Leveraging the DHI Databases to Estimate Missing Milk FT-MIR Spectra: A Strategy Toward Improving the Reliability of Breeding Values and Predictive Models

H. Soyeurt[1], X.-L. Wu[4,5], C. Grelet[2], M.L. van Pelt[3], N. Gengler[1], F. Dehareng[2], C. Bertozzi[6], and J. Burchard[4]

[1] Research and Teaching Centre (TERRA), Gembloux Agro-Bio Tech, University of Liège, 5030 Gembloux, Belgium; [2] Walloon Agricultural Research Center, 5030 Gembloux, Belgium; [3] Cooperation CRV, Animal Evaluation Unit, PO Box 454, 6800 AL Arnhem, the Netherlands;
[4] Council of Dairy Cattle Breeding, Bowie, MD 20716, USA;
[5] Department of Animal and Dairy Sciences, University of Wisconsin, Madison, WI 53706, USA;
[6] Walloon Breeders Association, 5590 Ciney, Belgium.

There use of milk Fourier Transform mid-infrared (**FT-MIR**) spectroscopy is on the rise in the dairy sector, thanks to the availability of new predicted phenotypes. For genetic evaluations, having a deep spectral dairy herd improvement (**DHI**) database is crucial for improving the reliabilities of breeding values. Unfortunately, the raw spectral data used to generate these FT-MIR phenotypes are not routinely stored. Moreover, many reference measurements of those traits of interest are available from past research activities but lack spectral records, making it impossible to use them to improve the FT-MIR models. Consequently, there is a growing interest in estimating those missing spectra. This study aimed to leverage the large spectral DHI database to estimate missing spectra by selecting probable spectra using, as the match criteria, four common dairy traits recorded for a long time by DHI organizations. Four matching combinations were tested to estimate missing spectra using the spectral DHI database: **Combi1** required equal fat and protein contents between the sample for which a spectrum was to be estimated and the reference samples, **Combi 2** requested additional equal urea content, Combi3 required equality for fat, protein, and lactose contents, and **Combi4** required equality for all criteria. When more than one spectrum was matched, their average was the estimated spectrum. Based on a search from 2,000,000 spectral DHI records, 1,700 missing spectra were estimated. For assessing the prediction quality, 11 phenotypes were predicted using FT-MIR models, related to the milk fat and mineral composition, lactoferrin content, the amount of eructed methane, the body weight, and the dry matter intake. Combi2 and Combi4 were too strict to estimate a spectrum for most samples. Combi3 had a poorer prediction performance than Combi1 due to fewer matched samples available to compute the missing spectrum. A new combination that allowed a range for matching lactose content ($\pm$0.1 g/dL of milk) increased the number of matched samples, leading to a better prediction. This performance was further improved by performing queries on the entire Walloon DHI spectral database (6,625,570 spectra). The prediction performance varied among the studied phenotypes. Without considering the traits used for the matching, the best predictions were obtained for the content of saturated fatty acids (Mean absolute error (MAE) = 0.15 g/dL of milk) and body weight (MAE = 12.80 kg). However, the predictions for the unsaturated fatty acids were less satisfactory (MAE = 0.13 and 0.018 g/dL of milk for monounsaturated and polyunsaturated fatty acids) mainly because of the poorer predictions of spectral regions related to the unsaturation of carbon chain. The same reason also accounted for the poorer methane predictions (MAE = 47.02 g/day). In conclusion, increasing the number of relevant matching criteria helps improve the quality of FT-MIR-

predicted phenotypes and the number of spectra used during the search, highlighting the interest in creating large-scale international spectral databases. These estimated missing spectra can be used to predict FT-MIR phenotypes in order to improve the reliability of breeding values estimated for FT-MIR-based phenotypes and the robustness of their related predictive models.