Implementing Automatic Relevance Determination (ARD) for RNNs for Bioprocess modelling

Author:	Harini Narayanan, Jean Demaël
Affiliation:	DataHow AG, R&D Team
Scope:	Internship
Skills:	Strong ML background, Python/PyTorch, Critical thinking (Bioprocess knowledge not crucial)
Date:	March 10, 2026
Contact:	h.narayanan@datahow.ch

Executive Summary

The primary objective of this project is to integrate Automatic Relevance Determination (ARD) into Recurrent Neural Network (RNN) architectures to enhance modeling efficiency in bioprocess applications. By leveraging ARD, we aim to systematically prune irrelevant input features and redundant hidden units, transitioning from "black-box" models to more interpretable, sparse architectures. This approach will be developed and evaluated for a variety of datasets (in-silico and potentially real processes) and benchmarked against our current implementation of RNN.

Background

Brief introduction / Context

Bioprocesses relies 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. The complexity of these systems prevents the development of pure knowledge–based models while the expense of collecting data limits the application of purely data-driven machine learning (ML) model. Hybrid models are a paradigm of models that augment ML models with a knowledge backbone, available in the form of system of differential equations.

$$\frac{dX_i}{dt} = f(X, Z) + \mathrm{In}_i - \mathrm{Out}_i$$

$$f(X, Z) = NN(X, Z)$$

Current approaches / State of the art

The straightforward augmentation into the ML framework can be realized by discretizing the differential equations and learning a mapping between the current time to a future time point.

$$X_{i,t+1} = X_{i,t} + \Delta t \times (f(X_t, Z) + \mathrm{In}_i - \mathrm{Out}_i)$$

We have an existing implementation with a recurrent neural network, specifically, an LSTM to learn the function f(X,Z).

Challenges

We have observed in some of our usecases that the presence of irrelevant features hampers the performance of the model. However, manual selection of the data is laborious and often subjective. Thus, we would like to develop a framework to automatically select the relevant input features for the hybrid-RNN.

Project rationale / Approach

Particularly, we would like to explore existing ARD mechanisms for RNNs and test them within our framework. By placing relevance hyperparameters on input features, the architecture will "learn" to prune non-contributing connections during training. The goal is to improve accuracy while also improving efficiency. This method will be compared to other techniques such as variational dropouts, attentional mechanisms, or group Lasso L1 regularization.

Objectives

1. **Implement automated feature selection methods:** Get familiarized with the current framework. Then evaluate the impact of different feature selection techniques: PCA, Lasso regularization, attention etc., on the predictive performance and efficiency. Test on in-silico datasets and confirm trends on benchmark datasets from real processes.
2. **Implement alternative ARD methods:** Implement the alternative ARD methods, adapt and integrate them within the existing framework.
3. **Predictive Performance assessment:** Study the performance of the model for base predictive capabilities and benchmark against current RNN for different datasets. Computational time and memory-efficiency are important parameter to monitor.
4. **Capability assessment:** Evaluate the pro-cons of the different frameworks and determine a pragmatic choice to carry forward.

Methods and Work Plan

Data and Resources

- Datasets: Existing in-silico bioprocess datasets with varying noise and sampling; selected de-identified real processes for external validity (subject to availability).
- Software: Python, PyTorch

Timeline

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

Phase	Tasks	Target Dates
Exploration	Literature review, baseline reproduction.	Wk 1–3
Assess	Impact of different feature selection methods.	Wk 4–6
Benchmark	Build the alternative ARD RNN; Establish benchmarks.	Wk 7–10
Stress Tests	Pros-cons evaluation of frameworks.	Wk 11
Wrap-up	Documentation; presentation; next steps.	Wk 12

Expected Outcome

At the end of this project, we hope to have a clear assessment of the abilities of different reasonably characterized implementations of feature selection for application to bioprocesses modeling using Hybrid-RNN.

References

[1] Maxim Kodryan et al. "Efficient Language Modeling with Automatic Relevance Determination in Recurrent Neural Networks". In: *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*. Ed. by Isabelle Augenstein et al. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 40–48. DOI: 10.18653/v1/W19-4306. URL: https://aclanthology.org/W19-4306/.

[2] Rendani Mbuvha, Illyes Boulkaibet, and Tshilidzi Marwala. *An Automatic Relevance Determination Prior Bayesian Neural Network for Controlled Variable Selection*. 2020. arXiv: 2001.01765 [stat.ML]. URL: https://arxiv.org/abs/2001.01765.

[3] Yao Qin et al. *A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction*. 2017. URL: https://arxiv.org/pdf/1704.02971.

[4] Surojit Saha, Sarang Joshi, and Ross Whitaker. *ARD-VAE: A Statistical Formulation to Find the Relevant Latent Dimensions of Variational Autoencoders*. 2025. arXiv: 2501.10901 [cs.LG]. URL: https://arxiv.org/abs/2501.10901.

[5] Hongwei Zhang et al. *Efficient Network Automatic Relevance Determination*. 2025. arXiv: 2506.12352 [cs.AI]. URL: https://arxiv.org/abs/2506.12352.