The unified framework for processing and storing large-scale sniffers-based emission data

Milkevych Viktor*[1], Ashayeri Leila[2], Annadatha Sowmya[1], Stephansen Rasmus Bak[1], Villumsen Trine Michelle[1], Bjerring Martin[3], Javed Muhammad[1]

[1] Center for Quantitative Genetics and Genomics, Aarhus University, [2] Center for Quantitative Genetics and Genomics, Aarhus University, Denmark, [3] Department of Animal and Veterinary Sciences, Aarhus University

Accurate quantification of CH4 and CO_2 emissions is essential for evaluating environmental impacts and developing mitigation strategies. In this context, sniffers play substantial role. The quality of such records is a well-known challenge and is critical to address. This study presents the framework dedicated for storing and processing a sniffers-based emission data. It integrates the GEDA pipeline for processing raw measurements within HPC environment and database storage.

The pipeline is based on a stochastic model that represents noisy gas emission signals as a function of automatic milking system's (AMS) data. This formulation enables the workflow that integrates data quality control, errors correction, synchronization of AMS and sniffers data following reliability assessment. The background concentrations are estimated through a weighted regression model following data corrections. Emission phenotypes are estimated as mean concentrations or emission intensities at visit or daily scales. The phenotypes quality is evaluated using linear mixed models to derive local repeatability indicator. GEDA enables workflow that integrates data quality control, errors correction, synchronization of AMS and sniffers data following reliability assessment. The background concentrations are estimated through a weighted regression model following data corrections. Emission phenotypes are estimated as mean concentrations or emission intensities at visit or daily scales. The phenotypes quality is evaluated using linear mixed models to derive local repeatability indicator. The storage system implements a four-tier architecture that covers data ingestion, storage, application logic, and a user interface. There are two distinct pipelines. The staging pipeline automatically ingests raw data into a PostgreSQL/TimescaleDB staging schema with full provenance metadata and idempotent loading. The curated pipeline then exports validated staging data for GEDA processing. A FastAPI backend serves curated data through RESTful APIs with role-based access control, while a frontend provides interactive dashboards and a data explorer with filtering, visualizations, and data export capabilities.

A study, using the four-year data from 38 commercial farms (the number of unique cows is 15335, the average number of observations per cow is 148.45, and the total number of observations is 2276528), assessed the pipeline's efficiency and quality. The estimated heritability values for daily averaged phenotypes are in the range of 0.13-0.16, and the repeatability values are in the range of 0.32-0.41. GEDA provides a novel data pipeline for large-scale processing of sniffer-based emission data. By systematically addressing data heterogeneity, noise, and misalignment, the pipeline produces high-quality emission phenotypes. Its HPC and storage integration represents the unified framework that supports more accurate genetic evaluations and strengthens research targeting mitigation of greenhouse gas emissions in dairy cattle. Overall, GEDA provides a novel, theoretically grounded automated pipeline for large-scale processing of sniffer-based gas emission data. By systematically addressing data heterogeneity, noise, and misalignment, the pipeline produces aligned datasets and high-quality emission phenotypes. Its HPC and storage integration represents the unified framework that supports more accurate genetic evaluations and strengthens research targeting mitigation of greenhouse gas emissions in dairy cattle.