Bouvier, F.; Brunschwig, P.; Caillat, H.; Douguet, M.; Faucon-Lahalle, F.; Gelé, M.; Thomas, G.; Trommenschlager, J.; Larroque, H. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 2013, 97, 17-35.
34. Gottardo, P.; Penasa, M.; Righi, F.; Lopez-Villalobos, N.; Cassandro, M.; De Marchi, M. Fatty acid composition of milk from Holstein-Friesian, Brown Swiss, Simmental and Alpine Grey cows predicted by mid-infrared spectroscopy. Ital. J. Anim. Sci. 2017, 16, 380-389.
35. Rovere, G.; de Los Campos, G.; Lock, A.L.; Worden, L.; Vazquez, A.I.; Lee, K.; Tempelman, R.J. Prediction of fatty acid composition using milk spectral data and its associations with various mid-infrared spectral regions in Michigan Holsteins. J. Dairy Sci. 2021, 10, 11242-11258.
36. Wang, Q.; Hulzebosch, A.; Bovenhuis, H. Genetic and environmental variation in bovine milk infrared spectra. J. Dairy Sci. 2016, 99, 6793-6803.
37. Toledo-Alvarado, H.; Vazquez, A.I.; De los Campos, G.; Tempelman, R.J.; Bittante, G.; Cecchinato, A. Diagnosing pregnancy status using infrared spectra and milk composition in dairy cows. J. Dairy Sci. 2018, 101, 2496-2505.
38. NRC. 2001. Nutrient requirements of dairy cattle (Seventh revised edition). Washington, D.C, National Academy Press.
39. Mattarozzi, M.; Riboni, N.; Maffini, M.; Scarpella, S.; Bianchi, F.; Careri, M. Reversed-phase and weak anion-exchange mixed-mode stationary phase for fast separation of medium-, long- and very long chain free fatty acids by ultra-high-performance liquid chromatography-high resolution mass spectrometry. J. chromatogr A. 2021, 1648, 462209.

40. Hayes, A.F. Truths and Myths about Mean Centering. Introduction to mediation, moderation, and conditional process analysis. The Guiford Press, New York, USA. 2013, 282-288.
41. Kumar, M.; Shanker Rao, G. Chapter 12 - Analysis of Time Series. Statistical Techniques for Transportation Engineering. Elsevier. 2017, 463-489.
42. Alseekh, S.; Scossa, F.; Fernie, A.R. Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research. Nat. Methods. 2021, 18, 733-746.
43. Collomb, M.; Bütikofer, U.; Sieber, R.; Jeangros, B.; Bosset, J.O. Correlation between fatty acids in cows' milk fat produced in the Lowlands, Mountains and Highlands of Switzerland and botanical composition of the fodder. Int. Dariy. J. 2002, 12, 661-666.
44. Soyeurt, H.; Dardenne, P.; Gillon, A.; Croquet, C.; Vanderick, S.; Mayerses, P.; Bertozzi, C.; Gengler, N. Variation in Fatty Acid Contents of Milk and Milk Fat Within and Across Breeds. J. Dairy Sci. 2006, 89, 4858-4865.
45. Bainbridge, M.L.; Cersosimo, L.M.; Wright, A.D.G.; Kraft, J. Content and Composition of Branched-Chain Fatty Acids in Bovine Milk Are Affected by Lactation Stage and Breed of Dairy Cow. PloS One. 2016, 11, e0150386.
46. White, S.L.; Bertrand, J.A.; Wade, M.R. Comparison of Fatty Acid Content of Milk from Jersey and Holstein Cows Consuming Pasture or a Total Mixed Ration[J]. J. Dairy Sci. 2001, 84 ,2295-2301.
47. Fengen, W.; Meiqing, C.; Runbo, L.; Guoxin, H.; Xufang, W.; Nan, Z.; Yangdong, Z.; Jiaqi, W. Fatty acid profiles of milk from Holstein cows, Jersey cows, buffalos, yaks, humans, goats, camels, and donkeys based on gas chromatography-mass spectrometry. J. Dairy Sci. 2022, 105, 1687-1700.
48. Grummer, R.Ric. Effect of feed on the composition of milk fat. J. Dairy Sci. 1991, 74, 3244-3257.
49. Soyeurt, H.; Dehareng, F.; Gengler, N.; Mcparland, S.; Wall, E.; Berry, D.P.; Coffey, M.; Dardenne, P. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries. J. Dairy Sci. 2011, 94, 1657-1667.
50. Mele, M.; Dal Zotto, R.; Cassandro, M.; Conte, G.; Serra, A.; Buccioni, a.; Bittante, G.; Secchiari, P. Genetic parameters for conjugated linoleic acid, selected milk fatty acids, and milk fatty acid unsaturation of Italian Holstein-Friesian cows. J. Dairy Sci. 2009, 92, 392-400.
51. Ferrand, M.; Huquet, B.; Barbey, S.; Barillet, F.; Faucon, F.; Larroque, H.; Leray, O.; Trommenschlager, J.M.; Brochard, M. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemometr Intell Lab Syst. 2011, 106, 183-189.
52. Caredda, M.; Addis, m.; Ibba, i.; Leardi, R.; Scintu, M.F.; Piredda, G.; Sanna, G. Prediction of fatty acid content in sheep milk by Mid-Infrared spectrometry with a selection of wavelengths by Genetic Algorithms. LWT-Food Sci Technol. 2016, 65, 503-510.
53. Tian, Z.; Zhang, Y.; Zhang, H.; Sun, Y.; Mao, Y.; Yang, Z.; Li, M. Transcriptional regulation of milk fat synthesis in dairy cattle. J. Funct. Foods. 2022, 96, 105208.
54. Rutten, M.J.M.; Bovenhuis, B.; Hettinga, K.A.; van Valenberg, H.J.F.; van Arendonk, J.A.M. Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. J. Dairy Sci. 2009, 12, 6202-6209.
55. Ho, P.; Bonfatti, V.; Luke, T.; Pryce, J. Classifying the fertility of dairy cows using milk mid-infrared spectroscopy. J. Dairy Sci. 2019, 102, 10460-10470.
56. Luke, T.; Rochfort, S.; Wales, W.; Bonfatti, V.; Marett, L.; Pryce, J. Metabolic profiling of early-lactation dairy cows using milk mid-infrared spectra. J. Dairy Sci. 2019, 102, 1747-1760.
57. Tiplady, K.M.; Lopdell, T.J.; Sherlock, R.G.; Johnson, T.J.J.; Spelman, R.; Harris, B.L.; Davis, S.R.; Littlejohn, M.D.; Garrick, D.J. Comparison of the genetic characteristics of directly measured and Fourier-transform mid-infrared-predicted bovine milk fatty acids and proteins. J. Dairy Sci. 2022, 105, 9763-9791.
58. Duffy and Norman, V. Interpretation of infrared spectra. Journal of Chemical Education. 1972, 49, 30-45.
59. Soyeurt, H.; Misztal, I.; Gengler, N. Genetic variability of milk components based on mid-infrared spectral data. J. Dairy Sci. 2010, 93, 1722-1728.
60. Rovere, G.; de Los Campos, G.; Tempelman, R.J.; Vazquez, A.I.; Miglior, F.; Schenkel, F.; Cecchinato, A.; Bittante, G.; Toledo-Alvarado, H.; Fleming, A. A landscape of the heritability of Fourier-transform infrared spectral wavelengths of milk samples by parity and lactation stage in Holstein cows. J. Dairy Sci. 2019, 102, 1354-1363.
61. Soyeurt, H.; Gillon, A.; Vanderick, S.; Mayeres, P.; Bertozzi, C.; Gengler, N. Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. J. Dairy Sci. 2007, 90, 4435-4442.
62. Narayana, S.G.; Schenkel, F.S.; Fleming, A.; Koeck, A.; Malchiodi, F.; Jamrozik, J.; Johnston, J.; Sargolzaei, M.; Miglior, F. Genetic analysis of groups of mid-infrared predicted fatty acids in milk. J. Dairy Sci. 2017, 100, 4731-4744.

63. Kaylegian, K.E.; Lynch, J.M.; Fleming, J.R.; Barbano, D.M. Influence of fatty acid chain length and unsaturation on mid-infrared milk analysis. J. Dairy Sci. 2009, 92, 2485-2501.

64. Sun, D.W. Infrared spectroscopy for food quality analysis and control: Fourier transform infrared (FT-IR) spectroscopy. Academic/Press/Elsevier, London, UK, 2009.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.