**Supplementary File S1**

For the prediction of milk composition, partial least squares regression (PLSR) using the R-package pls (Mevik and Wehrens, 2015) was applied. To derive prediction equations, an in-house dataset with 1,740 MIR spectra with known fatty acid concentrations was used, while the 15,596,136 MIR records used contained no labeled data.

To ensure optimal model fit, individual spectra were pre-processed by additional scaling, smoothing, and filtering with the ideal settings determined by cross-validation on the panel of 1.740 labeled spectra. For this, 20% of the internal dataset was randomly masked, and settings were compared based on the correlation between the masked response variable and fitted value using 100 replicates for each setting.

In the finally used model, individual wave lengths were mean-centered and divided by the empirical standard deviation in the individual wave length to increase the impact of wave length with lower variance, but never by a factor lower than 0.01 to avoid excessively increasing the weight of low-variance regions. Subsequently, individual spectra were smoothed, using a Nadaraya-Watson estimation (Nadaraya, 1964) with a bandwidth of 10 and a Gaussian kernel. Lastly, wave lengths 170 to 208 and 548 to 1060 were removed from the dataset, resulting in a final panel of 508 wave lengths per sample. Regions from 548 to 1060 are commonly removed from fatty acid prediction due to this association with water content (Grelet et al., 2015). In contrast, wave lengths 170 to 208 were removed due to insufficient training data as the 1.740 labeled spectra did not cover the full range of observed intensity values for the given wave length in the target dataset. Since no appropriate training data was available to ensure a good fit for these wave lengths on the full data panel, any real signal from these regions may have been lost for the analysis. We would still expect the impact of this region to be low as cross-validation accuracies were not substantially impacted by the removable of wave length in this region.

For the final model, predictions using the first 70 principal components were used to derive estimates. An overview of the average observed prediction accuracy is given in Supplementary Figure S20. In short, the removal of additional wave length substantially reduces the number of PCs required to obtain reliable predictions while smoothing and scaling improved prediction accuracies with the same number of PCs included while reducing overfitting. Average prediction accuracies increased by 9.2% by the use of all three preprocessing techniques compared to the plain use of a PLSR model across all fatty acids, with smoothing and filtering alone improving predictions by 6.8% and 5.9%, respectively. Predictions for fat and protein percentage obtained perfect predictions in all models, while short-chain fatty acids were predicted with an accuracy of ~85% and long-chain fatty acids ~70% (Supplementary Figure S21). Note again that these prediction accuracies were derived from the smaller panel of 1,740 MIR spectra, whereas the full dataset includes additional sources of noise, such as differences in machines, farms, and years, for which no training data existed to correct for these effects. Nonetheless, at minimum predictions for F% and P% should be highly reliable, showing correlations of 0.96 to test-day records used by CRV.

References

Grelet, C., J. F. Pierna, P. Dardenne, V. Baeten, and F. Dehareng. 2015. Standardization of milk mid-infrared spectra from a European dairy network. Journal of Dairy Science 98(4):2150–2160.

Mevik, B.-H., and R. Wehrens. 2015. Introduction to the pls Package. Help section of the “Pls” package of R studio software:1–23.

Nadaraya, E. A. 1964. On estimating regression. Theory of Probability & Its Applications 9(1):141–142.