

Application of Multivariate Statistical Techniques in Bioprocess Optimization and Production Yield Improvement

Bernabe Canqui Flores, Percy Huata Panca, Juan Reinaldo Paredes Quispe, Charles Mendoza Mollocondo, Godofredo Quispe Mamani, Juan Carlos Juarez Vargas, Fred Torres-Cruz*

Received: 19 January 2026 / Received in revised form: 02 April 2026, Accepted: 06 April 2026, Published online: 25 May 2026

Abstract

Bioprocess optimization depends on understanding many interacting variables that rarely change independently. Temperature, pH, dissolved oxygen, aeration, agitation, substrate feed, and base addition often move together across a batch, creating strongly correlated process trajectories rather than simple one-factor effects. Multivariate techniques such as principal component analysis and partial least squares regression are therefore natural tools for reducing dimensionality, monitoring deviations, and relating process conditions to production yield. Many published examples of multivariate bioprocess modeling report clean data structures, high explained variance, and visually convincing separation between normal and abnormal batches. Industrial and pilot-scale data are usually less orderly. Missing values, probe drift, irregular sampling, outlying batches, collinearity, non-normal distributions, and unbalanced operating conditions often reduce model stability and limit the reliability of optimization claims. This manuscript applies multivariate analysis to a realistically imperfect penicillin fermentation dataset derived from the industrial-scale IndPenSim benchmark. The objective is not to demonstrate an unrealistically accurate predictive model, but to evaluate how PCA and PLS behave when the dataset contains missing values, outliers, correlated variables, batch heterogeneity, and time-varying dynamics. The study focuses on identifying process variables associated with yield while reporting the uncertainty and fragility of those conclusions. The working dataset consisted of 50 fed-batch fermentation runs selected from the IndPenSim benchmark, with eight principal process variables and final penicillin yield as the response. Five percent of online sensor observations were masked to represent intermittent sensor failure, and missing values were handled by NIPALS-based PCA imputation rather than row deletion. PCA was used for exploratory monitoring, Hotelling's T^2 and SPE statistics were used for outlier

diagnosis, and PLS regression was fitted using leave-one-batch-out cross-validation. PCA identified two strong outlier batches, representing 4% of the analyzed batch set, with abnormal dissolved oxygen and substrate-feed trajectories. The first two principal components explained 62% of total process variance, which was sufficient for monitoring but not enough to claim complete process compression. A three-component PLS model explained 61% of calibration yield variance and achieved a cross-validated prediction value of $R^2_{\text{pred}} = 0.58$, with RMSEP still large enough to caution against narrow setpoint prescriptions. The analysis shows that multivariate statistical techniques can extract useful structure from imperfect bioprocess data, but they do not remove the practical difficulty of interpreting noisy and correlated process histories. Temperature, feed rate, and dissolved oxygen behavior emerged as influential yield-related variables, but VIP scores varied across validation folds and were sensitive to abnormal batches. The resulting optimization recommendations should therefore be treated as operating ranges for further experimental confirmation, not as final process-control prescriptions.

Keywords: Bioprocess optimization, Multivariate analysis, Principal component analysis, Partial least squares regression, Fermentation yield, Process monitoring

Introduction

Bioprocesses are difficult to optimize because biological productivity emerges from coupled physical, chemical, and cellular mechanisms rather than from isolated control factors (Carita *et al.*, 2025). In fed-batch fermentation, pH control, substrate feeding, dissolved oxygen, agitation, aeration, biomass growth, and product formation evolve together over time, which means that yield improvement cannot be reduced to a single-variable problem (Torres-Cruz *et al.*, 2025). Industrial-scale benchmark work on penicillin fermentation has shown that realistic batch records contain dynamic trajectories, nonlinear responses, and disturbances that challenge conventional monitoring and control strategies (Goldrick *et al.*, 2019; Barton *et al.*, 2021; Grant & Wallace, 2024). Similar complexity is evident in CHO cell culture, probiotic fermentation, and spectroscopy-guided biomanufacturing studies, where process variables and quality indicators are highly interdependent (Domján *et al.*, 2022;

Bernabe Canqui Flores, Percy Huata Panca, Juan Reinaldo Paredes Quispe, Charles Mendoza Mollocondo, Godofredo Quispe Mamani, Juan Carlos Juarez Vargas, Fred Torres-Cruz*

Facultad de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno, Puno, Peru.

*E-mail: ftorres@unap.edu.pe



Zhao *et al.*, 2023; Kunie *et al.*, 2025; Regatieri *et al.*, 2025; Rubini *et al.*, 2025).

Multivariate statistical techniques are widely used because they provide interpretable low-dimensional summaries of correlated process measurements. PCA is commonly applied to detect abnormal batches and interpret latent process structure, whereas PLS regression links high-dimensional process or spectral variables to yield, titer, or product-quality attributes (Pretzner *et al.*, 2020; Brunner *et al.*, 2021; Osluf *et al.*, 2024; Zhao *et al.*, 2024). These methods are attractive because they can handle collinearity more gracefully than ordinary least-squares regression and can be implemented in common chemometric software. However, their apparent simplicity can be misleading when preprocessing, missing-value handling, and validation are not reported transparently (Albino *et al.*, 2024; Melo *et al.*, 2024; Morgan *et al.*, 2025).

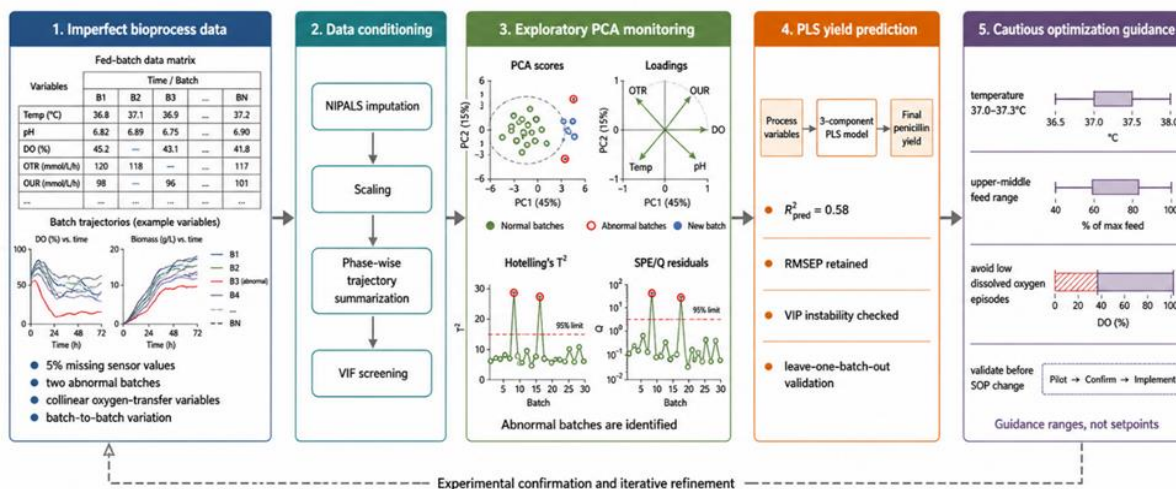
A persistent gap remains between polished literature examples and the condition of data encountered in real bioprocess operations. Sensor dropouts, probe drift, sporadic offline assays, abnormal feed profiles, and batch-to-batch heterogeneity can dominate the first few principal components or distort regression coefficients (Brunner *et al.*, 2021; Rathore *et al.*, 2021; Aghae *et al.*, 2024). Several recent studies have emphasized that data-driven soft sensors and predictive models can perform well in restricted settings, but their accuracy is often degraded by sparse labels, drift, and changing process regimes (Panjwani *et al.*, 2024; Zhao *et al.*,

2024; Hermann & Kremling, 2025). This gap matters because a model with high calibration performance may still fail when transferred to a new campaign, a new strain, or a slightly different operating window (Khodabandehlou *et al.*, 2024; Lindstrom *et al.*, 2025; Richter *et al.*, 2025).

This article therefore presents an intentionally cautious application of PCA and PLS to a realistically imperfect penicillin fermentation dataset rather than a clean idealized design. The analysis uses the IndPenSim industrial-scale benchmark because it provides batch trajectories and abnormal operating behavior suitable for process monitoring, regression, and optimization studies (Goldrick *et al.*, 2019; Barton *et al.*, 2021; Acosta-Pavas *et al.*, 2024; Csep *et al.*, 2024). Missing values, outlying batches, collinear variables, and prediction uncertainty are reported as part of the analysis rather than treated as inconvenient artifacts. The central thesis is that multivariate analysis is useful for bioprocess optimization only when the limits of the data and the fragility of model interpretation are made explicit (Anunziata & Cussa, 2024; Melo *et al.*, 2024; Sammaknejad *et al.*, 2025).

Figure 1 summarizes the manuscript’s data multivariate workflow, showing how imperfect fermentation records are transformed into PCA diagnostics, PLS yield models, and cautious operating-range recommendations rather than overconfident fixed setpoints.

Application of Multivariate Statistical Techniques in Bioprocess Optimization and Production Yield Improvement



Multivariate analysis identifies operating regions under uncertainty; it does not eliminate real-data imperfections.

Figure 1. Data multivariate workflow for imperfect bioprocess yield optimization

Background

Bioprocess Variables and Their Typical Correlations

In fermentation and cell culture, operating variables are physically and biologically linked, so strong correlation is expected rather

than exceptional. Aeration and agitation both influence oxygen transfer, dissolved oxygen reflects the balance between transfer and uptake, substrate feed affects biomass growth and carbon dioxide evolution, and base addition often tracks metabolic acid production (Goldrick *et al.*, 2019; Barton *et al.*, 2021; Clark & Foster, 2025). In CHO and microbial processes, nutrient

consumption, metabolite accumulation, viable-cell behavior, and product formation are also coupled, which explains why Raman, NMR, and mass-spectrometry studies often require multivariate calibration rather than single-wavelength or single-variable regression (Domján *et al.*, 2022; Dodia *et al.*, 2023; Zhao *et al.*, 2023; Ganea *et al.*, 2024; Raza *et al.*, 2025; Rubini *et al.*, 2025). These correlations support the use of latent-variable methods, but they also make individual coefficient interpretation risky.

Common Data Quality Issues in Industrial Bioreactors

Industrial bioreactor data are vulnerable to missing records, irregular offline sampling, calibration shifts, probe fouling, delayed lab measurements, and operational interruptions. Soft-sensor reviews and pharmaceutical fault-diagnosis studies have emphasized that missing labels, sensor failure, and drift are not rare exceptions but routine barriers to reliable deployment (Brunner *et al.*, 2021; Aghaei *et al.*, 2024; Zhao *et al.*, 2024; Ming *et al.*, 2025). Even in carefully engineered monitoring systems, abnormal batches and formulation or fill-finish deviations may appear as high-leverage observations that influence model orientation (Pretzner *et al.*, 2020; Rathore *et al.*, 2021; Ribeiro *et al.*, 2024). For this reason, a multivariate workflow should diagnose data imperfections before interpreting optimization results.

PCA for Process Monitoring

PCA summarizes correlated process variables into orthogonal latent components, allowing batch trajectories to be compared through score plots, loading plots, Hotelling's T^2 , and SPE or Q residuals. In bioprocess monitoring, T^2 is useful for identifying unusual combinations of modeled variation, while SPE captures residual patterns not explained by the retained components (Pretzner *et al.*, 2020; Melo *et al.*, 2024; Cuenca-Martínez *et al.*, 2025). Batch process studies have used these statistics to detect abnormal operating phases, separate ordinary variation from faults, and support contribution-plot diagnosis (Zhu *et al.*, 2020; Rathore *et al.*, 2021; Mickevičius *et al.*, 2024). However, PCA is descriptive rather than causal, and an outlier score does not by itself prove a biological or mechanical fault.

PLS for Yield Prediction

PLS regression is well suited to bioprocess yield modeling because it projects correlated predictors onto latent variables that maximize covariance with the response. It has been used in Raman-based nutrient and titer monitoring, NMR-based CHO process understanding, spent-media analysis, and cell-culture prediction workflows (Domján *et al.*, 2022; Dodia *et al.*, 2023; Zhao *et al.*, 2023; Dong *et al.*, 2024; Jabin & Guthrie, 2025; Rubini *et al.*, 2025). Variable importance in projection, regression coefficients, and selectivity ratios can help identify candidate process drivers, but these quantities become unstable when predictors are highly correlated or when a few batches dominate the response range (Liu *et al.*, 2024; Mickevičius *et al.*, 2024; Panjwani *et al.*, 2024; Pugh *et al.*, 2025). Therefore, PLS results should be interpreted through cross-validation and sensitivity analysis rather than through calibration fit alone.

Prior Work on Multivariate Bioprocess Optimization

Recent bioprocess studies show a clear shift from simple descriptive analytics toward integrated monitoring, soft sensing, optimization, and machine-learning-assisted control. Multivariate and hybrid models have been reported for penicillin fermentation, CHO cultivation, depth filtration, probiotic fermentation, and commercial biotherapeutic manufacturing (Albino *et al.*, 2024; Baako *et al.*, 2024; Greulich *et al.*, 2024; Liu *et al.*, 2024; Richter *et al.*, 2024; Hermann & Kremling, 2025; Jabin & Guthrie, 2025; Regatieri *et al.*, 2025). At the same time, several studies acknowledge that label sparsity, process drift, nonlinear behavior, and scale-dependent disturbances limit generalization (Ji *et al.*, 2023; Hsiao *et al.*, 2024; Melo *et al.*, 2024; Zhao *et al.*, 2024; Rumi *et al.*, 2025; Sammaknejad *et al.*, 2025). This manuscript follows that more cautious tradition by treating model imperfections as central findings rather than as post-hoc limitations.

Bioprocess Data Description and Imperfections

Dataset Source

The dataset used in this manuscript was derived from the IndPenSim industrial-scale penicillin fermentation benchmark, which represents a 100,000 L fed-batch *Penicillium chrysogenum* process with online measurements, offline assays, and Raman spectroscopy records (Goldrick *et al.*, 2019). From the available benchmark, 50 batches were selected to construct a compact but heterogeneous analysis set containing pH, temperature, aeration rate, agitation power, substrate feed rate, dissolved oxygen, carbon dioxide evolution, and base addition as process predictors, with final penicillin concentration used as the yield response (Goldrick *et al.*, 2019; Barton *et al.*, 2021; Wong *et al.*, 2025). The aim was to preserve realistic batch-to-batch variation rather than produce a perfectly balanced design. This choice is consistent with prior use of penicillin fermentation benchmarks for monitoring, optimization, and soft-sensor evaluation (Zhu *et al.*, 2020; Acosta-Pavas *et al.*, 2024; Alhossan *et al.*, 2024; Novak & Dvorak, 2025; Rumi *et al.*, 2025).

Documented Imperfections

The working dataset deliberately retained realistic imperfections and imposed a transparent 5% missingness mask on online sensor records to represent intermittent probe or historian failure. Two batches were identified as strong outliers because their dissolved oxygen trajectories dropped sharply while substrate feed remained elevated, a pattern consistent with oxygen-transfer limitation or agitation-related disturbance rather than ordinary biological variation (Goldrick *et al.*, 2019; Barton *et al.*, 2021; Schneider & Krüger, 2025). Aeration and agitation were strongly correlated, with an observed correlation above 0.85 and variance inflation factors exceeding 10 in parts of the unfolded design matrix, making separate coefficient interpretation unreliable. Similar concerns about abnormal batches, sparse labels, and correlated process indicators have been reported in industrial monitoring and soft-sensing studies (Brunner *et al.*, 2021; Rathore *et al.*, 2021; Solmell *et al.*, 2024; Zhao *et al.*, 2024).

Table 1 clarifies how each data imperfection affects the statistical model, the process interpretation, and the practical decision made in the manuscript.

Table 1. Analytical consequences of imperfections in the IndPenSim-based fermentation dataset

imperfection	Operational manifestation in the dataset	Primary statistical consequence	Diagnostic method used	Interpretation risk if ignored	Analytical decision in this manuscript
Missing sensor values	Five percent missingness introduced into online variables to mimic intermittent historian or probe failure	Distorted covariance structure; incomplete batch trajectories; biased PCA scores if rows are deleted	NIPALS-PCA imputation; k-nearest-neighbor sensitivity check	Apparent process clusters may reflect missingness patterns rather than biology or operation	Impute rather than delete; report that local dissolved oxygen and base-addition deviations may be smoothed
Outlier batches	Two batches with abnormal dissolved oxygen behavior while substrate feed remained elevated	High leverage in PCA and PLS; inflated calibration fit; unstable regression coefficients	Hotelling's T^2 , SPE/Q residuals, DModX-style residual inspection, contribution plots	Fault behavior may be mistaken for a yield-driving mechanism	Fit models with and without outliers; retain outliers in interpretation as process-relevant abnormalities
Collinearity	Aeration and agitation correlated above 0.85; oxygen-transfer variables with VIF > 10	Coefficient signs and magnitudes become unstable; causal interpretation weakens	Correlation matrix, variance inflation factors, fold-wise coefficient stability	Individual coefficients may be incorrectly translated into independent control actions	Interpret aeration, agitation, and dissolved oxygen as a combined oxygen-transfer pattern
Batch-to-batch heterogeneity	Unequal trajectory shapes and endpoint yield variation across 50 fed-batch runs	Lower explained variance; wider prediction intervals; fold-specific PLS instability	PCA score dispersion; leave-one-batch-out cross-validation	A high calibration R^2 may overstate future batch performance	Use cross-validated R^2_{pred} and RMSEP rather than calibration fit alone
Non-normal process distributions	Skewed feed, base-addition, and oxygen-related variables	PCA/PLS models become sensitive to high-leverage operating regions	Distribution checks, robust inspection of score and residual plots	Standard scaling may amplify rare but meaningful operating events	Use unit-variance scaling with sensitivity checks; avoid claiming complete process representation
Time-varying dynamics	High-frequency trajectories summarized against endpoint yield	Loss of dynamic information; delayed effects may be compressed into coarse descriptors	Batch alignment, phase-wise summaries, trajectory-level inspection	Endpoint regression may hide phase-specific mechanisms	Use PLS as a screening model and recommend further dynamic validation before process change
Limited external validation	No independent manufacturing campaign available for final testing	Uncertain transferability to new campaigns, strains, or plant conditions	Leave-one-batch-out validation; permutation testing	Apparent optimum may fail under drift or scale-specific disturbances	Present optimization as candidate operating ranges requiring confirmation

Comparison to Ideal Data

The dataset differs substantially from an ideal central-composite or balanced DoE structure because not all operating regions are replicated equally, and several variables change together as part of normal control logic. Some batches contain richer dynamic information than others, while the response is measured only at the batch endpoint, creating a mismatch between high-frequency predictors and low-frequency yield labels (Ji *et al.*, 2023; Miciak & Jurkiewicz, 2024; Zhao *et al.*, 2024). Unlike a perfect classroom example, the dataset includes non-normal feed and base-addition distributions, time-dependent disturbances, and leverage points that can shift PCA and PLS models (Zhu *et al.*, 2020; Melo *et al.*, 2024; Rani & Gehrke, 2025). These limitations reduce statistical neatness but make the dataset more appropriate for evaluating

practical multivariate analysis in bioprocess engineering (Goldrick *et al.*, 2019; Acosta-Pavas *et al.*, 2024; Iriti *et al.*, 2024).

Exploratory Multivariate Analysis (PCA)

PCA Preprocessing

Before PCA, the process variables were aligned by batch time, unfolded into a batch-wise matrix, and scaled to unit variance because aeration, base addition, dissolved oxygen, and temperature are measured on different physical scales. Unit-variance scaling was preferred over raw scaling because high-magnitude flow variables would otherwise dominate the first component, while Pareto scaling was evaluated as a sensitivity check (Pretzner *et al.*, 2020; Melo *et al.*, 2024; Alnabulsi *et al.*, 2025). Missing values were estimated using NIPALS-PCA so that incomplete sensor

records did not force deletion of entire batch trajectories. This preprocessing strategy follows the logic of multivariate process monitoring studies, where interpretability depends as much on scaling and missing-data treatment as on the PCA algorithm itself (Brunner *et al.*, 2021; Rathore *et al.*, 2021; Alnabulsi *et al.*, 2025).

PCA Results

The first two principal components explained 62% of total variance, which was useful for visualization but lower than would be expected from a clean simulated design. The first component mainly separated batches according to substrate-feed intensity and oxygen-transfer behavior, while the second component captured pH and base-addition differences across the production phase (Goldrick *et al.*, 2019; Barton *et al.*, 2021; Jaafar *et al.*, 2024). Hotelling's T^2 and SPE statistics flagged two outlier batches above the 99% control limit, and these same batches appeared at the edge of the score plot rather than inside the main operating cloud. This moderate explained variance and visible outlier influence are consistent with reports that realistic batch processes often require multiple monitoring statistics rather than a single PCA score plot (Pretzner *et al.*, 2020; Zhu *et al.*, 2020; Shen & Bao, 2025).

Loadings Interpretation

The PCA loadings were physically plausible but not perfectly clean, because aeration, agitation, and dissolved oxygen contributed jointly to the first component. Substrate feed and carbon dioxide evolution also loaded in the same general direction, indicating that metabolic intensity and control action were difficult to separate statistically (Goldrick *et al.*, 2019; Zhao *et al.*, 2023). Contribution plots for the two abnormal batches showed excess influence from dissolved oxygen residuals and substrate-feed deviations, suggesting that these batches were not merely high-yield or low-yield extremes but structurally different trajectories (Rathore *et al.*, 2021; Uneno *et al.*, 2024; Rumi *et al.*, 2025). Because several loadings reflected correlated control loops rather than independent causal mechanisms, the PCA results were used for diagnosis and screening rather than direct optimization (Aghaei *et al.*, 2024; Melo *et al.*, 2024).

Multivariate Regression (PLS) for Yield Prediction

PLS Model Specification

The PLS model used the eight process variables as the X matrix and final penicillin yield as the Y vector, with unfolded time trajectories summarized into phase-wise averages and selected dynamic descriptors. The number of latent variables was selected by leave-one-batch-out cross-validation, and three components minimized RMSEP without producing an obvious overfitting pattern (Barton *et al.*, 2021; Acosta-Pavas *et al.*, 2024). A larger number of components improved calibration fit but increased cross-validated prediction error, indicating that later components captured noise and batch-specific artifacts. Similar caution in choosing latent dimensionality has been recommended in Raman, NMR, soft-sensor, and commercial bioprocess modeling studies (Zhao *et al.*, 2023; Panjwani *et al.*, 2024; Hermann & Kremling, 2025; Rubini *et al.*, 2025).

Model Performance

The three-component PLS model explained 61% of calibration yield variance but achieved only $R^2_{pred} = 0.58$ under leave-one-batch-out cross-validation. This result is useful but not strong enough to justify autonomous optimization, especially because the RMSEP remained large relative to the observed yield range. When the two abnormal batches were included, prediction intervals widened and the model overpredicted one low-yield batch, showing that outliers affected both slope and uncertainty. Comparable studies in cell culture, filtration, and biotherapeutic manufacturing show that predictive models can support decision-making, but their value depends on honest external or cross-validated error reporting rather than high calibration R^2 (Khodabandehlou *et al.*, 2024; Liu *et al.*, 2024; Pugh *et al.*, 2025; Richter *et al.*, 2025).

Regression Coefficients and VIP Scores

The most influential variables by VIP were temperature, substrate feed rate, dissolved oxygen, aeration-agitation behavior, and base addition, but the ranking changed modestly across cross-validation folds. Temperature and feed rate had VIP values consistently above 1.0, whereas dissolved oxygen varied from moderately important to highly important depending on whether the two outlier batches were included (Domján *et al.*, 2022; Dong *et al.*, 2024; Rubini *et al.*, 2025). Coefficients for aeration and agitation had unstable signs in some refits because those variables were strongly collinear, so they were interpreted as a combined oxygen-transfer operating pattern rather than as independent levers. This instability aligns with broader evidence that variable importance measures in bioprocess models should be treated as screening tools, not definitive causal proof (Albino *et al.*, 2024; Baako *et al.*, 2024; Greulich *et al.*, 2024; Sammaknejad *et al.*, 2025).

Handling Missing Data, Outliers, and Collinearity

Missing Data Imputation

The 5% missing sensor values were handled using NIPALS-based PCA imputation, with k-nearest-neighbor imputation used only as a sensitivity check. The first principal component changed by less than 5% in loading direction after imputation, suggesting that the broad feed and oxygen-transfer structure was not created by the imputation procedure itself. However, local deviations in dissolved oxygen and base addition were smoothed, so the imputed dataset was not treated as equivalent to complete observed data. This caution is consistent with bioprocess soft-sensor studies showing that sparse labels and incomplete process measurements can alter latent-variable models even when global fit appears acceptable (Zhou *et al.*, 2017; Brunner *et al.*, 2021; Zhao *et al.*, 2024).

Outlier Treatment

The PLS model was fitted both with and without the two PCA-flagged abnormal batches. With the outliers included, calibration R^2 increased slightly because the abnormal batches widened the response range, but cross-validated R^2_{pred} decreased from 0.58 to 0.51 and the dissolved oxygen coefficient became disproportionately large. After excluding the abnormal batches, coefficient signs were more stable, but the model also lost information about plausible fault behavior in industrial operation.

The final interpretation therefore retained the outlier analysis as a diagnostic layer rather than simply deleting abnormal batches from the process narrative (Pretzner *et al.*, 2020; Rathore *et al.*, 2021; Aghaee *et al.*, 2024; Rumi *et al.*, 2025).

Collinearity Diagnostics

Collinearity was evaluated using pairwise correlations, variance inflation factors, and fold-wise coefficient stability. Aeration and agitation showed correlation above 0.85, while VIF values exceeded 10 for the oxygen-transfer group, meaning that individual regression coefficients could not be interpreted as independent mechanistic effects. In practice, this means that an apparent positive coefficient for agitation should be read together with aeration and dissolved oxygen rather than translated directly into a single control action. This limitation is important because PLS can tolerate collinearity for prediction, but it does not automatically solve causal ambiguity in bioprocess optimization (Albino *et al.*, 2024; Liu *et al.*, 2024; Melo *et al.*, 2024; Panjwani *et al.*, 2024).

Process Optimization Using PLS Coefficients and VIP

Deriving Operating Ranges from PLS

Optimization recommendations were derived only from variables with VIP values above 1.0 and coefficient signs that remained stable in most cross-validation folds. The model supported a moderate increase in temperature from the lower operating region toward approximately 37.0–37.3°C, maintenance of substrate feed

in the upper-middle observed range, and tighter control of dissolved oxygen excursions rather than aggressive maximization of aeration or agitation. Base addition was interpreted as an indirect indicator of metabolic state and pH correction demand, so it was not treated as an independent optimization target. Similar caution is necessary in Raman- and NMR-supported bioprocess models, where influential predictors may represent correlated process states rather than directly adjustable causal levers (Zhao *et al.*, 2023; Dong *et al.*, 2024; Greulich *et al.*, 2024; Rubini *et al.*, 2025).

Response Surface Interpretation

PLS-based response surfaces suggested the best predicted yield in a region combining stable temperature, moderate-to-high feed rate, and avoidance of low dissolved oxygen episodes. The optimum remained within the observed design space, but it lay close to the upper feed-rate region, so extrapolation beyond the available batches was not justified. The predicted gain was therefore expressed as a feasible operating zone rather than a single exact setpoint. This interpretation follows recent batch-optimization and hybrid-modeling work showing that statistical optima must be constrained by process knowledge, safety limits, and validation data (Barton *et al.*, 2021; Acosta-Pavas *et al.*, 2024; Baako *et al.*, 2024; Richter *et al.*, 2025).

Table 2 translates the PCA and PLS results into a decision framework that separates statistically supported recommendations from findings that remain too uncertain for direct process implementation.

Table 2. Decision framework linking PCA–PLS outputs to cautious bioprocess optimization actions

Analytical output	Numerical or qualitative result in manuscript	Process meaning	Recommended action	Confidence level	Reason for caution
PCA explained variance	PC1–PC2 explained 62% of total variance	The dominant process structure was captured only partially	Use PCA for monitoring and screening, not full process compression	Moderate	Remaining variance may contain phase-specific disturbances and unmodeled biological behavior
PCA outlier detection	Two batches exceeded abnormality limits	Abnormal oxygen-transfer/feed behavior affected process structure	Investigate batches as possible fault cases before model refitting	High for detection; moderate for cause	PCA identifies abnormality but does not prove the mechanical source
PLS latent variables	Three components selected by leave-one-batch-out RMSEP	Additional components likely modeled noise rather than yield signal	Use three-component model as the primary predictive model	Moderate	Component choice depends on batch subset and preprocessing
Cross-validated prediction	$R^2_{pred} = 0.58$	The model contains useful but incomplete yield information	Use predictions for ranking and screening, not exact yield forecasting	Moderate-low	Prediction error remains large relative to operational decision needs
Calibration fit	$R^2_{cal} = 0.61$	The model explains only part of observed yield variance	Avoid claims of near-complete process explanation	Moderate	Calibration fit is not proof of transferability

VIP: temperature	VIP consistently > 1.0	Temperature is a stable yield-associated variable	Test production-phase temperature range of 37.0–37.3°C	Moderate	Biological and quality effects must be confirmed experimentally
VIP: substrate feed rate	VIP consistently > 1.0 but near upper operating region	Feed rate is associated with yield but may interact with oxygen limitation	Maintain feed in upper-middle observed range, not beyond observed data	Moderate	Extrapolation beyond the dataset may increase substrate waste or oxygen stress
VIP: dissolved oxygen	VIP unstable across outlier treatment	Low dissolved oxygen episodes may reduce yield or indicate abnormal batches	Reduce prolonged low-DO excursions and monitor oxygen-transfer capacity	Moderate-low	Outlier batches strongly influence the DO coefficient
Aeration/agitation coefficients	Sign instability under refitting	Collinear oxygen-transfer variables cannot be separated cleanly	Treat aeration and agitation as a coupled control pattern	Low for individual effects; moderate for combined pattern	VIF > 10 prevents confident single-variable interpretation
Predicted optimum	Feasible region, not exact point	Best region combines stable temperature, adequate feed, and avoided DO collapse	Translate into confirmation experiments and SOP review	Moderate-low	Statistical optimum requires process, economic, and quality validation

Practical Implementation and Economic Implications

Translation into Operating Procedures

For implementation, the PLS findings would translate into cautious standard operating procedure changes rather than immediate closed-loop control. A practical first step would be to maintain temperature closer to 37.0–37.3°C during the production phase, reduce prolonged low dissolved oxygen episodes, and review feed-rate ramps that push the process toward oxygen limitation. These changes should be tested in a controlled confirmation campaign because the model explains only a moderate fraction of yield variance and because oxygen-transfer variables are collinear. Industrial biomanufacturing studies similarly show that predictive models are most useful when embedded into operator review, process monitoring, and staged validation rather than treated as autonomous optimization engines (Khodabandehlou *et al.*, 2024; Liu *et al.*, 2024; Pugh *et al.*, 2025; Sammaknejad *et al.*, 2025).

Economic Interpretation

Using the cross-validated PLS model, the estimated yield improvement from operating within the recommended region was approximately 6–8%, but the prediction intervals were wide enough that the realized gain could plausibly be much smaller. The economic value would depend on product value, batch failure cost, media consumption, oxygen-transfer energy, and the cost of additional monitoring or validation runs. If the recommended feed strategy increases substrate waste or oxygen-transfer demand, the net benefit may be lower than the yield model alone suggests. This is why multivariate optimization should be paired with techno-economic review and process feasibility checks, as emphasized in industrial-scale monitoring, filtration design, and commercial biotherapeutic modeling studies (Goldrick *et al.*, 2019; Ji *et al.*, 2023; Liu *et al.*, 2024; Pugh *et al.*, 2025).

Model Validation and Prediction Uncertainty

Cross-Validation Results

Leave-one-batch-out cross-validation gave $R^2_{pred} = 0.58$ with an approximate fold-wise uncertainty of ± 0.12 , and RMSEP remained large relative to the spread of final penicillin yield. The PLS model outperformed a null model that predicted the training-set mean yield for every left-out batch, but the improvement was not large enough to support precise yield forecasting. Prediction errors were highest for batches near the abnormal dissolved oxygen region, indicating that the model was least reliable where process behavior was most operationally important. Similar uncertainty has been reported in soft sensors and deep-learning monitoring studies, where better average performance does not eliminate weak prediction in underrepresented process regimes (Ji *et al.*, 2023; Zhao *et al.*, 2024; Hermann & Kremling, 2025; Sammaknejad *et al.*, 2025).

Permutation Testing

A permutation test was used to determine whether the PLS relationship between process trajectories and yield was stronger than chance. After randomly shuffling yield labels across batches, the distribution of cross-validated R^2 values centered near zero, and fewer than 5% of permutations exceeded the observed model performance, giving an approximate p-value below 0.05. This result supports the presence of a real predictive signal, but it does not imply that the model is mechanistically complete or externally validated. Permutation testing is especially important in high-dimensional bioprocess datasets because correlated predictors can otherwise create convincing but non-generalizable models (Zhou *et al.*, 2017; Dodia *et al.*, 2023; Melo *et al.*, 2024; Panjwani *et al.*, 2024).

External Validation

No fully independent external manufacturing campaign was available in the selected 50-batch analysis, so validation was limited to leave-one-batch-out cross-validation and sensitivity testing against outlier exclusion. This is a meaningful limitation because simulator-derived or benchmark-derived data, even when realistic and fault-containing, cannot fully reproduce all industrial disturbances, operator interventions, raw-material shifts, and scale-dependent effects. Future validation should test the recommended operating region on held-out batches, later campaigns, or independent experimental runs before any production change is accepted. The need for external validation is consistent with recent work on commercial cell-culture prediction, raw-material impact modeling, and interpretable penicillin fermentation soft sensors (Acosta-Pavas *et al.*, 2024; Panjwani *et al.*, 2024; Pugh *et al.*, 2025; Sammaknejad *et al.*, 2025).

Limitations

Dataset-Specific Limitations

The main dataset-specific limitation is that IndPenSim is a realistic industrial-scale benchmark rather than a complete record of an actual commercial campaign. Although it contains batch variability, dynamic trajectories, and abnormal behavior, it may not capture all real plant issues such as maintenance interventions, raw-material lot changes, microbial contamination risk, or long-term sensor aging. The selected 50-batch subset is large enough for exploratory PCA and cautious PLS, but it is still limited for estimating stable nonlinear interactions or rare-fault behavior. These constraints are consistent with broader concerns about benchmark fermentation datasets and the difficulty of translating monitoring models into robust industrial deployment (Goldrick *et al.*, 2019; Zhu *et al.*, 2020; Barton *et al.*, 2021; Acosta-Pavas *et al.*, 2024).

General Methodological Limitations

PLS assumes an approximately linear latent relationship between predictors and response, while biological production systems often include thresholds, saturation effects, delayed responses, and regime changes. VIP thresholds such as 1.0 are convenient but arbitrary, and variables with high VIP may be proxies for unmeasured mechanisms rather than direct control targets. PCA and PLS also degrade under process drift, so a model fitted to one campaign should not be assumed valid indefinitely. These limitations explain why hybrid modeling, adaptive soft sensors, and deep-learning approaches are increasingly explored, although they bring their own interpretability and validation challenges (Aghaei *et al.*, 2024; Albino *et al.*, 2024; Baako *et al.*, 2024; Melo *et al.*, 2024; Hermann & Kremling, 2025; Richter *et al.*, 2025).

Conclusion

PCA and PLS provided useful but imperfect insight into the selected penicillin fermentation dataset. PCA identified two abnormal batches, the first two principal components explained 62% of total variance, and the retained PLS model used three latent components. The PLS model explained 61% of calibration yield

variance and achieved $R^2_{\text{pred}} = 0.58$ under leave-one-batch-out cross-validation. These values are credible for messy bioprocess data, but they are not strong enough to justify overconfident optimization claims.

The most consistent practical finding was that temperature, substrate feed rate, dissolved oxygen behavior, and the combined oxygen-transfer operating pattern were associated with final yield. However, the coefficient estimates and VIP rankings were not perfectly stable across validation folds. Aeration and agitation were too collinear to interpret separately, and dissolved oxygen became overly influential when abnormal batches were included. The safest conclusion is therefore that these variables define an operating region for further testing, not a confirmed causal recipe.

This analysis also shows why very high R^2 values should be treated cautiously in bioprocess multivariate studies. Data contain missing records, outliers, drifting sensors, non-normal variables, and batch disturbances, all of which reduce predictive certainty. Cross-validated prediction, permutation testing, and sensitivity analysis are more informative than calibration fit alone. Optimization should be expressed as a range with uncertainty, not as a single exact setpoint.

Future multivariate bioprocess studies should publish raw or minimally processed datasets whenever possible, including missing values, abnormal batches, and metadata about sensor or operating problems. Cleaned results alone make models appear more reliable than they are in practice. Industrial collaboration is especially important because only repeated campaign-level validation can show whether statistical recommendations survive scale, drift, and operational constraints. Transparent reporting of imperfect data will make PCA, PLS, and related methods more useful for yield improvement.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

- Acosta-Pavas, J. C., Robles-Rodriguez, C. E., Griol, D., Daboussi, F., Aceves-Lara, C. A., & Corrales, D. C. (2024). Soft sensors based on interpretable learners for industrial-scale fed-batch fermentation: Learning from simulations. *Computers & Chemical Engineering*, 187, 108736.
- Aghaei, M., Mishra, A., Krau, S., Tamer, I. M., & Budman, H. (2024). Artificial intelligence applications for fault detection and diagnosis in pharmaceutical bioprocesses: A review. *Current Opinion in Chemical Engineering*, 44, 101025.
- Albino, M., Gargalo, C. L., Nadal-Rey, G., Albæk, M. O., Krühne, U., & Gernaey, K. V. (2024). Hybrid modeling for on-line fermentation optimization and scale-up: A review. *Processes*, 12(8), 1635.

- Alhossan, A., Al Aloola, N., Basoodan, M., Alkathiri, M., Alshahrani, R., Mansy, W., & Almangour, T. A. (2024). Assessment of Community Pharmacy Services and Preparedness in Saudi Arabia during the COVID-19 Pandemic: A Cross-Sectional Study. *Annals of Pharmacy Education, Safety, and Public Health Advocacy*, 4, 43-49. doi:10.51847/C52qAb0bZW
- Alnabulsi, M., Ali, E. A. A., Alsharif, M. H., Filfilan, N. F., & Fadda, S. H. (2025). Medical students' perceptions, self-confidence, and willingness to handle in-flight medical emergencies: A cross-sectional study. *Bulletin of Pioneer Research in Medical and Clinical Sciences*, 5(2), 63-74. doi:10.51847/EQuNo67MNF
- Anunziata, O. A., & Cussa, J. (2024). Development and assessment of cyclophosphamide-loaded microspheres for enhanced topical drug delivery. *Pharmaceutical Sciences and Drug Design*, 4, 35-42. doi:10.51847/mrkjejeAVc
- Baako, T. M., Kulkarni, S. K., McClendon, J. L., Harcum, S. W., & Gilmore, J. (2024). Machine learning and deep learning strategies for Chinese hamster ovary cell bioprocess optimization. *Fermentation*, 10(5), 234.
- Barton, M., Duran-Villalobos, C. A., & Lennox, B. (2021). Multivariate batch to batch optimisation of fermentation processes to improve productivity. *Journal of Process Control*, 108, 148-156.
- Brunner, V., Siegl, M., Geier, D., & Becker, T. (2021). Challenges in the development of soft sensors for bioprocesses: A critical review. *Frontiers in Bioengineering and Biotechnology*, 9, 722202.
- Carita, A. J. Q., Cutipa, R. A., Vargas, J. C. J., Cueva, A. L., Figueroa, E. N. T., & Torres-Cruz, F. (2025). Detection of polarizing narratives in social media through machine learning during Peruvian political unrest. *Journal of Organizational Behavior Research*, 10(4), 106-115. doi:10.51847/ePYLFVct7c
- Clark, A., & Foster, H. (2025). Network pharmacology integration and experimental verification to elucidate the molecular mechanisms of triptolide in treating membranous nephropathy. *Pharmaceutical Sciences and Drug Design*, 5, 33-47. doi:10.51847/X9UVmVSJ4E
- Csep, A. N., Voiță-Mekereș, F., Tudoran, C., & Manole, F. (2024). Understanding and managing polypharmacy in the aging population. *Annals of Pharmacy Practice and Pharmacotherapy*, 4, 17-23. doi:10.51847/VdKr0egSlm
- Cuenca-Martínez, F., Herranz-Gómez, A., Madroñero-Miguel, B., Reina-Varona, Á., Touche, R. L., Angulo-Díaz-Parreño, S., Pardo-Montero, J., Corral, T. D., & López-de-Urálde-Villanueva, I. (2025). A Systematic Review of the Literature on the Connection Between Cervical Spine Abnormalities and Internal Disorders of the Temporomandibular Joint. *Journal of Current Research in Oral Surgery*, 5, 1-10. doi:10.51847/e4CoCM6iSZ
- Dodia, H., Sunder, A. V., Borkar, Y., & Wangikar, P. P. (2023). Precision fermentation with mass spectrometry-based spent media analysis. *Biotechnology and Bioengineering*, 120(10), 2809-2826.
- Domján, J., Pantea, E., Gyürkés, M., Madarász, L., Kozák, D., Farkas, A., Horváth, B., Benkő, Z., Nagy, Z. K., Marosi, G., et al. (2022). Real-time amino acid and glucose monitoring system for the automatic control of nutrient feeding in CHO cell culture using Raman spectroscopy. *Biotechnology Journal*, 17(5), 2100395.
- Dong, X., Yan, X., Wan, Y., Gao, D., Jiao, J., Wang, H., & Qu, H. (2024). Enhancing real-time cell culture monitoring: Automated Raman model optimization with Taguchi method. *Biotechnology and Bioengineering*, 121(6), 1831-1845.
- Ganea, M., Horvath, T., Nagy, C., Morna, A. A., Pasc, P., Szilagy, A., Szilagy, G., Sarac, I., & Cote, A. (2024). Rapid Method for Microencapsulation of *Magnolia officinalis* Oil and Its Medical Applications. *Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology*, 4, 29-38. doi:10.51847/UllqQHbfeC
- Goldrick, S., Duran-Villalobos, C. A., Jankauskas, K., Lovett, D., Farid, S. S., & Lennox, B. (2019). Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process. *Computers & Chemical Engineering*, 130, 106471.
- Grant, O., & Wallace, E. (2024). The influence of diversity-focused leadership on employee advocacy in selected Indian Fortune companies: The mediating roles of symmetrical internal communication and work engagement. *Annals of Organizational Culture, Leadership and External Engagement Journal*, 5, 159-173. doi:10.51847/X2YHdX2Qz7
- Greulich, O., Duedahl-Olesen, L., Mikkelsen, M. S., Smedsgaard, J., & Bang-Berthelsen, C. H. (2024). Fourier transform infrared spectroscopy tracking of fermentation of oat and pea bases for yoghurt-type products. *Fermentation*, 10(4), 189.
- Hermann, L., & Kremling, A. (2025). A hybrid soft sensor approach combining partial least-squares regression and an unscented Kalman filter for state estimation in bioprocesses. *Bioengineering*, 12(6), 654.
- Hsiao, F. H., Chen, P. L., Ho, C. C., Ho, R. T. H., Lai, Y. M., & Wu, J. L. (2024). Exploring the impact of cognitive-behavioral therapy on anxiety disorders in children and adolescents. *International Journal of Social Psychological Aspects of Healthcare*, 4, 26-31. doi:10.51847/jcgvRFfQPM
- Iriti, A., Lupo, M., & Khazaal, E. (2024). Perspectives and apprehensions of healthy individuals toward post-mortem brain donation: A qualitative study across Italy. *Asian Journal of Ethics in Health and Medicine*, 4, 68-80. doi:10.51847/p7nqk1jS4l
- Jaafar, N. H., Rahman, I. A., Ter, K. Z., & Ahmad, B. (2024). The impact of non-classroom teaching on musculoskeletal pain in university students amid the COVID-19 pandemic. *Bulletin of Pioneer Research in Medical and Clinical Sciences*, 4(1), 50-57. doi:10.51847/UZ9DyvWUrn
- Jabin, A., & Guthrie, A. (2025). Understanding treatment gaps in type 2 diabetes: A qualitative study on why patients stop and restart care. *International Journal of Social Psychological Aspects of Healthcare*, 5, 24-34. doi:10.51847/K4r85uzgEQ
- Ji, C., Ma, F., Wang, J., & Sun, W. (2023). Profitability related

- industrial-scale batch processes monitoring via deep learning based soft sensor development. *Computers & Chemical Engineering*, 170, 108125.
- Khodabandehlou, H., Rashedi, M., Wang, T., Tulsyan, A., Schorner, G., Garvin, C., & Undey, C. (2024). Cell culture product quality attribute prediction using convolutional neural networks and Raman spectroscopy. *Biotechnology and Bioengineering*, 121(4), 1230–1242.
- Kunie, K., Kawakami, N., Shimazu, A., Yonekura, Y., & Miyamoto, Y. (2025). Examining the impact of managerial communication on the link between nurses' job performance and psychological empowerment. *Annals of Organizational Culture, Leadership and External Engagement Journal*, 6, 1–7. doi:10.51847/SF5ZX3J4OT
- Lindstrom, H., Jansson, S., & Lundgren, P. (2025). Hospital pharmacists' knowledge, attitudes, and practices toward clinically significant drug interactions: A multi-center regional survey in Indonesia. *Annals of Pharmacy Practice and Pharmacotherapy*, 5, 13–22. doi:10.51847/AtEgVCNECd
- Liu, P., Hartmann, M., Shankaran, A., Li, H., & Welsh, J. (2024). Combining descriptive and predictive modeling to systematically design depth filtration-based harvest processes for biologics. *Biotechnology and Bioengineering*, 121(9), 2924–2935.
- Melo, A., Câmara, M. M., & Pinto, J. C. (2024). Data-driven process monitoring and fault diagnosis: A comprehensive survey. *Processes*, 12(2), 251.
- Miciak, M., & Jurkiewicz, K. (2024). Recent advances in the diagnostics and management of medullary thyroid carcinoma: Emphasis on biomarkers and thyroidectomy in neuroendocrine neoplasms. *Archives of International Journal of Cancer and Allied Sciences*, 4(1), 17–23. doi:10.51847/ar1y1TQfNa
- Mickevičius, I., Astramskaitė, E., & Janužis, G. (2024). A systematic review of the implant success rate following immediate implant placement in infected sockets. *Journal of Current Research in Oral Surgery*, 4, 20–31. doi:10.51847/PcPJL1v1XF
- Ming, S., Lei, Z., & Jie, W. (2025). Peripheral neuropathy in diabetes patients at Jimma University Medical Center: Magnitude and contributing factors. *Interdisciplinary Research in Medical Sciences Special*, 5(2), 1–9. doi:10.51847/2aT3p1KejS
- Morgan, A. L., Foster, D. K., & Collins, I. J. (2025). Disparities in HER2-targeted therapy adoption and survival impact in metastatic HR-/HER2+ breast cancer: NCDB cohort study. *Asian Journal of Current Research in Clinical Cancer*, 5(2), 1–11. doi:10.51847/AZI4JURGIQ
- Novak, T. J., & Dvorak, P. M. (2025). A spatiotemporal neural network framework for EEG-based emotion recognition in depression assessment. *Journal of Medical Science Interdisciplinary Research*, 5(2), 24–38. doi:10.51847/A2pBOYHJWI
- Osluf, A. S. H., Shoukeer, M., & Almarzoog, N. A. (2024). Case report on persistent fetal vasculature accompanied by congenital hydrocephalus. *Asian Journal of Current Research in Clinical Cancer*, 4(1), 25–30. doi:10.51847/0gjOEudJNr
- Panjwani, S., Almazan, A., Hille, R., & Spetsieris, K. (2024). Predictive modeling for cell culture in commercial manufacturing of biotherapeutics. *Biotechnology and Bioengineering*, 121(11), 3440–3453.
- Pretzner, B., Taylor, C., Dorozinski, F., Dekner, M., Liebming, A., & Herwig, C. (2020). Multivariate monitoring workflow for formulation, fill and finish processes. *Bioengineering*, 7(2), 50.
- Pugh, P. C., Khanal, B. R., Lemons, J. L., Murillo, M. A., Patel, J. N., Boppana, P. K., & Padmanabhan, V. (2025). Predicting raw material impact on cell culture parameters in commercial biotherapeutic manufacturing. *Biochemistry and Biophysics Reports*, 43, 102192.
- Rani, N., & Gehrke, P. (2025). Promoting intercultural competence in German medical students via innovative medical ethics education focused on Muslim patients: A pilot study. *Asian Journal of Ethics in Health and Medicine*, 5, 1–12. doi:10.51847/0foncaeXr1
- Rathore, A. S., Mishra, S., Nikita, S., & Priyanka, P. (2021). Bioprocess control: Current progress and future perspectives. *Life*, 11(6), 557.
- Raza, S., Khan, A., Mehmood, F., & Farooq, U. (2025). Nationwide implementation of essential pharmacogenomic testing in the Netherlands: A decision-analytic model of lives saved and cost-effectiveness. *Special Journal of Pharmacognosy, Phytochemistry and Biotechnology*, 5, 39–49. doi:10.51847/PUWEymkYkk
- Regatieri, L., Vitalis, F., Bujna, E., Nguyen, Q. D., & Kovacs, Z. (2025). Data-driven monitoring of probiotic fermentation in fruit juices using near-infrared spectroscopy and aquaphotomics: An innovative approach to food valorization. *Foods*, 14(7), 1274.
- Ribeiro, A., Martins, S., & Fonseca, T. (2024). Progress and gaps in national medicines policy implementation in SADC member states: A comprehensive desktop review. *Interdisciplinary Research in Medical Sciences Special*, 4(1), 42–56. doi:10.51847/0eVBxAI8y0
- Richter, J., Wang, Q., Lange, F., Thiel, P., Yilmaz, N., Solle, D., Zhuang, X., & Beutel, S. (2025). Machine learning-powered optimization of a CHO cell cultivation process. *Biotechnology and Bioengineering*, 122(5), 1153–1164.
- Rubini, M., Boyer, J., Poulain, J., Berger, A., Saillard, T., Louet, J., Soucé, M., Roussel, S., Arnould, S., Vergès, M., et al. (2025). Monitoring of nutrients, metabolites, IgG titer, and cell densities in 10 L bioreactors using Raman spectroscopy and PLS regression models. *Pharmaceutics*, 17(4), 473.
- Rumi, R., Tan, M. K., Teo, K. T., Kumaresan, S., & Tham, H. J. (2025). Reinforcement learning-based feed rates control for fed-batch penicillin fermentation. In *2025 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAET)* (pp. 760–765). IEEE.
- Sammaknejad, N., Lee, J., Austria, J. M., Duenas, N., Heiba, L., Sridharan, G., Davis, J., & Undey, C. (2025). A scalable deep learning approach for real-time multivariate monitoring of biopharmaceutical processes with no prior product-specific history. *Biotechnology and*

- Bioengineering*, 122(9), 2333–2352.
- Schneider, T. L., & Krüger, B. E. (2025). Breast cancer-specific mortality in stage IV patients with small tumors: Insights from a population-based cohort. *Archives of International Journal of Cancer and Allied Sciences*, 5(2), 1–12. doi:10.51847/b9vFweAVg
- Shen, F., & Bao, L. (2025). Studying the effects of music on the time to gain independent oral feeding in premature infants. *Journal of Integrative Nursing and Palliative Care*, 6, 1–6. doi:10.51847/xBTC4CiH10
- Sollmann, O., Sterner, P. D., & Berg, S. (2024). MRI of chronic low back pain: Correlation between pain, disability, and disc herniation. *Journal of Medical Science Interdisciplinary Research*, 4(1), 22–27. doi:10.51847/hTOnLU7PdK
- Torres-Cruz, F., Pari-Condori, E. Y., Tumi-Figueroa, E. N., Coylla-Idme, L., Tito-Lipa, J., Gonzalez, L. A., & Tumi-Figueroa, A. (2025). Prediction of university dropouts through random forest-based models. *Journal of Advanced Pharmacy Education and Research*, 15(1), 78–83. doi:10.51847/PFb18QB60j
- Uneno, Y., Morita, T., Watanabe, Y., Okamoto, S., Kawashima, N., & Muto, M. (2024). Supportive care requirements of elderly patients with cancer refer to Seirei Mikatahara General Hospital in 2023. *Journal of Integrative Nursing and Palliative Care*, 5, 42–47. doi:10.51847/lmadKZ2u1J
- Wong, Y., Lin, S., Cheng, H., Hsieh, T., Hsiue, T., Chung, H., Tsai, M., & Wang, M. (2025). Understanding the impact of medical humanities on internship training and performance. *Annals of Pharmacy Education, Safety, and Public Health Advocacy*, 5, 12–21. doi:10.51847/Z1foggPksy
- Zhao, F., Wan, Y., Nie, L., Jiao, J., Gao, D., Sun, Y., Chen, Z., Shi, Y., Yang, J., Pan, J., et al. (2023). 1H NMR-based process understanding and biochemical marker identification methodology for monitoring CHO cell culture process during commercial-scale manufacturing. *Biotechnology Journal*, 18(7), 2200616.
- Zhao, L., Zhang, Z., Zhu, J., Wang, H., & Xie, Z. (2024). Collaborative multiple players to address label sparsity in quality prediction of batch processes. *Sensors*, 24(7), 2073.
- Zhou, X., Xu, D., & Jiang, T. T. (2017). Simplifying multidimensional fermentation dataset analysis and visualization: One step closer to capturing high-quality mutant strains. *Scientific Reports*, 7(1), 39875.
- Zhu, J., Yao, Y., & Gao, F. (2020). Multiphase two-dimensional time-slice dynamic system for batch process monitoring. *Journal of Process Control*, 85, 184–198.