Ensembling Techniques for Robust Hybrid Modeling

Author:	Lucas Cosier
Affiliation:	DataHow AG, R&D Team
Scope:	**Internship**
Skills:	Curiosity to tackle challenging research problems
	Strong background and experience in Machine Learning
	Good programming skills in Python
Date:	September 26, 2025
Contact:	h.narayanan@datahow.ch

Executive Summary

Hybrid bioprocess models combining kinetics based-mechanistic model with neural networks face significant challenges in predictive uncertainty and generalization across diverse experimental conditions. This project aims to investigate ensemble learning strategies to move beyond simple model averaging, inspired by recent work [1] that proposes stacking in time series forecasting. Building on our existing infrastructure of recurrent hybrid continuous models, we will develop hierarchical ensemble architectures that learn optimal model combinations through learned weighting schemes, selective model inclusion, and intelligent aggregation methods that learn how to best combine diverse model outputs. The research will explore dynamic estimator selection strategies, experiment with out-of-bag scoring through data subsampling, and implement advanced model scoring strategies for L2-level stacking. The goal of the project is to improve the robustness and performance in terms of predictive accuracy and confidence interval calibration of our recurrent hybrid models.

Background

Brief introduction / Context

Hybrid bioprocess models, which integrate mechanistic differential equations with neural network components, have emerged as a powerful paradigm for capturing both known biological mechanisms and unmodeled system dynamics. In our framework, these autoregressive hybrid models are unrolled through time to generate multiple Monte Carlo trajectory realizations, providing a natural mechanism for uncertainty quantification through the construction of confidence intervals from the ensemble of predictions. The critical question of how these stochastic trajectories should be optimally combined directly impacts both predictive accuracy and the calibration of uncertainty estimates.

Ensemble learning offers a principled approach to address these limitations by combining predictions from multiple diverse models. Our experiments have shown that ensembles demonstrate improved robustness and better-calibrated confidence intervals compared to single models, as the diversity in model predictions naturally captures epistemic uncertainty arising from limited training data and model specification choices.

Current approaches / State of the art

Current ensemble strategies in hybrid bioprocess modeling primarily rely on simple averaging techniques, where K-fold cross-validation could be used to generate multiple models whose predictions are combined through arithmetic or weighted means:

$$\hat{y}_{\text{ensemble}} = \frac{1}{K} \sum_{k=1}^{K} \hat{y}_k$$

While effective, these approaches fail to exploit the rich information contained in individual model behaviors. Recent advances in time series forecasting [1] demonstrate that hierarchical stacking—training a meta-model to optimally combine base model predictions—can achieve significant performance improvements over simple aggregation methods by learning combination functions from out-of-fold predictions

Challenges

Despite proven benefits, current ensemble implementations face several limitations: (i) uniform weighting schemes ignore model-specific strengths across different operating regimes, (ii) fixed ensemble sizes may be suboptimal for datasets with varying complexity, and (iii) simple averaging fails to leverage correlations between model errors for improved uncertainty quantification. Furthermore, the limited availability of bioprocess data constrains the application of sophisticated multi-layer stacking architectures that require extensive validation sets. Finally, the incompatibility between bootstrap sampling and stacking methods requires careful selection of which ensemble strategy to apply.

Project rationale / Approach

This project will build upon and extend our existing ensemble architecture to investigate stacking methodologies within the constraints of bioprocess modeling. We will investigate two complementary ensemble generation strategies: (i) K-fold cross-validation for stacking approaches requiring complete out-of-fold predictions, and (ii) bootstrap sampling for simple averaging ensembles with flexible model counts (B=5-20) to determine optimal ensemble sizes. The proposed framework will leverage our established cross-validation infrastructure while introducing bootstrap sampling methods to generate diverse model portfolios, enabling comprehensive evaluation of how trajectory aggregation strategies affect confidence interval calibration and overall model robustness.

Objectives

1. **Implementation of ensemble frameworks:** For K-fold CV models, develop L2-level stacking architectures including greedy selection and linear meta-models. For bootstrap ensembles, implement weighted averaging schemes to handle larger model counts without stacking complexity.

2. **Performance benchmarking across ensemble strategies:** Systematically evaluate stacking methodologies against baseline averaging across multiple bioprocess datasets. Quantify improvements in predictive accuracy and computational overhead.

3. **Uncertainty quantification through trajectory aggregation:** Assess impact of stacking approaches on confidence interval calibration when combining Monte Carlo trajectories. Evaluate coverage probability and epistemic uncertainty decomposition.

4. **Adaptive ensemble selection:** Develop algorithms for dynamic determination of optimal estimator counts based on dataset characteristics. Establish guidelines for method selection under data and computational constraints.

Methods and Work Plan

Data and Resources

Datasets: Existing insilico datasets with varying noise levels, as well as customer (confidential) data.
Software: Python ecosystem leveraging PyTorch for neural network components, existing hybrid model infrastructure.

Timeline

Project scope: 12 weeks full-time (extensible to semester project).

Phase	Tasks	Target Dates
Foundation	Literature review; reproduce baseline ensembles; implement performance metrics.	Wk 1–2
Core Implementation	Build stacking framework; implement greedy selection, weighted aggregation, linear meta-models.	Wk 3–5
Validation Infrastructure	Develop out-of-fold prediction pipeline for K-fold CV; out-of-bag validation for bootstrap ensembles.	Wk 6–7
Benchmarking	Systematic evaluation across datasets; ablation studies; computational profiling.	Wk 8–10
Advanced Features	Adaptive ensemble selection; trajectory-specific weighting (different weights per Monte Carlo trajectory); uncertainty decomposition.	Wk 11
Documentation	Code packaging; results synthesis; guidelines development.	Wk 12

Expected Outcome

Delivery of a validated stacking framework demonstrating measurable improvements in predictive accuracy and uncertainty quantification over baseline averaging. Key deliverables include: (i) open-source implementation integrated with existing infrastructure, (ii) comprehensive benchmarking results across bioprocess datasets, and (iii) practical guidelines for ensemble strategy selection based on data characteristics.

References

References

[1] N. Bosch, O. Shchur, N. Erickson, M. Bohlke-Schneider, and A. C. Turkmen, "Multi-layer stack ensembles for time series forecasting," in *AutoML 2025 Methods Track*, 2025. [Online]. Available: https://openreview.net/forum?id=ve5Q1q1W5n.