Near-infrared spatially-resolved spectroscopy for milk quality analysis

Authors

Jose A. Diaz-Olivares[1a], Martin J. Gote[a], Wouter Saeys[b], Ines Adriaens[a], Ben Aernouts[a]

[a] KU Leuven, Department of Biosystems, Division of Animal and Human Health Engineering, Campus Geel, Kleinhoefstraat 4, 2440 Geel, Belgium

[b] KU Leuven, Department of Biosystems, Division Mechatronics, Biostatistics and Sensors, Kasteelpark Arenberg 30, 3001 Leuven, Belgium

[1] Corresponding author: Jose A. Diaz-Olivares, KU Leuven, Department of Biosystems, Division of Animal and Human Health Engineering, Campus Geel, Kleinhoefstraat 4, 2440 Geel, Belgium, +32 (0)14 72 14 67, jose.diaz@kuleuven.be

---

Abstract

To support in-line quality control of raw milk, we propose a fiber-optic spatially-resolved spectroscopy (SRS) setup. This setup allows to vary the interaction of long-wave near-infrared (LW-NIR) light with the milk by submerging a separate optical illumination and detection fiber into the sample and altering their relative distance to optimize measurements for specific milk component(s). We evaluated this approach for predicting milk fat, protein, and lactose content and determined the optimal illumination-to-detection distances for each milk component. The region between 1.1 to 1.8 mm was optimal for lactose, and between 2.2 and 3.8 mm for fat and protein. These distance ranges resulted in a root-mean-square error of prediction (RMSEP) of less than 0.10% weight/weight (wt/wt) for milk fat, and lower than 0.13% (wt/wt) for protein and lactose. Integration of these distances into a fiber-optic SRS reflectance probe would allow to

simultaneously determine the fat, protein, and lactose content of raw milk in-line with high accuracy.

Keywords

SRS; spatially resolved spectroscopy; near-infrared spectroscopy; food quality control; reflectance; milk;

1 Introduction

The demand for high-quality dairy products is rising at a global scale (OECD, 2022). Hence, there is an increasing need for higher and more efficient production capacity (Thornton, 2010) and for the capability to monitor and differentiate the quality of these high-value foodstuffs (Walstra et al., 2005). To ensure high-quality dairy products and optimize the processing chain, milk composition is a crucial indicator that should be regularly evaluated at every stage, from on-farm production to manufacturing and end-product storage. Regular milk composition monitoring can ensure the quality of dairy products, which is essential for public health, safety and consumer assurance (Gunasekaran, 1996; Kong et al., 2019; Paiva, 2013). In addition, changes over time in milk fat, protein and lactose of individual cows can be linked to their nutritional and metabolic status, as well as to their udder health (Mäntysaari et al., 2019). Therefore, monitoring the produced milk is an efficient way to also evaluate the performance and welfare of dairy cows in addition to ensuring and optimizing dairy quality (McParland et al., 2014).

The implementation of in-line measurements is necessary when considering the increasing automation of the dairy sector and the food industry (Rodenburg, 2017). Non-destructive sensors integrated into a milk pipeline can enable high-frequency and autonomous composition monitoring, which can have measurable positive impacts on dairy quality control and processing efficiency, animal health and welfare, and dairy production economics (Kunes et al., 2021). In-line milk quality analysis can further benefit dairy processing systems by enabling milk sorting infrastructures to classify milk into different batches at the point of production (Augustin et al.,

2013), based on required quality properties. This technology can also facilitate the creation of new dairy products and the optimization of the production process for existing ones.

Non-destructive milk quality control technologies have been applied in the dairy sector to characterize the composition and morphology of dairy products (Ellis et al., 2012; Gowen et al., 2007; Karoui and de Baerdemaeker, 2007; Marcone et al., 2013). Amongst these technologies, near-infrared (NIR) spectroscopy has been proven to be an advantageous analytical method in the food production and processing industry due to the minimal sample preparation and labor requirements, and the absence of chemical consumables needed for operation (Shenk et al., 2007). Various researchers (Kawamura et al., 2007; Kawasaki et al., 2008; Melfsen et al., 2012; Saranwong and Kawano, 2008; Tsenkova et al., 1999) have already examined the application of NIR spectroscopy for on-site monitoring of milk quality. In previous works (Aernouts et al., 2011; Diaz-Olivares et al., 2020), the authors highlighted the advantage of the long-wave NIR (LW-NIR) range (1000 to 1700 nm) in reflectance and transmittance spectroscopic analysis of raw milk. Specifically, on-site LW-NIR analyses showed an outstanding performance for raw milk composition prediction, with a coefficient of determination (R^2) greater than 0.95 for fat and protein and 0.69 for lactose, with root-mean-squared errors (RMSE) below 0.08% weight/weight (wt/wt) for all three components (Diaz-Olivares et al., 2020).

However, conventional transmittance and diffuse reflectance LW-NIR spectroscopy have not been widely implemented in dairy production or processing lines due to the turbid nature of raw milk. In this turbid media, varying physical structures such as fat globules and casein micelles strongly scatter light, while water molecules strongly absorb LW-NIR radiation, making the quantification of milk components challenging (Bogomolov and Melenteva, 2013). Furthermore,

the sample thickness in LW-NIR transmittance measurements is limited to 2 mm (Aernouts et al., 2011; Tsenkova et al., 1999). In contrast, in-line measurement of diffuse reflectance is more straightforward, but the acquired signals are dominated by LW-NIR light that has been superficially scattered by the numerous fat globules and casein micelles in the milk. As a result, it provides limited information on the absorption by chemical compounds, particularly of the non-scattering components dissolved in the milk serum (Aernouts et al., 2015a, 2011).

In spatially-resolved spectroscopy (SRS), light that has interacted with the sample is collected at different distances from the illumination point (Torricelli et al., 2013). By increasing the physical distance between illumination and detection, the captured light will have interacted more and with deeper layers of the sample. However, this longer illumination-to-detection distance also reduces the intensity of the measured signal significantly, which may jeopardize the signal-to-noise ratio (SNR). To find a balance between signal intensity and information content, it is necessary to optimize the optical configuration for a specific application (van Beers et al., 2015). SRS can also separate the information on light scattering and absorption related to the composition and physical structure of samples by combining spectroscopic measurements collected at various illumination-to-detection distances (Wetzel and LeVine, 2000). Several research groups have investigated the potential of this technique for food quality control (Huang et al., 2018; Ma et al., 2022; Nguyen Do Trong et al., 2014a, 2014b; Vanoli et al., 2020).

When applied to solid samples, SRS typically utilizes a contactless optical configuration using lasers for illumination and cameras for detection (Aernouts et al., 2015b). SRS on liquids is usually performed with optical fibers that make direct contact with or are submerged into the sample (Crofcheck et al., 2002; Foschum et al., 2011; Watté et al., 2016). Fiber-optic SRS can be

implemented in a reflectance probe design with different optical fibers serving as illumination and detection points. Multiple detection fibers can be distributed linearly on a flat surface and located parallel and at different distances from a single, unique illumination fiber (Bendoula et al., 2019). An SRS reflectance probe configuration would be straightforward to install in a milk pipeline and could support effective in-line milk analysis. As the path length of light traveling through the milk sample increases with distance, its interaction with the milk serum also intensifies. This could be beneficial for predicting lactose content, which is heavily diluted and dissolved in the milk serum (Aernouts et al., 2015a, 2011).

Previous studies have focused on applying SRS to milk quality analysis in the visible (400 – 780 nm) and short-wave NIR (780 – 1010 nm) range (Bogomolov et al., 2017; Crofcheck et al., 2002). In this wavelength range, milk components have lower absorption coefficients, resulting in the absorption of only a small fraction of the incident light. This leaves a substantial amount of the interacted light available for detection and simplifies the process of making accurate SRS measurements. However, the low absorption coefficients of the milk components within this range of wavelengths lead to SRS spectra dominated by light scattering effects and contain only minimal information about the chemical compounds of the sample. As a result, SRS for milk quality analysis within this range is limited to the quantification and characterization of milk components that scatter light. For instance, Crofcheck et al. (2002) studied milk fat globule size distribution effects on visible and short-wave NIR SRS spectra, while Bogomolov et al. (2017) combined those spectra to predict the milk fat and protein content with good accuracy (RMSE ≤ 0.09% wt/wt).

Nonetheless, to successfully detect milk components that mainly absorb light at higher wavelengths, such as lactose, the LW-NIR range must be considered (Aernouts et al., 2015a, 2011). Watté et al. (2016) explored LW-NIR SRS to quantify the bulk optical properties of milk based on a simulation study. They concluded that the most informative illumination-to-detection distances for bulk optical properties determination were 1.1 to 2.5 mm, and recommended the use of an additional detection fiber at a further distance to differentiate between scattering effects and absorption.

To the best of our knowledge, no studies have been reported on the application and evaluation of SRS for milk quality analysis in the LW-NIR range. It is hypothesized that the illumination-to-detection distance resulting in the best predictions may vary for the three milk components. A deeper analysis of the observations is expected to provide valuable insights into how optical food sensors can be improved concerning the characteristics of the sample and the quality attributes of interest. Therefore, this study aimed to (1) develop a system that can collect high-quality LW-NIR SRS spectra of turbid liquids at different illumination-to-detection distances; (2) measure LW-NIR SRS spectra of raw milk samples with a wide variety in chemical composition; (3) develop and evaluate models to predict milk fat, protein and lactose from the spectra acquired at single illumination-to-detection distances; (4) study the effect of the illumination-to-detection distances on their prediction performance; and (5) provide recommendations for the design of a fiber-optic LW-NIR SRS reflectance probe to increase its technology readiness level (TRL) from laboratory validation to eventually an autonomous in-line milk quality analysis application (Mankins, 1995).

2 Materials and methods

2.1 Spatially resolved spectroscopy setup

The experimental setup used for the SRS measurements on raw milk is schematically illustrated in Figure 1.

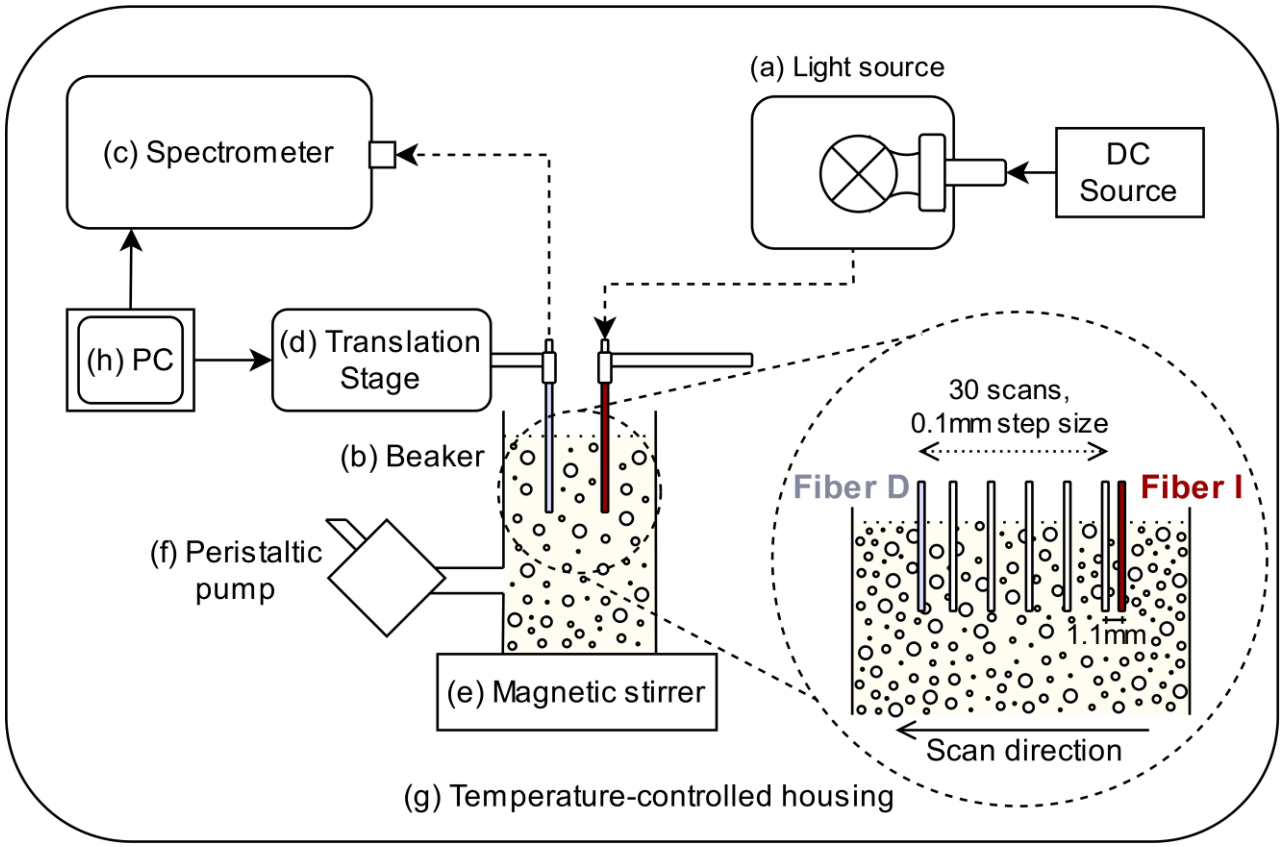

Figure 1: *Conceptual depiction of the experimental fiber-optic SRS setup. Bold black arrows indicate control flow, while dashed black arrows signify the flow of light. A detailed view of the measurement process in the beaker is presented at the bottom-right corner.*

The light source (a) of the setup consists of a 100-Watt halogen lightbulb (64341 HLX-Z/C, Osram Licht AG, Munich, Germany) housed in an aluminum cube (LC6W, Thorlabs) and powered by a

constant direct current driver (HLG-185-12-AB, Mean Well Enterprises, Taipei, Taiwan). The light emitted by the lightbulb is collected and collimated towards an optical fiber (Fiber I).

The illumination (Fiber I) and detection (Fiber D) fibers are low-OH fibers with a 600 µm core diameter and NA of 0.22 (091119, RoMack Inc., Virginia, USA). The sample-side terminations of Fibers I and D are protected by factory-installed stainless steel ferrules with an outer diameter of 1.1 mm and a length of 50 mm. The fiber tips are polished and in direct contact with the sample. The milk sample is presented in a polypropylene beaker (b) with a 75 mm height and a 50 mm diameter (446081, VITLAB GmbH, Grossostheim, Germany). The ends of Fibers I and D that are not in contact with the sample have SMA terminations (SM1SMA, Thorlabs) for connecting them respectively to the light source and to a cooled diode array NIR spectrometer (c) that measures from 970 to 1690 nm with a 2.81 nm pixel resolution (1.7-256 Plane Grating Spectrometer, Carl Zeiss, Jena, Germany).

Fiber I is held in a calibrated fixed position, while Fiber D is mounted on a movable platform (d) controlled by a 25 mm motorized X-translation stage (MTS25/M-Z8, Thorlabs) to move Fiber D in the scan direction. Fiber I and Fiber D were positioned at the same height. To ensure good contact between the fibers and the sample, as well as to minimize the presence of air bubbles at the surface of the fiber tips, both fibers were inserted into the milk sample with their tips at least 25 mm below the surface of the sample and 40 mm away from the bottom of the beaker. Fiber I was positioned at the center of the beaker in the horizontal plane, while Fiber D was positioned a minimum of 20 mm away from the wall of the beaker at all times. Preliminary tests on milk samples with a very high (5.46%) and a very low fat content (2.46%) confirmed that the interaction

of the detected light with the beaker wall, bottom, and sample surface is negligible in this configuration.

A magnetic stirrer (e) is located under the beaker to mix the milk sample at a nominal rotation of 250 RPM (MIXdrive 1 eco, 2Mag AG, Munich, Germany), to prevent changes in the morphology of the sample while measurements are performed. Additionally, a two-channel peristaltic pump (ISMATEC MS-2/12, Cole-Parmer GmbH, Wertheim, Germany) pumps air into the sample at a flow rate of 0.032 ml/min on the side of the beaker to assist in mixing (f). All the elements of the experimental setup are contained in a compact and temperature-controlled (PowerCool, Laird, Liberec, Czech Republic) housing (g) regulated to 36±0.5°C.

For each milk sample introduced into the beaker, the NIR spectrometer recorded a LW-NIR spectrum from 970 to 1690 nm at 30 different distances between Fiber I and Fiber D. In its spatial resolution, the illumination-to-detection distance ranged from 1.1 mm to 4 mm between the centers of fibers I and D, with a scan step size of 0.1 mm. Each measurement was acquired with an integration time of 260 ms. To increase the SNR, 50 repeated measurements were taken and averaged for each of the 30 illumination-to-detection distances per sample. The translation stage and the NIR spectrometer were controlled by a dedicated computer (h) with dedicated acquisition software in LabView 2020 (National Instruments, Austin, USA).

2.2 Acquisition and laboratory analysis of milk samples

The raw milk samples for this experiment were collected with a VMS™ Classic (DeLaval, Tumba, Sweden) automatic milking system (AMS) at the experimental dairy farm 'Hooibeekhoeve' in the province of Antwerp (Geel, Belgium). The AMS was equipped with a Herd Navigator™ sampler

(HNS Supra+, DeLaval) collecting and mixing a representative milk sample (±300 mL) of every single milking (ICAR, 2017a). For each milking, this sampler was configured so that a representative fraction (±30 mL) of the sample was collected in a plastic cup for laboratory analysis. The remainder of the milk sample (at least 200 ml) was collected in a glass container for SRS measurements. The container was submerged in a water bath (Circulator DC10, Haake GmbH, Karlsruhe, Germany) at a nominal temperature of 36°C. No more than 15 minutes elapsed between the end of the milking process and the analysis by the SRS experimental setup. As the milk sample was thoroughly mixed before splitting it, the samples taken for laboratory analysis are representative of those taken for the SRS analyses.

On average, 17 milk samples were extracted and analyzed per day, and measurements were performed during 13 working days spread over a period of 3 weeks. The milk samples taken for laboratory analyses were preserved with bronopol (0.3 mg/mL) at 4°C and analyzed at the Milk Control Center (MCC Vlaanderen, Lier, Belgium) within 3 days after sample collection. In this experiment, the laboratory analysis (fat, protein, and lactose) of each sample was performed according to ISO 9622 (ISO, 2013). These laboratory analyses have an accuracy of < 1.0% coefficient of variation (CV) for fat and < 0.9% CV for protein and lactose.

Thirty-five samples were excluded from the further analysis due to unsuccessful laboratory analyses, either because insufficient milk was available for laboratory analysis or due to contamination with flies. The final dataset consisted of 186 raw milk samples for which both SRS data and laboratory analyses were available.

2.3 Spatially resolved spectroscopy measurements

Once a milk sample for SRS measurement was collected, it was taken to the SRS experimental setup and introduced into the beaker. The acquisition of SRS spectra at the 30 consecutive illumination-to-detection distances was fully automated and took approximately 420 seconds per milk sample. For each of the 186 individual milk samples measured with the SRS setup, 30 different spectra were captured, corresponding to the 30 different illumination-to-detection distances. For each spectrum measured, 256 different spectral data points were obtained in the wavelength range from 970 to 1690 nm. This results in a two-dimensional array of 30 (distances) by 256 (wavelengths) for each of the 186 milk samples. When considering these SRS data as a function of an increasing illumination-to-detection distance, the obtained profiles follow an exponential-like decrease and are further referred to as 'SRS profiles'.

At the end of the sample measurement, the system autonomously returned to the initial position to allow for the manual cleaning of the sample beaker and fibers, followed by the removal of the beaker and thereafter measurement of the dark and white reference spectra. After cleaning, the sample beaker was removed from the measurement position and manually substituted by a dark reference to quantify the stray light and background noise of the spectrometer. To accomplish this, a 1% Acktar Black (Acktar, Hohenaspe, Germany) coated plate was mounted in a 30 mm diameter optical tube (SM30L30, Thorlabs) and positioned 1 mm below the tip of the detection fiber, covering it and blocking stray light. After performing the dark reference measurement, a white reflectance standard (Spectralon SRS-99, Labsphere, North Sutton, USA) was placed 1 mm below the fibers to quantify the possible drift of the system. These reference measurements were

acquired at the closest illumination-to-detection distance (1.1 mm center-to-center) with the same integration time (260 ms) and number of averaged measurements (50) as the sample measurements.

The raw spectral data of the samples and the white and dark reference measurements have integer values ranging from 0 to 65,656 digital counts (DC) originating from the 16-bit analog-to-digital converter of the NIR spectrometer. These raw data were loaded and processed in R version 4.2.1. (R Core Team, 2022). The initial step of data processing involved subtracting the obtained dark reference spectra from the corresponding sample and white reference spectra to correct for stray light in the SRS system and background noise in the NIR spectrometer channels.

2.4 Dynamic range correction

As the distance between the illumination and detection fibers increases with each scan step during the SRS measurement process, the intensity of the light captured by the detection fiber (Fiber D) follows an exponential-like decrease (Crofcheck et al., 2002). This signal loss is attributed to the geometrical dispersion of light with distance in the medium and the combined action of scattering and absorbance by the different milk components, with fat being a primary contributor (Postelmans et al., 2020). This translates into a suboptimal use of the dynamic range of the spectrometer, which decreases the SNR, for larger illumination-to-detection distances.

Therefore, the evaluation of how accurately the milk composition could be predicted from the spectra collected at each individual illumination-to-detection distance would be biased if no correction would be applied for this suboptimal use of the dynamic range for larger distances. It is hypothesized that by performing a dynamic range correction, the milk composition prediction

performance based on the spectra collected at the different distances will be less conditioned by the loss of signal as distance increases with each scan step. To test this, two different scenarios were evaluated: in the non-corrected scenario, the exponential decrease of the signal associated with the illumination-to-detection distance was not compensated; in the corrected scenario, as described later, a correction is performed to equalize the dynamic range and associated SNR of the spectrometer for all illumination-to-detection distances. Given the inability to increase the SNR of the pre-existing SRS data, the SNR for shorter illumination-to-detection distances is reduced to match the SNR at longer distances.

To perform the dynamic range correction, the overall maximum SRS profile was first identified. This is because the maximum signal intensity is the main constraint when optimizing the illumination intensity or the spectrometer response (e.g. gain, integration time) to avoid saturation of the spectrometer detector. To obtain this maximum overall SRS profile, the maximum value over all 186 samples and all 256 wavelengths together was calculated for each of the 30 illumination-to-detection distances individually (Figure 2.a). Next, the maximum value at the largest illumination-to-detection distance was divided by the maximum values of all individual distances to obtain a one-dimensional array of length 30, the dynamic range correction factor (Figure 2.b). This correction factor has values between 0 and 1, being 1 for the largest illumination-to-detection distance. The sample SRS data were corrected for the dynamic range scale of the largest distance by multiplying the SRS profiles with the correction factor and accordingly rounding the spectral intensity values (in DC) to the nearest integer.

2.5 Normalization of the sample spectra

The average white reference spectrum and the average dark reference spectrum for each measurement day were calculated and used to correct and normalize the sample SRS spectra for that day. Specifically, the dark-subtracted sample spectra were divided by the average dark-subtracted white reference spectrum of that day. This normalization technique corrected for drift in the light source intensity and spectrometer sensitivity over the different measurement days (Andersen et al., 2013) without introducing excessive noise from the manual manipulations and the spectrometer. This normalization procedure was applied to both the original sample spectra and the corrected spectra.

2.6 Development and validation of the prediction models

The prediction models were developed and validated using a customized chemometrics toolbox for R (Aernouts et al., 2020). The procedure to develop and validate the prediction models was the same for the non-corrected and corrected scenario.

By applying the Duplex algorithm (Snee, 1977) on the laboratory analysis data (fat, protein and lactose) of the 186 samples, this dataset with its corresponding SRS spectra was split into representative calibration and test sets with respectively two-thirds and one-third of the samples. The Duplex algorithm utilized the Mahalanobis distance in the three-dimensional space of the laboratory analyses to emulate correlations, descriptive statistics, and the entire variability of the dataset, without considering measurement time discrimination.

To remove spectral regions that have an excessively reduced SNR, the wavelengths where the signal intensity of the dark-corrected raw sample spectra was lower than 50 DC for the lowest intensity sample at the shortest distance were excluded, for all samples and all distances. As a result, the outermost regions of the spectral range (1680-1690 nm) were removed due to the lower sensitivity of the spectrometer at these wavelengths. Additionally, the spectral region from 1360 to 1500 nm was discarded due to weak SRS signals, caused by the high absorption of water molecules.

The resulting spectra were preprocessed to reduce non-linearity and scattering effects with techniques that provided the best performance in previous studies involving the same milk components, spectral wavelength range and resolution (Diaz-Olivares et al., 2020). Standard normal variates (SNV) weighting was applied, followed by a first-order Savitzky-Golay derivative involving a second-order polynomial filter with a spectral window length of 15 wavelengths (±40 nm) for fat and protein. For lactose, a second-order Savitzky-Golay derivative with the same window length was applied. Mean centering was performed on the preprocessed spectra and the milk components right before building the partial least squares regression (PLSR) models on the spectra of the calibration set. Individual PLSR models were constructed for each milk component and each individual illumination-to-detection distance, for both non-corrected and corrected spectra, with up to 20 latent variables.

A 10-fold cross-validation with random groups was repeated 100 times for the samples in the calibration set to study the effect of the model complexity (number of latent variables) on the model performance for each PLSR model in terms of the root-mean-squared error of cross-validation (RMSECV). The selection of the number of latent variables prioritized achieving the best

results (lowest RMSECV) for both the original data as well as those corrected for the dynamic range over the entire illumination-to-detection distance.

Milk composition predictions were obtained by applying the PLSR models on the SRS spectra of the samples in the test set and the respective residuals were derived. To evaluate which distances give the best performance for milk composition prediction, the best illumination-to-detection distance was selected as the one with the lowest root-mean-square error of prediction (RMSEP) for samples in the test set. The best distance was selected for each combination of milk components, number of latent variables in the selected range and correction scenario. Next, one-sided paired T-tests ($\alpha = 0.05$) were applied to the squared residuals of the test samples to evaluate whether the best distance has significantly lower residuals compared to the other 29 distances (Cederkvist et al., 2005). The same procedure was also applied to the cross-validation residuals to compare the best illumination-to-detection distance (lowest RMSECV) with the other 29 distances. The results of the latter were reported in the supplementary materials.

3 Results and discussion

3.1 Long-wave near-infrared spatially resolved reflectance spectra

In Figure 2.a, the collected SRS data at the different wavelengths are plotted as a function of the illumination-to-detection distance. These SRS profiles show an exponential-like decrease of the signal with increasing distance. This clearly illustrates that the longer illumination-to-detection distances make less efficient use of the dynamic range (0 to 65,656 DC) of the spectrometer compared to the shorter distances. The maximum signals are obtained at around 1050 nm, where the detector sensitivity is high and the light absorption by the samples is relatively low. The overall maximum SRS profile is represented by a black line in Figure 2.a, while the derived dynamic range correction factor is shown in Figure 2.b. This correction factor increases from approximately 0.1 for the shortest illumination-to-detection distance (1.1 mm) to 1 for the largest distance (4 mm).

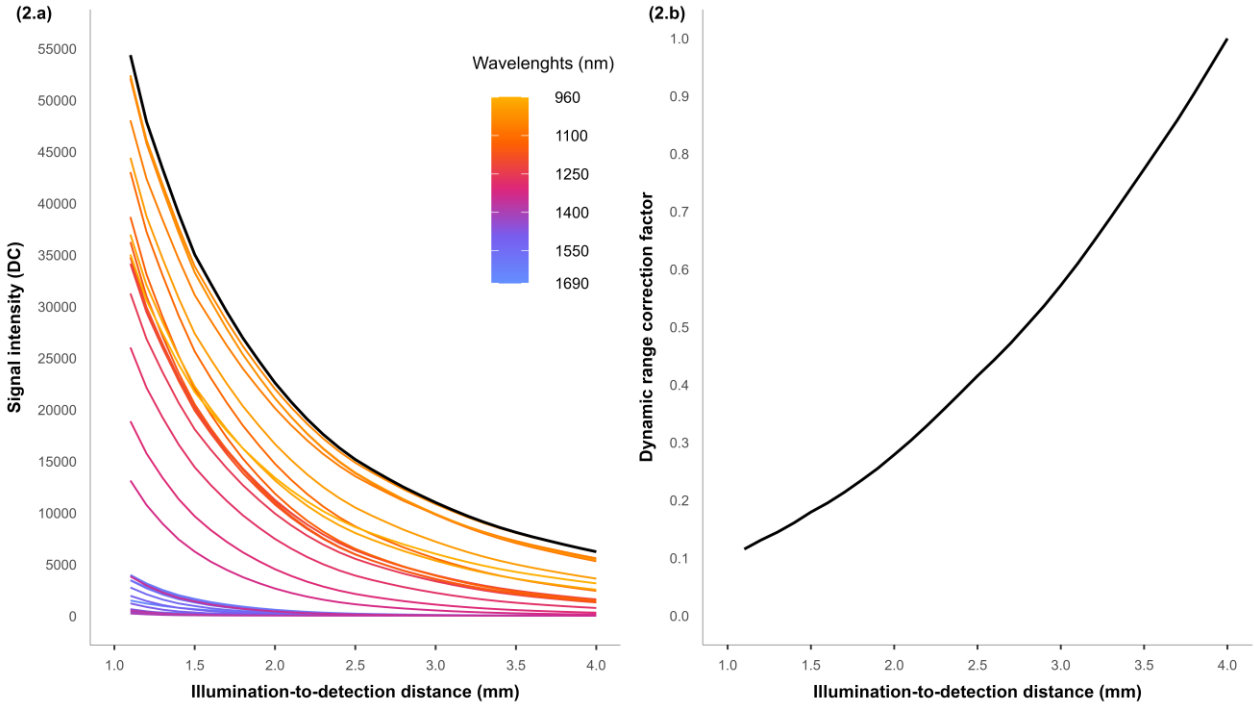

Figure 2. *(2.a) Depiction of the SRS profiles demonstrating the decreasing signal intensity in digital counts (DC) as the illumination-to-detection distance (in mm) increases, for the milk sample with the highest fat content (5.46%). The line colors of the SRS profile vary in function of the spectral wavelength, with only every eighth wavelength shown to avoid overcrowding the figure. The overall maximum SRS profile is shown as a thicker black line; (2.b) Dynamic range correction factor applied to compensate for the poorer use of the dynamic range of the spectrometer as the illumination-to-detection distance increases.*

In Figure 3, the SRS spectra of all 186 milk samples are presented for three different equidistant measuring points (1.1, 2.5 and 4 mm), for both non-corrected (3.a.1, 3.a.2 and 3.a.3) and corrected (3.c.1, 3.c.2 and 3.c.3) spectra.

For all scenarios, the SRS signal decreases around 970, 1200 and 1450 nm, indicating absorbance by covalent O-H bonds of water. The highest absorbance of light can be observed between 1360 to 1500 nm, causing a low SNR. This range was thus removed when developing the PLSR models.

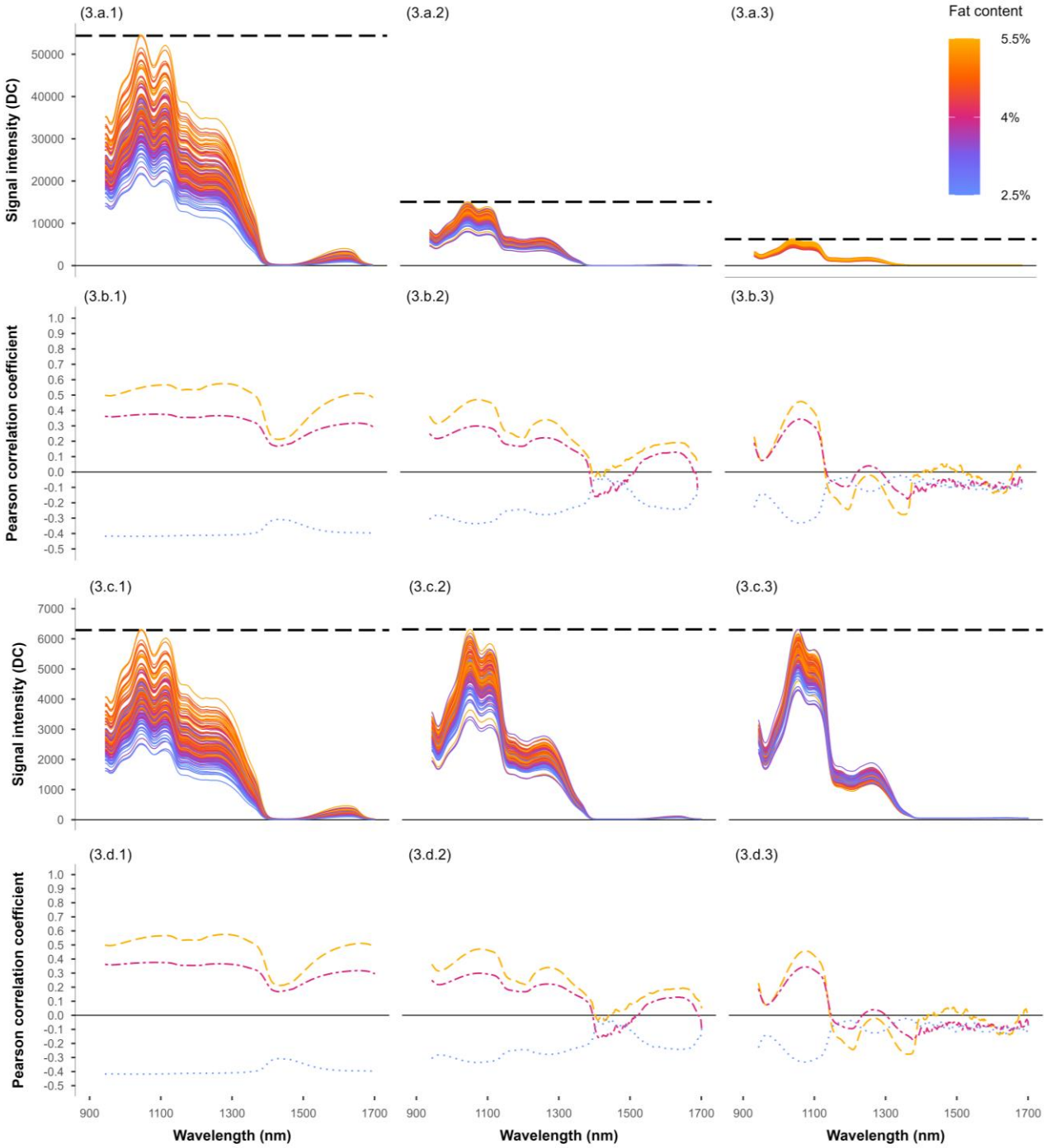

Figure 3. *Illustration of the SRS spectra collected after removing dark reference (solid orange-to-blue gradient lines) for all 186 milk samples at three different equidistant illumination-to-detection distances: 1.1, 2.5 and 4 mm. Orange represents higher fat content, while blue indicates lower fat*

content. The non-corrected SRS spectra for the three distances are depicted in Figures 3.a.1, 3.a.2 and 3.a.3, while Figures 3.c.1, 3.c.2 and 3.c.3 illustrate the dynamic range corrected SRS spectra. In these six subfigures, the overall maximum signal intensity is indicated with a horizontal black dashed line. The Pearson correlation coefficients between the SRS spectra and the milk components are depicted for the non-corrected scenario (Figures 3.b.1, 3.b.2 and 3.b.3) and the corrected scenario (Figures 3.d.1, 3.d.2 and 3.d.3). These coefficients represent the linear correlation between the SRS spectra at specific illumination-to-detection distances for all the milk samples and their fat (orange dashed line), protein (magenta dot-dashed line), and lactose (blue dotted line) concentration, for the entire spectral range measured.

Figures 3.a.1, 3.a.2 and 3.a.3 clearly illustrate that the overall maximum signal level decreases with the increasing illumination-to-detection distance. After applying the dynamic range correction, the overall maximum signal level is the same for every distance (Figures 3.c.1, 3.c.2 and 3.c.3). All data in Figures 3.a.1, 3.a.2, 3.a.3, 3.c.1, 3.c.2 and 3.c.3 are integers. As a result, all distances in Figures 3.c.1, 3.c.2 and 3.c.3 make use of the same dynamic range, albeit with a higher variability observed at shorter distances (Figure 3.c.1). No significant difference is observed when comparing the correlations between signal and composition for the corrected and non-corrected scenarios. This could have been expected as the scaling step of the dynamic range correction is a linear operation, while the rounding operation has nearly no effect on the correlations.

The orange-to-blue color gradient represents the variability in fat content and its influence on the acquired spectra. In general, an increased fat content results in stronger light scattering due to the higher number of fat particles. At short distances, there is an overall increase of light scattered

into the detection fiber with increasing fat content of the milk sample over nearly the entire wavelength range (Figures 3.a.1 and 3.c.1), in line with the observation reported by Crofcheck et al. (2002). This translates into a positive correlation between the fat content and the SRS signal intensity at short distances which is quasi-constant ($r > 0.5$, except for the water absorption band between 1360 and 1500 nm) over the entire wavelength range (Figures 3.b.1 and 3.d.1). Conditioned to the influence of fat particles, a weak positive correlation can be observed between the protein content and the signal intensity for short distances ($r > 0.35$, except between 1360 and 1500 nm). This relationship can be mainly attributed to the typical genetic correlation between the protein and fat content of milk originating from individual cows (Soyeurt et al., 2007). A weak negative correlation ($r < 0.25$, except between 1360 and 1500 nm) was found between milk lactose content and the SRS spectra. The correlation coefficients are nearly the opposite of the ones for fat, suggesting a negative correlation between the lactose and fat concentration in milk samples. None of the correlation coefficients display clear absorption peaks linked to the dry matter components of raw milk.

In contrast to the shorter illumination-to-detection distances, the correlation between the signal intensity and fat, protein and lactose content at longer distances is less consistent over the measured wavelength range (Figures 3.b.3 and 3.d.3). Similar correlations as those at short distances can be found at longer distances for spectral regions where absorption is relatively low (1030 to 1120 nm). However, for the other spectral regions where absorption is more dominant, the correlation changes. For fat, this correlation is negative for almost all wavelengths above 1150 nm. This phenomenon has also been described by Crofcheck et al. (2002) and Kumar and Schmitt (1996). Nonetheless, the signal intensity of measured SRS spectra at long distances is mainly

limited by the water absorption, as the longer travel path of the light increases the chance of absorption. Moreover, it is observed that at shorter illumination-to-detection distances some spectral signal can be measured in the wavelength range between 1390 and 1690 nm, while this signal is almost completely lost at the longest illumination-to-detection distance (4 mm). This wavelength range from 1390 to 1690 nm overlaps largely with the absorption bands of lactose linked to the O-H and C-H stretching vibrations (Scheibelhofer et al., 2018). It was hypothesized that a larger illumination-to-detection distance would capture the increased interaction of light with milk serum, carrying more information about lactose content. However, due to the loss of signal intensity, the prediction performance of lactose at these longer distances may be compromised.

3.2 Development and validation of the prediction models

The sample dataset was split into a calibration (n = 120) and test set (n = 66). The descriptive statistics of the composition of these sets and their correlations are summarized in Table 1. Both populations were compared with a two-sample t-test, showing there was no significant difference (α = 0.05) between them. Therefore, the data splits made by the duplex algorithm are representative of the whole dataset, which is crucial for building a robust prediction model (Saeys et al., 2008). In addition, both groups have composition values and variability in line with those of the cow population that is monitored in the Flemish milk recording programs (Aernouts et al., 2011).

Table 1. *Basic statistics and Pearson correlations for the main milk components in the calibration and test sets.*

Component	Calibration (n = 120)						Test (n = 66)					
	Basic statistics (% wt/wt)				Pearson corr.		Basic statistics (% wt/wt)				Pearson corr.	
	Mean	SD	Min	Max	Fat	Protein	Mean	SD	Min	Max	Fat	Protein
Fat	4.04	0.58	2.49	5.34	1	-	3.99	0.66	2.46	5.46	1	-
Protein	3.37	0.25	2.55	3.98	0.41	1	3.37	0.28	2.74	4.03	0.46	1
Lactose	4.69	0.13	4.38	5.07	-0.17	-0.19	4.71	0.15	4.31	5.05	-0.21	-0.24

Figure 4 shows the prediction performance on the test set of the different models for fat (4.a), protein (4.b) and lactose (4.c), for both the non-corrected (4.a.1, 4.b.1 and 4.c.1) and corrected (4.a.2, 4.b.2 and 4.c.2) scenarios. The results are plotted for a range of latent variables of the PLSR models, including the number of latent variables giving the best performance (lowest RMSECV and RMSEP) for both the non-corrected and corrected scenario. By considering a range of latent variables, the effect of local minima can be separated from the actual trends to evaluate the model performances for a certain component and scenario.

The illumination-to-detection distance with the lowest RMSEP is represented with a squared point, while all other distances that show no significant difference in their prediction performance are indicated with a triangular point. Distances that display a significantly worse performance are indicated with a hollow circular point.

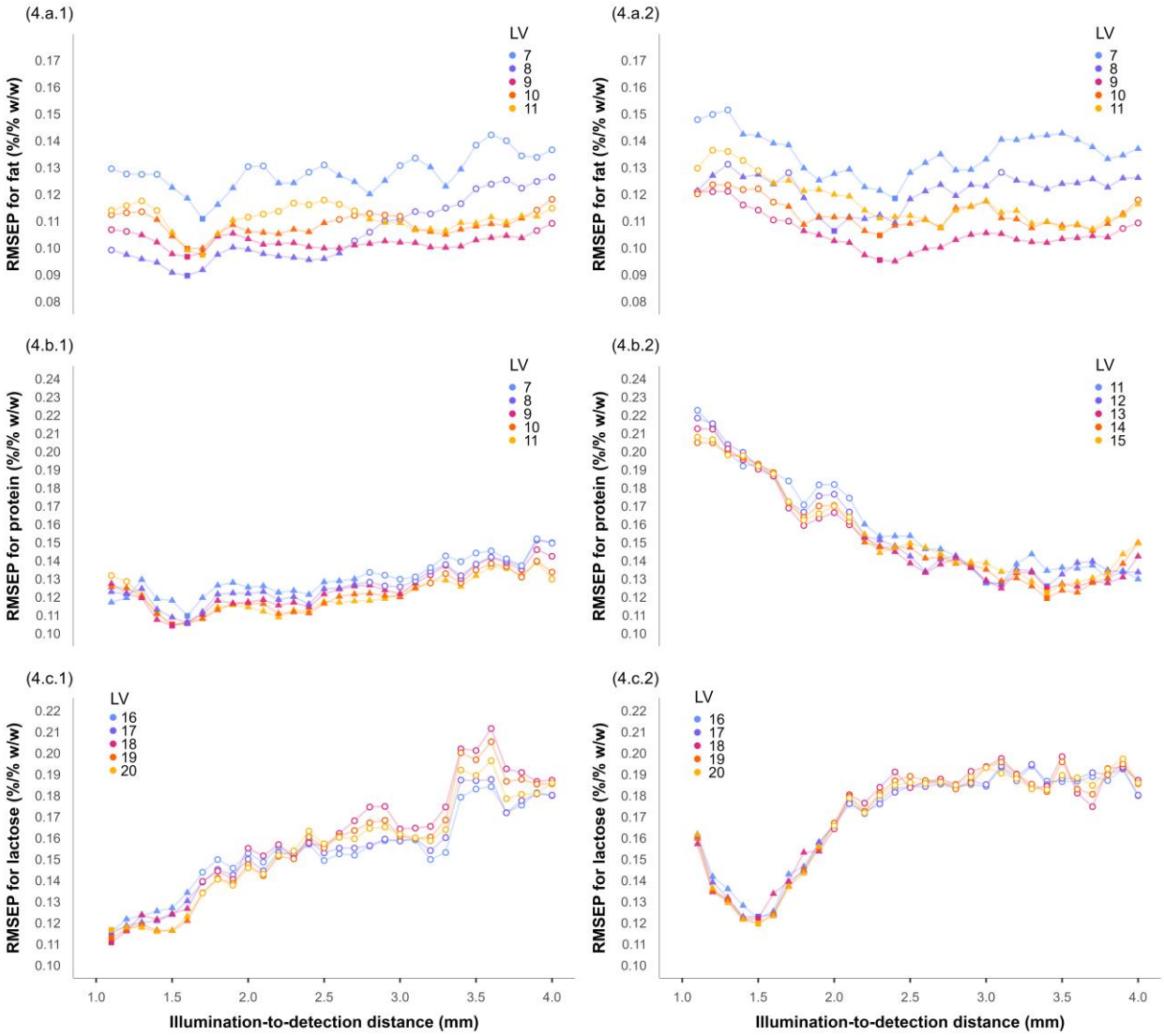

Figure 4. *Plots of the root-mean-square error of prediction (RMSEP) of the milk composition content for the test set without dynamic range correction for (4.a.1) fat, (4.b.1) protein, and (4.c.1) lactose and with correction (4.a.2, 4.b.2 and 4.c.2). The minimum RMSEP values are indicated by squares, while triangles indicate no significant ($\alpha = 0.05$) differences between the prediction residuals of each scan distance and the minimum. LV = number of latent variables used by the*

model, with a selected range of latent variables around the optimal number to not overcrowd the figure.

For all three milk components, good predictions are obtained at their best prediction distance, with RMSEP values under 0.13% (wt/wt), for protein and lactose, and lower than 0.10% (wt/wt) for fat. In general, it can be observed that for the non-corrected scenario, the lowest RMSEP values for all components are obtained at relatively short distances. This supports the hypothesis that the poorer use of the dynamic range of the spectrometer at the longer illumination-to-detection distances penalizes the milk composition predictions based upon these measurements when compared with shorter distances. This may be attributed to the reduced SNR resulting from a suboptimal use of the dynamic range at longer distances.

To exemplify, better prediction results are observed when short illumination-to-detection distances (1.3 – 2.6 mm) are selected in the non-corrected scenario for fat (Figure 4.a.1). Once the dynamic range correction is applied (Figure 4.a.2), no significant differences in the distance used for fat composition prediction can be observed, except for short or extremely long distances. The entire span of distances from 1.8 to 3.8 mm performs equally well for predicting the fat composition when the optimal number of latent variables (8 and 9 respectively) is selected.

For protein, the difference between scenarios is distinct (Figures 4.b.1 and 4.b.2). The non-corrected scenario favors shorter illumination-to-detection distances (1.1 – 3.1 mm), while the dynamic range correction scenario prefers distances from 2.2 to 4 mm, for the optimal number of latent variables (9 and 13 respectively).

For lactose, the difference in the distance with the lowest RMSEP selected between correction scenarios is less significant (Figures 4.c.1 and 4.c.2). In the non-corrected scenario, adequate distances to predict lactose are found in the 1.1 to 1.6 mm interval with the best performance at 1.1 mm. The dynamic range correction slightly increased the size of this interval (1.1 – 1.9 mm) and the best performance moved to 1.5 mm. The PLSR models for lactose had the highest complexity (18 and 19 latent variables respectively). As seen in Figure 3, only the SRS data acquired at short illumination-to-detection distances provide useful signals between 1390 and 1690 nm. This wavelength interval is relevant for the prediction of lactose, as it carries information related to absorption bands of O-H and C-H functional groups (de Lima et al., 2018). That is likely the reason why the best lactose predictions are obtained at short distances, even after implementing the dynamic range correction. Therefore, our hypothesis that lactose predictions would improve based on SRS spectra collected at longer distances, acquiring light that penetrated deeper into the milk sample, has been disproven.

Considering the previous results, there is no overlap between the best protein and lactose prediction distance ranges, while the determination of fat percentage seems to perform well in the 1.8 to 3.8 mm range disregarding the selected distance. Therefore, to obtain the best results in raw milk quality analysis, it is recommended to combine a short illumination-to-detection distance (between 1.1 to 1.8 mm) to measure lactose and a long distance (between 2.2 and 3.8 mm) to measure fat and protein. This disposition will allow for the simultaneous determination of fat, protein and lactose content while minimizing the number of fibers and measurement time.

The requirements of the International Committee for Animal Recording (ICAR) for on-farm milk analyzers are used to evaluate the predictive capabilities of the approach in different on-farm

situations. As the RMSEP values for protein and lactose were lower than 0.13% (wt/wt) and those for fat were below 0.10% (wt/wt), which is well below the ICAR limits for at-line (0.2%) and in-line (0.25%) milk analyses (ICAR, 2017b).

The current approach was limited by the loss of signal intensity with increasing distance, due to the fixed illumination intensity and spectrometer integration time employed for all spectral acquisitions. Increasing the integration time as distance increases could overcome this, but it would result in an excessively long measurement time, where the morphology of the sample could change significantly during the measurement. Therefore, to counteract this light loss, it is recommended to focus on augmenting the injected light into the sample for larger illumination-to-detection distances to flatten the exponential-like shape of the measured SRS profiles. This physical compensation is expected to enhance the SNR at longer illumination-to-detection distances, which could not be obtained with the dynamic range correction implemented in this study. It is hypothesized that this would especially improve the prediction capability for protein, as this measurement performs better at longer distances. Furthermore, it would allow a more thorough examination of lactose prediction capabilities, as the spectral signal and SNR between 1390 and 1690 nm will be higher at far illumination-to-detection distances when illumination intensities increase for farther distances.

Moreover, it should be noted that this study was limited to the prediction of the milk components using the SRS spectra collected at single illumination-to-detection distances. Even better prediction performances might be obtained by employing a multiblock PLS approach (Biancolillo and Næs, 2019; Liland et al., 2016) combining the information acquired at different distances to perform a low-level data fusion (Cocchi, 2019; Hayes et al., 2023). This approach not only

improves performance but also allows for individual examination of the redundancy or uniqueness of each distance or block of distances (Westerhuis et al., 1998).

As currently presented, this technology can be implemented in-line within milking machines or dairy processing lines to produce autonomous, non-destructive milk quality predictions. Nevertheless, exploring the applicability of this approach to other turbid foods would be of interest, serving to validate its use in the field of food quality control.

4 Conclusions

This study has assessed the applicability of a fiber-optic LW-NIR SRS approach for raw milk quality analysis. The PLSR prediction accuracy using the best illumination-to-detection distances was good for all three milk components (RMSEP < 0.10% for fat, < 0.13% for protein and lactose) as these performances meet the ICAR guidelines for in-line and at-line milk analysis systems. With regards to illumination-to-detection distance performance, two regions of interest have been identified: (1) between 1.1 to 1.8 mm for lactose and (2) between 2.2 and 3.8 mm for fat and protein. A fiber-optic SRS reflectance probe including these distances would allow to simultaneously determine the fat, protein, and lactose content of raw milk with high accuracy.

5 Acknowledgments

José A. Díaz Olivares received funding from the Research Foundation Flanders (FWO, Belgium) through PhD fellowship strategic basic research No. 1S76320N. This research was financially supported by KU Leuven internal funding (C3 project C3/19/037). The authors gratefully acknowledge the experimental dairy farm of the province of Antwerp, 'Hooibeekhoeve' (Geel, Belgium), for permitting the collection of raw milk samples. The authors have not stated any conflicts of interest.

---

6 References

Aernouts, B., Adriaens, I., Diaz-Olivares, J.A., Saeys, W., Mäntysaari, P., Kokkonen, T., Mehtiö, T., Kajava, S., Lidauer, P., Lidauer, M.H., Pastell, M., 2020. Mid-infrared spectroscopic analysis

of raw milk to predict the blood nonesterified fatty acid concentrations in dairy cows. J Dairy Sci 103, 6422–6438. https://doi.org/10.3168/JDS.2019-17952

Aernouts, B., Polshin, E., Lammertyn, J., Saeys, W., 2011. Visible and near-infrared spectroscopic analysis of raw milk for cow health monitoring: Reflectance or transmittance? J Dairy Sci 94, 5315–5329. https://doi.org/10.3168/jds.2011-4354

Aernouts, B., van Beers, R., Watté, R., Huybrechts, T., Lammertyn, J., Saeys, W., 2015a. Visible and near-infrared bulk optical properties of raw milk. J Dairy Sci 98, 6727–6738. https://doi.org/10.3168/jds.2015-9630

Aernouts, B., van Beers, R., Watté, R., Huybrechts, T., Jordens, J., Vermeulen, D., van Gerven, T., Lammertyn, J., Saeys, W., 2015b. Effect of ultrasonic homogenization on the Vis/NIR bulk optical properties of milk. Colloids Surf B Biointerfaces 126, 510–519. https://doi.org/10.1016/J.COLSURFB.2015.01.004

Andersen, H.V., Wedelsback, H., Hansen, P.W., 2013. A white paper from FOSS: NIR spectrometer technology comparison, in: FOSS P/N. pp. 1–14.

Augustin, M.A., Udabage, P., Juliano, P., Clarke, P.T., 2013. Towards a more sustainable dairy industry: Integration across the farm–factory interface and the dairy factory of the future. Int Dairy J 31, 2–11. https://doi.org/10.1016/J.IDAIRYJ.2012.03.009

Bendoula, R., Ducanchez, A., Herrero-Langreo, A., Guerrero-Castro, P., Roger, J.M., Gobrecht, A., 2019. Effect of the architecture of fiber-optic probes designed for soluble solid content prediction in intact sugar beet slices. Sensors 19, 2995. https://doi.org/10.3390/S19132995

Biancolillo, A., Næs, T., 2019. The Sequential and Orthogonalized PLS Regression for Multiblock Regression: Theory, Examples, and Extensions. Data Handling in Science and Technology 31, 157–177. https://doi.org/10.1016/B978-0-444-63984-4.00006-5

Bogomolov, A., Belikova, V., Galyanin, V., Melenteva, A., Meyer, H., 2017. Reference-free spectroscopic determination of fat and protein in milk in the visible and near infrared region below 1000 nm using spatially resolved diffuse reflectance fiber probe. Talanta 167, 563–572. https://doi.org/10.1016/J.TALANTA.2017.02.047

Bogomolov, A., Melenteva, A., 2013. Scatter-based quantitative spectroscopic analysis of milk fat and total protein in the region 400–1100 nm in the presence of fat globule size variability. Chemometrics and Intelligent Laboratory Systems 126, 129–139. https://doi.org/10.1016/J.CHEMOLAB.2013.02.006

Cederkvist, H.R., Aastveit, A.H., Næs, T., 2005. A comparison of methods for testing differences in predictive ability. J Chemom 19, 500–509. https://doi.org/10.1002/CEM.956

Cocchi, M., 2019. Introduction: Ways and Means to Deal With Data From Multiple Sources. Data Handling in Science and Technology 31, 1–26. https://doi.org/10.1016/B978-0-444-63984-4.00001-6

Crofcheck, C.L., Payne, F.A., Mengüç, M.P., 2002. Characterization of milk properties with a radiative transfer model. Applied Optics, 41, 2028–2037. https://doi.org/10.1364/AO.41.002028

de Lima, G.F., Andrade, S.A.C., da Silva, V.H., Honorato, F.A., 2018. Multivariate Classification of UHT Milk as to the Presence of Lactose Using Benchtop and Portable NIR Spectrometers. Food Anal Methods 11, 2699–2706. https://doi.org/10.1007/s12161-018-1253-7

Diaz-Olivares, J.A., Adriaens, I., Stevens, E., Saeys, W., Aernouts, B., 2020. Online milk composition analysis with an on-farm near-infrared sensor. Comput Electron Agric 178, 105734. https://doi.org/10.1016/j.compag.2020.105734

Ellis, D.I., Brewster, V.L., Dunn, W.B., Allwood, J.W., Golovanov, A.P., Goodacre, R., 2012. Fingerprinting food: Current technologies for the detection of food adulteration and contamination. Chem Soc Rev 41, 5706–5727. https://doi.org/10.1039/C2CS35138B

Foschum, F., Jäger, M., Kienle, A., 2011. Fully automated spatially resolved reflectance spectrometer for the determination of the absorption and scattering in turbid media. Review of Scientific Instruments 82, 103104. https://doi.org/10.1063/1.3648120

Gowen, A.A., O'Donnell, C.P., Cullen, P.J., Downey, G., Frias, J.M., 2007. Hyperspectral imaging - an emerging process analytical tool for food quality and safety control. Trends Food Sci Technol 18, 590–598. https://doi.org/10.1016/J.TIFS.2007.06.001

Gunasekaran, S., 1996. Computer vision technology for food quality assurance. Trends Food Sci Technol 7, 245–256. https://doi.org/10.1016/0924-2244(96)10028-5

Hayes, E., Greene, D., O'Donnell C., O'Shea N., Fenelon M. A., 2023. Spectroscopic technologies and data fusion: Applications for the dairy industry. Frontiers in Nutrition 9. https://doi.org/10.3389/fnut.2022.1074688

Huang, Y., Lu, R., Chen, K., 2018. Assessment of tomato soluble solids content and pH by spatially-resolved and conventional Vis/NIR spectroscopy. J Food Eng 236, 19–28. https://doi.org/10.1016/J.JFOODENG.2018.05.008

ICAR, 2017a. Guidelines for Testing. Approval and Checking of Milk Recording Devices. Section 11.

ICAR, 2017b. Guidelines for On-Farm Milk Analysis. Section 13.

ISO, 2013. Milk and Liquid Milk Products–Guidelines for the Application of Mid-Infrared Spectrometry. ISO Norm 9622: 2013/IDF 141: 2013.

Karoui, R., de Baerdemaeker, J., 2007. A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. Food Chem 102, 621–640. https://doi.org/10.1016/J.FOODCHEM.2006.05.042

Kawamura, S., Kawasaki, M., Hiroki, A.E., Ae, N., Natsuga, M., 2007. Near-infrared spectroscopic sensing system for online monitoring of milk quality during milking. Sensing and Instrumentation for Food Quality and Safety 2007 1, 37–43. https://doi.org/10.1007/S11694-006-9001-X

Kawasaki, M., Kawamura, S., Tsukahara, M., Morita, S., Komiya, M., Natsuga, M., 2008. Near-infrared spectroscopic sensing system for on-line milk quality assessment in a milking robot. Comput Electron Agric 63, 22–27. https://doi.org/10.1016/J.COMPAG.2008.01.006

Kong, D., Shi, L., Yang, Z., 2019. Product recalls, corporate social responsibility, and firm value: Evidence from the Chinese food industry. Food Policy 83, 60–69. https://doi.org/10.1016/J.FOODPOL.2018.11.005

Kumar, G., Schmitt, J.M., 1996. Spectral Distortions in Near-Infrared Spectroscopy of Turbid Materials. Applied Spectroscopy, 50, 1066–1073. https://doi.org/10.1366/0003702963905295

Kunes, R., Bartos, P., Iwasaka, G.K., Lang, A., Hankovec, T., Smutny, L., Cerny, P., Poborska, A., Smetana, P., Kriz, P., Kernerova, N., 2021. In-Line Technologies for the Analysis of Important Milk Parameters during the Milking Process: A Review. Agriculture 2021, 11, 239. https://doi.org/10.3390/AGRICULTURE11030239

Liland, K.H., Næs, T., Indahl, U.G., 2016. ROSA—a fast extension of partial least squares regression for multiblock data analysis. J Chemom 30, 651–662. https://doi.org/10.1002/CEM.2824

Ma, T., Zhao, J., Inagaki, T., Su, Y., Tsuchikawa, S., 2022. Rapid and nondestructive prediction of firmness, soluble solids content, and pH in kiwifruit using Vis–NIR spatially resolved spectroscopy. Postharvest Biol Technol 186, 111841. https://doi.org/10.1016/J.POSTHARVBIO.2022.111841

Mankins, J.C., 1995. Technology readiness levels. White Paper, NASA, Washington, DC, 1995.

Mäntysaari, P., Mäntysaari, E.A., Kokkonen, T., Mehtiö, T., Kajava, S., Grelet, C., Lidauer, P., Lidauer, M.H., 2019. Body and milk traits as indicators of dairy cow energy status in early lactation. J Dairy Sci 102, 7904–7916. https://doi.org/10.3168/JDS.2018-15792

Marcone, M.F., Wang, S., Albabish, W., Nie, S., Somnarain, D., Hill, A., 2013. Diverse food-based applications of nuclear magnetic resonance (NMR) technology. Food Research International 51, 729–747. https://doi.org/10.1016/J.FOODRES.2012.12.046

McParland, S., Lewis, E., Kennedy, E., Moore, S.G., McCarthy, B., O'Donovan, M., Butler, S.T., Pryce, J.E., Berry, D.P., 2014. Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. J Dairy Sci 97, 5863–5871. https://doi.org/10.3168/jds.2014-8214

Melfsen, A., Hartung, E., Haeussermann, A., 2012. Accuracy of in-line milk composition analysis with diffuse reflectance near-infrared spectroscopy. J Dairy Sci 95, 6465–6476. https://doi.org/10.3168/jds.2012-5388

Nguyen Do Trong, N., Erkinbaev, C., Tsuta, M., de Baerdemaeker, J., Nicolaï, B., Saeys, W., 2014a. Spatially resolved diffuse reflectance in the visible and near-infrared wavelength range for non-destructive quality assessment of 'Braeburn' apples. Postharvest Biol Technol 91, 39–48. https://doi.org/10.1016/J.POSTHARVBIO.2013.12.004

Nguyen Do Trong, N., Rizzolo, A., Herremans, E., Vanoli, M., Cortellino, G., Erkinbaev, C., Tsuta, M., Spinelli, L., Contini, D., Torricelli, A., Verboven, P., de Baerdemaeker, J., Nicolaï, B., Saeys, W., 2014b. Optical properties–microstructure–texture relationships of dried apple slices: Spatially resolved diffuse reflectance spectroscopy as a novel technique for analysis and process control. Innovative Food Science & Emerging Technologies 21, 160–168. https://doi.org/10.1016/J.IFSET.2013.09.014

OECD, 2022. OECD-FAO Agricultural Outlook 2022-2031. OECD-FAO Agricultural Outlook. https://doi.org/10.1787/19991142

Paiva, C.L., 2013. Quality Management: Important Aspects for the Food Industry. Food Industry. https://doi.org/10.5772/53162

Postelmans, A., Aernouts, B., Jordens, J., van Gerven, T., Saeys, W., 2020. Milk homogenization monitoring: Fat globule size estimation from scattering spectra of milk. Innovative Food Science & Emerging Technologies 60, 102311. https://doi.org/10.1016/J.IFSET.2020.102311

R Core Team, 2022. R: A Language and Environment for Statistical Computing.

Rodenburg, J., 2017. Robotic milking: Technology, farm design, and effects on work flow. J Dairy Sci 100, 7729–7738. https://doi.org/10.3168/JDS.2016-11715

Saeys, W., Beullens, K., Lammertyn, J., Ramon, H., Naes, T., 2008. Increasing Robustness against Changes in the Interferent Structure by Incorporating Prior Information in the Augmented Classical Least-Squares Framework. Anal Chem 80, 4951–4959. https://doi.org/10.1021/AC800155N

Saranwong, S., Kawano, S., 2008. System Design for Non-Destructive near Infrared Analyses of Chemical Components and Total Aerobic Bacteria Count of Raw Milk. 16, 389–398. https://doi.org/10.1255/JNIRS.807

Scheibelhofer, O., Wahl, P., Larchevêque, B., Chauchard, F., Khinast, J., 2018. Spatially Resolved Spectral Powder Analysis: Experiments and Modeling. Appl Spectrosc 72, 000370281774983. https://doi.org/10.1177/0003702817749839

Shenk, J.S., Workman, J.J., Westerhaus, M.O., 2007. Application of NIR spectroscopy to agricultural products. Handbook of Near-Infrared Analysis, Third Edition 347–386. https://doi.org/10.1201/9781420007374

Snee, R.D., 1977. Validation of Regression Models: Methods and Examples. Technometrics 19, 415–428. https://doi.org/10.1080/00401706.1977.10489581

Soyeurt, H., Gillon, A., Vanderick, S., Mayeres, P., Bertozzi, C., Gengler, N., 2007. Estimation of heritability and genetic correlations for the major fatty acids in bovine milk. J Dairy Sci 90, 4435–4442. https://doi.org/10.3168/jds.2007-0054

Thornton, P.K., 2010. Livestock production: recent trends, future prospects. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 2853. https://doi.org/10.1098/RSTB.2010.0134

Torricelli, A., Spinelli, L., Vanoli, M., Leitner, M., Nemeth, A., Trong, N.N.D., Nicolaï, B., Saeys, W., 2013. Optical coherence tomography (OCT), space-resolved reflectance spectroscopy (SRS) and time-resolved reflectance spectroscopy (TRS): principles and applications to food microstructures. Food Microstructures: Microscopy, Measurement and Modelling 132–162. https://doi.org/10.1533/9780857098894.1.132

Tsenkova, R., Atanassova, S., Toyoda, K., Ozaki, Y., Itoh, K., Fearn, T., 1999. Near-Infrared Spectroscopy for Dairy Management: Measurement of Unhomogenized Milk Composition. J Dairy Sci 82, 2344–2351. https://doi.org/10.3168/JDS.S0022-0302(99)75484-6

van Beers, R., Aernouts, B., León Gutiérrez, L., Erkinbaev, C., Rutten, K., Schenk, A., Nicolaï, B., Saeys, W., 2015. Optimal Illumination-Detection Distance and Detector Size for Predicting Braeburn Apple Maturity from Vis/NIR Laser Reflectance Measurements. Food Bioproc Tech 8, 2123–2136. https://doi.org/10.1007/S11947-015-1562-4

Vanoli, M., van Beers, R., Sadar, N., Rizzolo, A., Buccheri, M., Grassi, M., Lovati, F., Nicolaï, B., Aernouts, B., Watté, R., Torricelli, A., Spinelli, L., Saeys, W., Zanella, A., 2020. Time- and spatially-resolved spectroscopy to determine the bulk optical properties of 'Braeburn' apples after ripening in shelf life. Postharvest Biol Technol 168, 111233. https://doi.org/10.1016/J.POSTHARVBIO.2020.111233

Walstra, P., Wouters, J.T.M., Geurts, T.J., 2005. Dairy Science and Technology. Dairy Sci Technol. https://doi.org/10.1201/9781420028010

Watté, R., Aernouts, B., van Beers, R., Postelmans, A., Saeys, W., 2016. Computational optimization of the configuration of a spatially resolved spectroscopy sensor for milk analysis. Anal Chim Acta 917, 53–63. https://doi.org/10.1016/J.ACA.2016.02.041

Westerhuis, J.A., Kourti, T., Macgregor, J.F., 1998. Analysis of multiblock and hierarchical PCA and PLS models. https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S

Wetzel, D.L., LeVine, S.M., 2000. Biological applications of infrared microspectroscopy. Infrared and raman spectroscopy of biological materials 101–142. ISBN 9780824704094.

7 Supplementary materials

For each milk component, number of latent variables in the selected range and correction scenario combination, the best illumination-to-detection distance was selected as the one with the lowest RMSECV. Next, for each combination, the testing was conducted to identify the other 29 distances that resulted in a prediction of the milk component that was not significantly worse compared to the best distance. To this end, one-sided paired T-tests ($\alpha = 0.05$) were applied to the squared residuals of the cross-validated samples of the calibration set to compare the best distance to the other distances (Cederkvist et al., 2005).

Good predictions were achieved for all three milk components at their optimal cross-validation distance, yielding RMSECV values below 0.10% (wt/wt) for fat, less than 0.14% (wt/wt) for protein, and under 0.12% (wt/wt) for lactose. In the non-corrected scenario, the lowest RMSECV values for all components were observed at relatively short distances, attributed to the poorer SNR at longer illumination-to-detection distances.

In this scenario, shorter illumination-to-detection distances (1.2 – 2.5 mm) displayed better prediction results for fat (Figure 5.a.1), whereas the significance of different distances was expanded (1.6 – 4 mm) in the corrected scenario (Figure 5.a.2), for 9 latent variables in both cases.

Regarding protein (Figures 5.b.1 and 5.b.2), the non-corrected scenario favored distances within the range of 1.1 – 2.8 mm, whereas the corrected scenario had the best performance in distances between 2.1 – 4 mm, with an optimal number of latent variables of 9 and 13, respectively.

For lactose (Figures 5.c.1 and 5.c.2), the non-corrected scenario identified suitable prediction distances within the interval of 1.1 to 1.5 mm, with the best performance achieved at 1.2 mm.

The application of dynamic range correction slightly expanded this interval to 1.1 – 1.9 mm, while the peak performance shifted to 1.5 mm. The PLSR models for lactose exhibited the highest complexity, requiring 18 latent variables

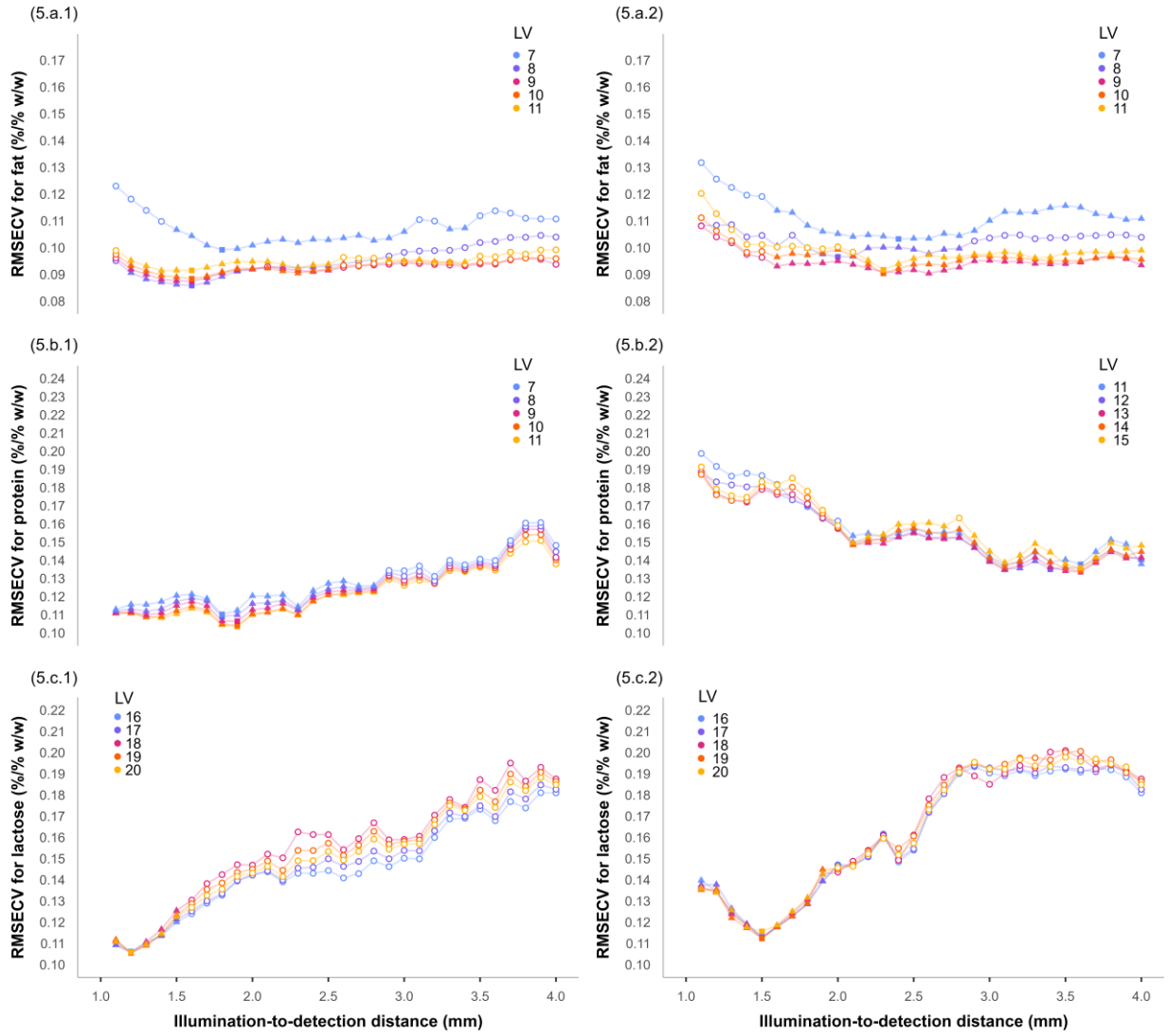

Figure 5. *Plots of the root-mean-square error of cross-validation (RMSECV) of the milk composition content for the calibration set without dynamic range correction for (5.a.1) fat, (5.b.1) protein, and (5.c.1) lactose and with correction (5.a.2, 5.b.2 and 5.c.2). The minimum RMSECV*

values are indicated by squares, while triangles indicate no significant (α = 0.05) differences between the cross-validation residuals of each scan distance and the minimum. LV = number of latent variables used by the model.