Multi-frequency data handling

Author: Guilherme Ramos, Tiago Mateus
Affiliation: DataHow AG, R&D Team
Scope: Internship
Skills: Coding in Python, Machine Learning, Time Series Handling
Date: September 26, 2025
Contact: h.narayanan@datahow.ch

Executive Summary

The primary goal of this project is to develop a flexible framework for combining high-frequency and low-frequency measurements to improve the fidelity and robustness of our bioprocess models. The proposed approach will be developed and evaluated for a variety of datasets (in-silico and potentially real processes) and benchmarked against our current approach.

Background

Brief introduction / Context

Bioprocesses rely on the use of living organisms to produce various products, presenting a dynamic system with complex interactions. Modeling is crucial for design, optimization, monitoring, and control of these processes. However, one of the major challenges in bioprocess modeling lies in the heterogeneous nature of available data, particularly the integration of high-frequency sensor data (e.g., temperature, pH, dissolved oxygen) with low-frequency analytical measurements (e.g., biomass concentration, metabolite levels). These data sources differ not only in their sampling rates but also in their levels of noise, reliability, and informational content.

Current approaches / State of the art

Traditional approaches often rely on preprocessing techniques such as resampling, interpolation, or smoothing to align data temporally. Although simple to implement, these methods can introduce bias or obscure important dynamics, especially when applied to sparse or nonlinear systems. More advanced methods include: multi-rate filtering and estimation, Gaussian processes, neural networks for irregular time series (latent ODEs, ODE-RNNs, attention-based models) and hybrid approaches (physics-informed neural networks, hybrid state estimators).

Challenges

Despite progress in handling irregular and multi-frequency data in other domains, several challenges remain when applying these techniques to bioprocess modeling: temporal misalignment, information imbalance, noise/uncertainty, modeling temporal dependencies and integration with process models.

Project rationale / Approach

To mitigate this, we want to assess alternatives to the traditional approaches, which could nicely integrate into our current modeling pipeline and compare them against the benchmark.

Objectives

1. **Implementation of a data frequency mismatch handler:** Explore different architectures to deal with data frequency mismatch and connect them to the modeling pipeline.
2. **Predictive performance benchmarking:** Study the performance of the model when using the new data preparation approach and compare it against the benchmark models. Computational time and memory-efficiency of the proposed approaches are important parameters to monitor.
3. **Capability assessment:** Inspect the effect of different noise levels, data frequency and missing data on the different data sources.
4. **Stress testing and limitations:** Identify potential shortcomings/pitfalls of the proposed approach and outline future work direction.

Methods and Work Plan

Data and Resources

Timeline

Timeline subject to the scope. Preference: Longer the better.

Phase	**Tasks**	**Target Dates**
Exploration	Literature review, baseline reproduction.	Wk 1–3
Prototype	Develop a new data handling pipeline; sanity checks.	Wk 4–6
Benchmark	Compare model performance when feeding with the new pipeline against the benchmark on multiple datasets.	Wk 7–9
Stress Tests	Noise/frequency/missing data effect studies; limitations.	Wk 10–11
Wrap-up	Documentation; presentation; next steps.	Wk 12

Expected Outcome

At the end of this project, we aim to have a clear assessment of which multi-frequency handling method fits better our current modeling pipeline for bioprocess applications.