Multi-Scale Modelling and Variable Transformations

Author:	Jakub Polak, Guilherme Ramos, Alice Rosa
Affiliation:	DataHow AG, R&D Team
Scope:	Internship
Skills:	Coding in Python, Machine Learning, Time Series Handling, Biotech knowledge beneficial
Date:	March 11, 2026
Contact:	h.narayanan@datahow.ch

Executive Summary

The goal of this project is to develop a principled, scale-aware feature engineering framework to improve transfer learning in bioprocess scale-up models. Current approaches either ingest variables as is, treat scale as a categorical variable, or apply global normalization strategies to reconcile data from different bioreactor scales, but do not leverage the underlying process physics that govern how variables change across scales. This project will systematically identify which process variables are scale-dependent, propose scale-invariant transformations grounded in engineering principles (e.g., dimensionless numbers, specific rates, volumetric transfer coefficients), and evaluate whether these representations improve the predictive accuracy and extrapolation capability of hybrid models. The outcome will be a practical feature engineering guide and an improved transfer learning pipeline validated on internal multi-scale datasets. In addition, a general assessment of the knowledge that is possible to transfer from one scale to another should be quantified.

Background

Brief introduction

Bioprocess scale-up remains one of the most resource-intensive stages of biopharmaceutical development. As vessel volume increases from microbioreactors (e.g., ambr15) to lab and pilot scale, biological systems respond nonlinearly to changes in mixing, oxygen transfer, shear stress, and feeding dynamics. Data-driven and hybrid models have shown great promise in supporting scale-up by learning cross-scale relationships from historical data. DataHow has demonstrated that including even a single large-scale run in model training can improve prediction accuracy by up to 50% compared to models trained exclusively on small-scale data.

Current approaches

A fundamental and largely unaddressed question in multi-scale modeling remains: which input features are truly scale-invariant, which are scale-dependent, and how should the latter be handled? The standard approach involves including all available features regardless of physical comparability across scales, applying global normalization uniformly, or appending a categorical scale indicator to the feature vector. These strategies are pragmatic but suboptimal, they treat scale as a nuisance variable rather than a structured source of information, and may fail

to generalize to unseen scales or configurations (Mercier et al., 2017; Möller et al., 2019). The bioprocess engineering literature provides rich guidance on how physical phenomena change with scale. The volumetric oxygen transfer coefficient (kLa), power input per unit volume (P/V), and mixing time all depend strongly on vessel geometry and agitation, meaning the same DO setpoint or RPM value represents fundamentally different cellular environments at ambr15 versus 200 L scale. Feed volumes added as bolus additions are not directly comparable across scales without normalization to viable cell density. Dimensionless numbers such as the Reynolds number encode the relationship between physical transport and biological reaction rates in a scale-independent way (Nienow, 2006; Junker, 2004). These principles are well established in the engineering literature but have not yet been systematically integrated into data-driven transfer learning pipelines.

Challenges

The main challenges of this project are the rigorous classification of process variables by their scale-dependence, the computation or estimation of scale-relevant parameters (kLa, P/V, mixing time) from available sensor and reactor geometry data which may be incomplete, and the design of a benchmarking framework that isolates the effect of feature engineering from other modeling choices. Ensuring the resulting framework is interpretable for process scientists and generalizable across process modalities are additional considerations.

Project rationale

This project pursues five interconnected objectives to deliver a rigorous and practically applicable scale-aware modeling framework:

Objectives

1. **Variable Classification:** Systematically classify commonly used bioprocess input variables (feeds, pH, agitation, gas flows, temperature, metabolites, online spectra) as scale-independent, scale-dependent, or conditionally scale-dependent, supported by engineering reasoning and literature.

2. **Scale-Invariant Feature Derivation:** For each scale-dependent variable, propose and implement scale-aware transformations. Candidates include specific feed rates normalized to viable cell density, P/V, kLa, dimensionless mixing times, and cell-specific metabolic rates.

3. **Comparative Modeling Study:** Train and evaluate hybrid models using (a) the current baseline with a categorical scale variable, (b) scale-specific normalization, and (c) the proposed scale-aware feature set. Compare performance on held-out large-scale runs using RMSE of final titer and key CQAs.

4. **Transfer Learning Assessment:** Evaluate how feature representation affects the minimum number of large-scale runs required to achieve a target prediction accuracy, directly quantifying the practical benefit of scale-aware features for reducing large-scale experimentation.

5. **Documentation:** Deliver a practical feature engineering guide and a reusable preprocessing pipeline with clear recommendations for integration into DataHowLab's transfer learning workflow.

Methods and Work Plan

Timeline

Timeline subject to the scope, the duration would be 3-6 months.

Phase	Tasks	Target Dates
Exploration	Literature review on scale-dependent variables and scale-up criteria (kLa, P/V, mixing time, dimensionless numbers); review of internal datasets.	Wk 1-3
Variable Classification	Develop classification scheme for input variables; identify data gaps; consult with process engineers.	Wk 4-5
Feature Engineering	Implement scale-invariant transformations in Python; validate against literature benchmarks.	Wk 6-7
Modeling	Train hybrid models with baseline and proposed feature sets.	Wk 8-10
Benchmarking	Evaluate prediction accuracy and extrapolation performance on internal multi-scale datasets.	Wk 10-11
Wrap-up	Documentation, final benchmarking report, recommendations for DataHowLab integration, demo cases.	Wk 12

Expected Outcome

By the end of the project, DataHow will have a principled classification of bioprocess input variables by scale-dependence, a validated set of scale-aware feature transformations ready for hybrid model training, and quantitative evidence on whether this approach reduces the large-scale data requirement compared to current methods. A reusable Python preprocessing module compatible with the existing DataHowLab infrastructure and a practical guide for process scientists will also be delivered.

References

- Sokolov, M., Morbidelli, M., Butte, A., Souquet, J., & Broly, H. (2018). Sequential Multivariate Cell Culture Modeling at Multiple Scales. Biotechnology Journal, 13, 1700461. https://doi.org/10.1002/biot.201700461
- DataHow AG & Wheeler Bio. (2024). Accelerating Scale-Up with Transfer Learning and DataHowLab [Case Study]. www.datahow.ch
- Nienow, A. W. (2006). Reactor engineering in large scale animal cell culture. Cytotechnology, 50, 9–33. https://doi.org/10.1007/s10616-006-9005-8
- Junker, B. H. (2004). Scale-up methodologies for Escherichia coli and yeast fermentation processes. Journal of Bioscience and Bioengineering, 97(6), 347–364. https://doi.org/10.1016/S1389-1723(04)70218-2

- Mercier, S. M. et al. (2014). Multivariate PAT solutions for biopharmaceutical cultivation. Biotechnology Advances, 32(6), 329 - 336. https://doi.org/10.1016/j.tibtech.2014.03.008
- Möller, J. et al. (2019). Model uncertainty-based evaluation of process strategies during scale-up. Computers & Chemical Engineering, 134, 106693. https://doi.org/10.1016/j.compchemeng.2019.106693