Evaluating the interpretability of the Hybrid-RNN Bioprocess Models

Author:	Harini Narayanan, Lucas Cosier
Affiliation:	DataHow AG, R&D Team
Scope:	Internship
Skills:	Curiosity and willingness to tackle challenging research problems Strong background and experience in Machine Learning Good programming skills in Python Familiarity with autodifferentiation
Date:	March 10, 2026
Contact:	h.narayanan@datahow.ch

Executive Summary

This project proposes developing a **interpretability** framework for hybrid bioprocess models that combine neural networks with mechanistic mass balances. While these models achieve strong predictive performance through learned residual dynamics, they remain fundamentally black boxes. Extracting and analyzing the Jacobian, can reveal metabolic control patterns and enable early detection of process deviations. Such framework will provide both interpretability tools for understanding model decisions and predictive capabilities for proactive process control, transforming black-box models into actionable grey-box systems that maintain performance while offering mechanistic insights. There are also additional interpretability metric that can potentially reveal complementary patterns in the hybrid modeling framework. The over all aim is to perform a thorough investigation of the different methods of interpretability and design a guided to approach to choosing the appropriate interpretbaility metric.

Background

Brief introduction / Context

Bioprocesses are dynamic systems with complex interactions, making them difficult to capture with either purely mechanistic or purely data-driven models. Hybrid models address this limitation by combining machine learning (ML) with a mechanistic backbone, thereby improving predictive performance. However, interpretability remains a key concern, as these models often function as black boxes. To support deeper process insight and effective control, it is essential not only to accurately forecast system trajectories but also to uncover the underlying factors that drive these outcomes.

Current approaches / State of the art

Among the models used for hybrid modeling, recurrent neural networks (RNNs) are particularly well suited. At the same time, their complexity has motivated the development of interpretability

approaches in deep learning, where techniques such as feature attribution or attention mechanisms aim to indicate which inputs most strongly influence predictions.

Challenges

Despite strong predictive performance, hybrid RNNs remain difficult to interpret as their predictions arise from complex, recurrent transformations across many timesteps. In addition, bioprocess data often contain discontinuities introduced by bolus feeding, for instance, which create non-smooth dynamics and complicate sensitivity analysis. The large number of features further results in large sensitivity matrices at every timestep, which are difficult to analyze without dimensionality reduction.

Objectives

1. **Set up the hybrid model:** Set up the hybrid-RNN model for in-silico dataset
2. **Literature Curation of interpretability metrics:** Survey and summarize the different plausible intepretability metric with theoretical pros-cons for our application. Select a subset of most relevant scoring with the capability to offer orthogonal/complementary perspective into the model. Some crucial categories to consider: (i) Interpretability in time, (ii) Interpretability into hidden state (and memory) and (iii) general feature importance.
3. **Jacobian extraction and analysis:** Implement routines to compute $\partial X_{t+1}/\partial X_t$ at every timestep of the RNN. Characterize dominant eigenvalues/vectors to identify metabolic control shifts.Handling discontinuities: Devise strategies to separate continuous Jacobians from discrete bolus-induced jumps.
4. **Evaluate other promising interpretability metric:** Implement additional metrics such as probing hidden states, memory retention, and general feature importance
5. **Compare different metric and draw final conclusions:** Compare and contrast different metrics to determine if they provide complementary interpretations

Methods and Work Plan

Data and Resources

- Datasets: In-silico datasets with bolus feeding
- Frameworks: Python, PyTorch

Timeline

Timeline subject to the scope. Preference: The longer, the better.

Replace example table with your plan (weeks or months).

Timeline

Phase	Tasks	Target Dates
Exploration	Literature review on different metrics; baseline hybrid RNN setup.	Wk 1–3
Implementation	Different metrics.	Wk 4–6
Analysis	Evaluate the different types of informationon provided by the metrics.	Wk 7–9
Applications	Interpretability demos; predictive depletion forecasting.	Wk 10–11
Wrap-up	Documentation, evaluation vs alternatives, final report.	Wk 12

Expected Outcome

At the end of this project, we will have developed a interpretability toolkit for hybrid RNN bioprocess models.