Identifying Drivers of Downstream Yield Variability Using Integrated Process Models: An Application to API Manufacturing

Tobias Overgaard*,, Maria-Ona Bertran***, John Bagterp Jørgensen*, Bo Friis Nielsen***

**Technical University of Denmark, Department of Applied Mathematics and Computer Science, Kgs. Lyngby, Denmark (tobov@dtu.dk)*
***Novo Nordisk A/S, PS API Manufacturing, Science & Technology, Bagsværd, Denmark*
****Novo Nordisk A/S, PS API Expansions, Kalundborg, Denmark*

Abstract: We introduce a novel two-level method to address systematic yield variability in biopharmaceutical batch processes. At the first level (inter-step), we utilize process-wide connectivity data to identify the specific process step where performance variability occurs. A sequential and orthogonalized partial least squares (SO-PLS) model is then developed to trace the origin of these variabilities, linking data blocks across the flowsheet and filtering correlated information. Once a critical step is identified, the second level (intra-step) employs unit-specific PLS models to capture the internal dynamics of that step, using entire batch trajectories for modeling. In collaboration with process experts, this level isolates variable trajectories that drive the systematic variability. Applied to a commercial batch process producing an active pharmaceutical ingredient (API), this method reveals that downstream yield is impacted by variability during cell culture production. Furthermore, a detailed analysis of bioreactor data identifies key manipulated variable trajectories, specifically the dosage of glucose and NH_3, impacting cell culture production. Validation of process improvement hypotheses is conducted in collaboration with process experts, enhancing transparency and yielding valuable insights.

Keywords: biopharmaceutical processes, batch process modeling and control, process optimization, data mining tools, artificial intelligence and machine learning, process and performance monitoring

1. INTRODUCTION

Batch processes are widely used in the pharmaceutical industry for their flexibility in producing various products with a single equipment stream, their faster time to market for new products, and their ease of adaptation to regulatory standards (Korovessi & Linninger, 2005). Ideally, each batch would be processed under optimal conditions to ensure consistent and stable product yield. However, the time-varying and non-linear nature of batch processes often leads to yield fluctuations, causing productivity losses.

Variability in raw materials, initial conditions, and upstream influence on downstream process steps can cause deviations from baseline yield (Barton et al., 2021). While statistical process control (SPC) is used to monitor process performance and detect *special cause variation* (Bisgaard & Kulahci, 2007), this work focuses on utilizing historical data from the entire manufacturing process to identify drivers of *common cause variation*. By tracking and adjusting low-performing batches from an intended yield target, we aim to enhance process understanding and identify opportunities for optimization.

One approach is to model the correlation between product yield and various process and quality variables across multiple process steps. Examples of such a model include *integrated process models* that represent processes with multiple steps or unit operations. These models transform the output of one unit into the input of the next, simulating material flow and capturing interactions between upstream and downstream equipment. Applications include control strategy design, risk analysis, experimental design, and process optimization (Marschall et al., 2022; Zahel et al., 2017; Diab et al., 2022), and they range from first-principles-based to fully data-driven. Due to the absence of detailed physical models for each unit operation and the high-dimensional, non-causal, non-full rank, and low signal-to-noise nature of data from industrial pharmaceutical processes, multivariate statistical techniques, and in particular multiblock partial least squares (MB-PLS) models, have proven effective for integrated process modeling (Brás et al., 2004). Such approaches organize information from various process steps into distinct blocks ($\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, ..., $\boldsymbol{X}^{(S)}$) to model process-wide effects. For example, input variables measured at the end of each batch can be divided into meaningful blocks that are then related to a response block $\boldsymbol{Y}$, such as product purity at a downstream step.

Recently, sequential-orthogonalized partial least-squares (SO-PLS) regression has been employed to capture the connectivity of process flow diagrams and translate it into data-driven models, serving as soft sensors in multi-unit processes (Zhu et al., 2024). The block order is similar to the MB-PLS setup, but orthogonalization between blocks is used to eliminate overlapping data, ensuring each block retains only unique information. This capability allows SO-PLS to quantify the incremental contributions of different blocks to the output of interest, similar to variance decomposition. Cattaldo et al. (2024) have demonstrated some dynamic extensions and applications, and Lauzon-Gauthier et al. (2018) compared SO-PLS to similar methods, like sequential MB-PLS (SMB-PLS), concluding similar performance while emphasizing the extra layer of interpretation offered by orthogonalization.

In industrial pharmaceutical processes, batch reports link a specific batch number to multiple process steps, enabling yield prediction using process variables from different unit operations. However, challenges arise when streams mix and

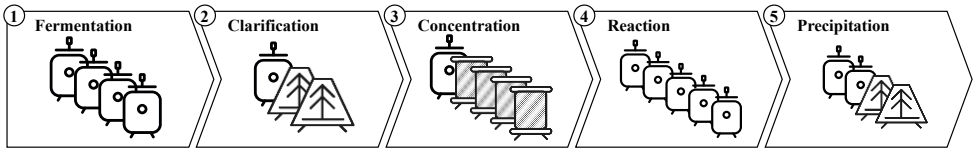

Figure 1. Process sequence involving fermentation and initial purification of an API manufacturing line.

split across units, complicating the tracking of batches and process conditions. In these cases, it is often best to investigate each unit separately, such as with individual PLS models (Kourti, 2006). This allows for the detection of specific events and identification of optimal variable trajectories to enhance overall productivity.

1.2 Contribution

To align with this recommendation from Kourti (2006), we propose a two-level approach that integrates process-wide connectivity data (*inter-step level*) with internal dynamics at a process step (*intra-step level*). At the inter-step level, we assume that systematic yield variability has been observed at a specific step, termed the *step of manifestation*, and fit an SO-PLS model on the entire process up to this step. The SO-PLS model's role is to isolate the *step of origin*, where variability likely begins. Once identified, we investigate the intra-step level to uncover internal dynamics and variable interactions that could cause yield variability. We apply this approach to identify yield variability drivers in a pharmaceutical batch process producing an API (Figure 1). Findings are validated and discussed between process experts and data scientists.

This study aims to answer the following questions: Which process steps most affect downstream product yield? Can their impacts be separated? Can process-wide data be linked to manipulated variable trajectories at a unit? How can expert input be integrated into the workflow? Using the two-level approach, we aim to identify key trajectories impacting yield and inform performance improvement decisions.

2. METHOD

Multiblock approaches, especially the SO-PLS model, allow integration of data from all process steps into a single model. This model considers the interactions between steps and their relative importance to the final product yield (inter-step model). Key process steps are then analyzed in detail to determine necessary changes in manipulated variable trajectories (intra-step model). The method involves three main stages (Figure 2):

1. *Scoping of the problem*
2. *Analysis of the integrated process model (inter-step level)*
3. *Analysis of unit operation model(s) (intra-step level)*

The methodology includes feedback loops when variance cannot be isolated within a single block, as detailed in the following subsections.

2.1 Scoping of the problem

Stage 1.a: The first stage involves detecting systematic yield variability in a process step, possibly using a control chart. This variability, influenced by known or unknown inputs, biases the yield in a specific direction, such as a seasonal effect or grouping. These differences can be tested using a t-test.

Stage 1.b: Next, we identify all relevant process steps and connecting streams leading to the observed step, guided by process flowsheets diagrams. The process is then divided into blocks representing individual steps.

Stage 1.c: Afterwards, we gather all necessary data for each step from plant and laboratory information systems. This includes process variables, raw material properties, and intermediate product properties. For yield modeling, key variables include the product concentration from each step.

2.2 Analysis of integrated process model (inter-step level)

In this stage, we identify process steps that most influence yield variability. This analysis considers how downstream units, intermediate products, and final products are affected by upstream raw material properties and process settings.

Stage 2.a: First, variables from each step are organized into matrices and preprocessed. In industrial batch processes, variables form a tensor with variable names, batches, and batch completion times as dimensions. Conversely, product property samples form a matrix, since they are often taken at the end of each operation (though occasionally they are available during processing). To align these variables, methods like unfolding can be used, treating each time instance as a new variable (Westerhuis et al., 1999). However, challenges such as uneven batch times and varying sample frequencies can complicate data alignment (Sartori et al., 2023). Since this stage is a

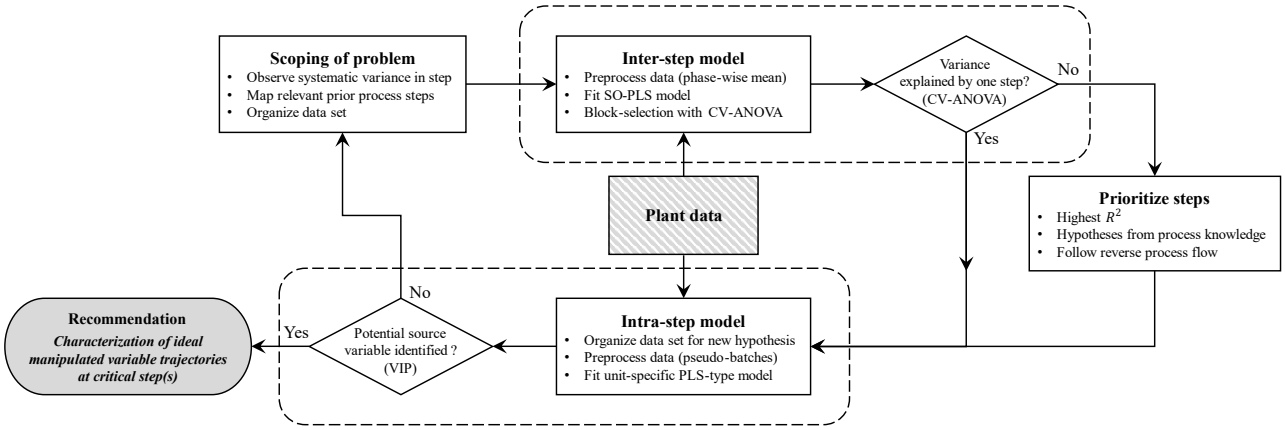

Figure 2. A methodology to identify drivers of downstream systematic variability in biopharmaceutical batch processes.

preliminary screening, a simpler approach is used, each process variable is aggregated into scalar values, using the average, minimum, and maximum across different processing phases. Higher-order moments of the variables may also be considered (Rendall et al., 2017). The data is then autoscaled. After preprocessing, relevant variables are selected using a backwards procedure. A PLS model is fitted for each block against the yield output, and variables are sorted by a *variable importance in projection* (VIP) index. Using a threshold (typically set at one), the least informative variables are eliminated, and the PLS model is refitted. This process is repeated until maximum model performance is achieved.

Stage 2.b: We then construct the integrated process model using the SO-PLS method to incorporate connectivity information across the manufacturing process, as proposed by Zhu et al. (2024). We start with an overview of PLS, the foundation of the SO-PLS algorithm.

The original PLS model (Wold et al., 2001) is a linear multivariate regression model that relates a matrix $\boldsymbol{X} \in \mathbb{R}^{N\times M}$ of M regressors (e.g., process variables) to a matrix $\boldsymbol{Y} \in \mathbb{R}^{N\times K}$ of K responses (e.g., product yields) for the same N observations (e.g., batches). PLS decomposes $\boldsymbol{X}$ and $\boldsymbol{Y}$ into a reduced space of V orthogonal latent variables (LVs) as follows

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\intercal} + \boldsymbol{E},$$
$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^{\intercal} + \boldsymbol{F}.$$

Here, $\boldsymbol{P}^{\intercal} \in \mathbb{R}^{V\times M}$ and $\boldsymbol{Q}^{\intercal} \in \mathbb{R}^{V\times K}$ are the transpose of the loading matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, $\boldsymbol{T} \in \mathbb{R}^{N\times V}$ and $\boldsymbol{U} \in \mathbb{R}^{N\times V}$ are the score matrices, and $\boldsymbol{E} \in \mathbb{R}^{N\times M}$ and $\boldsymbol{F} \in \mathbb{R}^{N\times K}$ are the residual matrices, minimized in a least squares sense. The loadings summarize the correlations among process variables, while the scores show the relationships among batches based on the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Based on this, the SO-PLS model can be defined. Given blocks ($\boldsymbol{X}^{(1)}$, $\boldsymbol{X}^{(2)}$, ..., $\boldsymbol{X}^{(S)}$), following a pre-defined block sequence which aligns with the flowsheet design of the process, the algorithm starts with the first block $\boldsymbol{X}^{(1)}$ where separate PLS model is fitted with $\boldsymbol{Y}$ as the response. The subsequent block $\boldsymbol{X}^{(2)}$ is then orthogonalized with respect to the scores of the PLS model from the previous block, and so is the response $\boldsymbol{Y}$. The algorithm repeats this process for all blocks in the system. This ensures that only new information not modeled by previous blocks remains in subsequent blocks. The pseudo code is shown in Table 1 (Smilde et al., 2022).

Table 1. Pseudo-code for (two-block) SO-PLS model

First block

1. $\text{PLS}\left(\boldsymbol{X}^{(1)}, \boldsymbol{Y}\right) \implies \boldsymbol{T}^{(1)}, \boldsymbol{P}^{(1)}$
   PLS regression for $\boldsymbol{X}^{(1)}$ *and* $\boldsymbol{Y}$
2. $\boldsymbol{X}^{(2),\text{ort}} = \left(\boldsymbol{I} - \boldsymbol{T}^{(1)}\left(\left(\boldsymbol{T}^{(1)}\right)^{\intercal}\boldsymbol{T}^{(1)}\right)^{-1}\left(\boldsymbol{T}^{(1)}\right)^{\intercal}\right)\boldsymbol{X}^{(2)}$
   Orthogonalization of $\boldsymbol{X}^{(2)}$ *with respect to* $\boldsymbol{T}^{(1)}$
3. $\boldsymbol{Y}^{\text{ort}} = \left(\boldsymbol{I} - \boldsymbol{T}^{(1)}\left(\left(\boldsymbol{T}^{(1)}\right)^{\intercal}\boldsymbol{T}^{(1)}\right)^{-1}\left(\boldsymbol{T}^{(1)}\right)^{\intercal}\right)\boldsymbol{Y}$
   Orthogonalization of $\boldsymbol{Y}$ *with respect to* $\boldsymbol{T}^{(1)}$

Second block

4. $\text{PLS}\left(\boldsymbol{X}^{(2),\text{ort}}, \boldsymbol{Y}^{\text{ort}}\right) \implies \boldsymbol{T}^{(2)}, \boldsymbol{P}^{(2)}$
   PLS regression for $\boldsymbol{X}^{(2),\text{ort}}$ *and* $\boldsymbol{Y}^{\text{ort}}$
5. $\boldsymbol{Y} = \boldsymbol{T}^{(1)}\boldsymbol{Q}^{(1)} + \boldsymbol{T}^{(2)}\boldsymbol{Q}^{(2)} + \boldsymbol{F}$
   Estimation in the least squares sense

SO-PLS offers several advantages. Unlike other multi-block techniques, it allows a different number of components for each block (Westerhuis et al., 1998). The orthogonalization step eliminates redundant information, ensuring additional explained variance from a new block represents new information not captured by previous blocks (Figure 3). However, the sequence of blocks can influence model performance in SO-PLS. Evidence suggests that the block order should align with the process flow diagram's topology (Næs et al., 2021). For parallel unit operations, if they are identical, data blocks are concatenated. If not, data blocks are serialized, and the order is chosen freely (Zhu et al., 2024). The model assumes no recycle streams in the process, but the impact of loops can be evaluated by summarizing the feedback as a variable (e.g., the amount of recycled material) and measuring its effect (van Kollenburg et al., 2021).

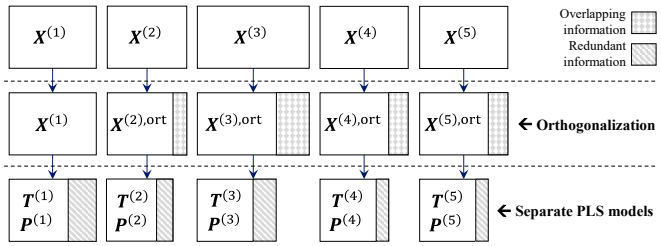

Figure 3. Information overlap handled by SO-PLS, adapted from Zhu et al. (2024).

Stage 2.c: Finally, critical source steps are identified. SO-PLS is useful for this due to its ability to divide the total sum-of-squares into contributions for each block, similar to ANOVA (Smilde et al., 2022). Future work aims to leverage this property to identify the block that contributes most to the output variance. However, calculating degrees of freedom for PLS models is a challenge (van der Voet, 1999). As an alternative, CV-ANOVA (Indahl & Næs, 1998) can be used. For a two-block SO-PLS model, this method compares predicted residuals of a one-block model with those of a two-block model, using a paired t-test to judge significance. This can be easily extended to more blocks. With this method, multiple significant process steps may be identified, requiring further intra-step analyses, as shown in Figure 2.

2.3 Analysis of unit operation model(s) (intra-step level)

After identifying process steps that significantly impact downstream yield, we proceed to the intra-step analysis. This level focuses on examining how manipulated variable trajectories influence the output of the critical process steps. We analyze the entire evolution of batches uncover internal dynamics and variable interactions causing variability downstream.

Stage 3.a: Inter-step analysis identifies process steps needing further examination. Intra-step analysis is unit-specific, requiring separate models for each unit. Therefore, a prioritized list of hypotheses must be created with process experts, based on existing knowledge. For example, if the inter-step analysis shows that the fermentation concentration profile impacts downstream yield, a hypothesis might be to explore which variable trajectories affect the concentration profile, making it the output of interest for the intra-step model. When there are competing critical process steps, one can prioritize based on the number of available hypotheses or choose the step closest to where variability is observed.

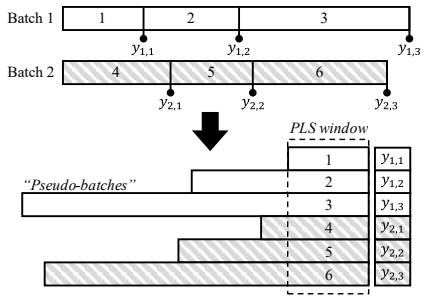

Figure 4. Pseudo-batch method for batch data alignment, adapted from López-Montero et al. (2015).

Stage 3.b: Next, the intra-step data is preprocessed. This data is often in tensor format, challenging to align due to varying sampling frequencies and batch lengths. Multi-way PLS methods can handle such data (Westerhuis et al., 1999) but usually require equal batch lengths. Solutions like truncating data or synchronizing around an indicator variable (Sartori et al., 2023) are unsuitable because they remove crucial end-of-batch data or distort the sampling frequency of the quality measurements, introducing unnecessary complexities. Instead, we use the pseudo-batch technique (López-Montero et al., 2015), as shown in Figure 4. This method unfolds input data from each batch into a vector, treating each time instance as a new variable (Westerhuis et al., 1999). The entire trajectory up to each sampling point is saved as a pseudo-batch. Data is then synchronized by removing excess information from the beginning of trajectories to fit a specific modeling window.

Stage 3.c: With the data transformed, autoscaled, and using the same variable selection procedure as in *Stage 2.a*, a standard PLS model can be fitted to identify the effects between input and output trajectories of the process step. While PLS models are excellent for process exploration, other models, including mechanistic ones, can also be applied if available. Variable trajectories that significantly impact the output are listed based on the VIP index and discussed with process experts. If no immediate decisions can be made after the intra-step analysis, the procedure returns to the inter-step level to investigate another step. This explains the second decision point in Figure 2. If no further steps explain the variance, it may be necessary to exit the workflow and conclude that additional data sources, such as raw materials data, are needed.

3. RESULTS AND DISCUSSION

3.1 Process description

This study focuses on a section of a commercial batch process producing an API for diabetes treatment. The proposed two-level method extends multivariate statistical analysis to include both the production stages of the API (**fermentation (1)**) and the purification steps, which include:

- **Clarification (2):** Removing yeast (and residues) from the broth. *Primary equipment: Tanks and centrifuges.*
- **Concentration (3):** Concentrating the product and reducing host cell proteins. *Primary equipment: Tanks and chromatography.*
- **Reaction (4):** Chemical transformation from precursor to the desired chemical form. *Primary equipment: Tanks.*
- **Precipitation (5):** Concentrating the product and preparing it for storage. *Primary equipment: Tanks, centrifuges, and containers.*

The objective is to model the product yield variability at the end of the multi-step process and evaluate each step's contribution to the yield. For confidentiality, the API, some variable names, absolute data values, and details about unit operations and their connectivity are not disclosed. Each step includes multiple unit operations arranged in parallel or series.

3.2 Drivers of downstream yield variability (inter-step)

The two-level approach is applied to identify drivers of yield variability at the precipitation step. The plant is equipped with online sensors recording process variables every second via the AVEVA PI system. Hourly measurements are extracted for analysis. Data from 693 precipitation batches, originating from 43 fermentation batches, are collected, including temperatures, pressures, flow rates, and controller setpoints. Three fermentation batches used in engineering runs are excluded as outliers. For this study, 62 process variables, selected based on expert input, undergo phase-wise aggregation resulting in 336 variables. A VIP-based selection procedure narrows this to 152 variables. Product samples taken at each production stage are analyzed in laboratories to determine product concentration and yield. Yield data are concatenated with online plant data for each stage. For fermentation, the concentration of the harvested product is used instead of yield. The yield at the precipitation step is used as the output vector $\boldsymbol{y} \in \mathbb{R}^{693\times1}$. To match the row numbers of the fermentation data with the purification data, each fermentation batch is repeated for each subsequent purification batch, resulting in the following set of matrices: $\boldsymbol{X}^{(1)} \in \mathbb{R}^{693\times32}$, $\boldsymbol{X}^{(2)} \in \mathbb{R}^{693\times36}$, $\boldsymbol{X}^{(3)} \in \mathbb{R}^{693\times35}$, $\boldsymbol{X}^{(4)} \in \mathbb{R}^{693\times24}$, and $\boldsymbol{X}^{(5)} \in \mathbb{R}^{693\times25}$. Note, this highlights the presence of dependent batches, which will be discussed later.

The end-process yield, $\boldsymbol{y}$, is estimated using SO-PLS based on the matrices $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}, \boldsymbol{X}^{(4)}$ and $\boldsymbol{X}^{(5)}$ which include yield and process data from prior steps. Model performance is measured by determination coefficients in calibration (R_Y^2) and cross-validation (Q^2). Cross-validation involves splitting the dataset into segments and holding out one segment for validation. The latent variable combination maximizing Q^2 is visualized using a Måge plot (Smilde et al., 2022).

The SO-PLS model explains 78.7% of the variance in the API's yield at precipitation during calibration (R_Y^2) and 70.8% during validation (Q^2). The optimal latent variable combination {4,0,0,0,3} is shown in Figure 5(a).

Table 2. Total explained variance of inter-step SO-PLS is 78.7%. The number of LVs is {4,0,0,0,3}, selected by cross-validation.

	LV 1	LV 2	LV 3	LV 4	Total
$\boldsymbol{X}^{(1)}$	38.8	21.5	6.40	1.90	68.6
$\boldsymbol{X}^{(5)}$	3.84	4.95	1.30		10.1

Table 2 reveals that the most impactful block is $\boldsymbol{X}^{(1)}$, explaining 68.6% of the variance. Specifically, the latent variable {1,0,0,0,0} explains 38.8% of the variance. This variable is mainly driven by an indicator variable that is set to 1 when the concentration profile of an undisclosed metabolite crosses a threshold. Batches where this indicator equals 1 are called *disturbance batches*, and their yield at precipitation is shown in Figure 5(b). A significant difference is confirmed by a Welch's two-sample t-test, with a t-value of -12.15 and 63.9 degrees of freedom, leading to p-value $< 2.2 \cdot 10^{-16}$. Note that in this case study, intermediate steps ($\boldsymbol{X}^{(2)}, \boldsymbol{X}^{(3)}, \boldsymbol{X}^{(4)}$) do

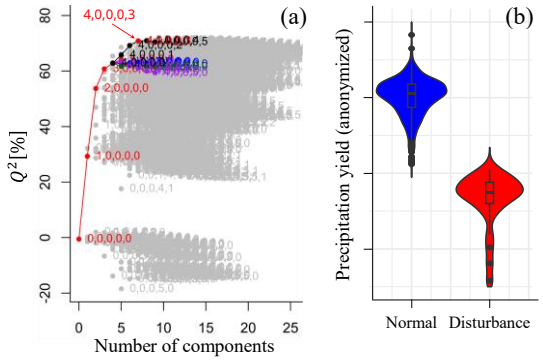

Figure 5. (a) Maximum obtained Q^2 is 70.8%. (b) Yields from the last harvested batch from each fermentation (**red**) compared to prior batches (**blue**).

not significantly contribute to model performance, as indicated by the absence of latent variables.

Thus, lower-performing precipitation batches seem to originate from process disturbances in fermentation. Process experts note that the disturbances are characterized by the production of certain metabolites, which can decrease productivity and biomass formation. We initiate an in-depth intra-step analysis of the fermentation step to investigate these issues.

3.3 Linking metabolite profile to process inputs (intra-step)

To understand the relationship between input dynamics of the fermentors and the process disturbance, we study the correlation between online process variables ($\widetilde{\boldsymbol{X}}$) and metabolite concentration profile ($\widetilde{\boldsymbol{y}}$). For the intra-step analysis, hourly online process variables are extracted, including pH, temperature, flow rate controllers (glucose, biomass, NH_3), and measurements of oxygen, ethanol, and CO_2. Data from 43 fermentation batches are collected – three removed as outliers. Unlike the inter-step analysis, no aggregation is performed, and the entire trajectory is kept. In total, we have 40 features, including transformations of existing inputs (e.g., dosage ratios).

In the fermentors, cells are grown in fed-batch mode until the desired tank weight and biomass concentration are reached. Then, the fermentors continue in continuous cultivation mode, delivering material to harvest tanks. Process variables are split into these two phases and autoscaled with respect to the mean and variance of each phase. Product samples are taken during the batch evolution, though less frequently than the online measurements. Metabolite concentration is determined from the chromatograms of product samples and autoscaled according to the mean and variance of the sample trajectories.

We use the pseudo-batch method and build a PLS model to estimate $\widetilde{\boldsymbol{y}}$ from $\widetilde{\boldsymbol{X}}$. The input data is unfolded, and the pseudo-batch transformation is applied, synchronizing the pseudo-batches to the average 12-hour time window between product samples. Testing minimum (9 hours) and maximum (16 hours) time windows showed no significant model differences. After the pseudo-batch transformation, we obtain a single column of metabolite concentration $\widetilde{\boldsymbol{y}} \in \mathbb{R}^{211\times 1}$. Due to non-linearities, the metabolite concentration is log-transformed. Following the VIP-based variable selection procedure, we obtain an input matrix of size $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{211\times 109}$.

Table 3 shows the performance of the PLS model. Latent variables are chosen using segment-based cross-validation, where entire batches – rather than individual pseudo-batches – are excluded for validation to preserve the inherent dependencies among pseudo-batches. The optimal PLS model employs 6 latent variables and explains 82.5% in calibration and 71.4% in validation. The low difference between R_Y^2 and Q^2 indicates no overfitting.

Table 3. Cumulative explained variance for PLS with 6 LVs, selected by cross-validation.

	R_X^2 [%]	R_Y^2 [%]	Q^2 [%]
LV 1	35.5	24.6	17.1
LV 2	47.9	41.1	26.5
LV 3	61.0	50.2	35.2
LV 4	73.5	70.8	52.8
LV 5	77.0	76.5	67.1
LV 6	83.7	82.5	71.4

The PLS model provides insights into the relationship between process variables and metabolite concentration. Figure 6 highlights variable importance using the VIP index, with a threshold of 1 to determine significance. Variables are as follows: **1)** *fermentor weight,* **2)** *tank agitator power,* **3)** *NH_3 flow controller,* **4)** *transfer flow controller,* **5)** *fermentor pH,* **6)** *dilution rate,* **7)** *glucose flow rate,* **8)** *measured CO_2 concentration,* **9)** *glucose dosage stop duration, and* **10)** *dosage ratio between NH_3 and glucose*. Significant variables include dilution rate, measured CO_2 concentration, and dosage ratio between NH_3 and glucose. Given the direct link between dilution rate and glucose feed, and the fact that CO_2 concentration is a response variable, we focus on the dosage ratio.

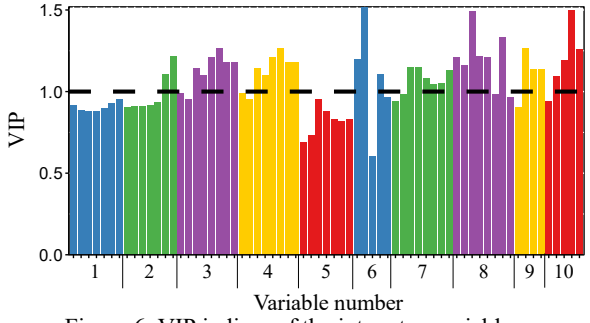

Figure 6. VIP indices of the intra-step variables

As seen in Figure 7, a low dosage ratio between NH_3 and glucose correlates with high metabolite concentration (which is also reflected by high CO_2 levels). To mitigate disturbances affecting productivity and yield, it has been discussed with process experts to maintain the dosage ratio above 0.19 (standardized value).

4. CONCLUSION

This paper introduces a novel methodology to identify drivers of yield variability in pharmaceutical batch processes. Using multivariate statistical methods, it suggests adjusting manipulated variable trajectories to mitigate systematic variability. The industrial case study demonstrates that the SO-PLS model highlights the fermentation step as crucial for the API's downstream yield in a five-step process. While other steps contribute, their impact is less significant. Identifying the most influential step allows for ideal variable trajectories to resolve yield variability, recommending maintaining the dosage ratio of NH_3 and glucose above a lower limit. This knowledge can enhance process performance and enables experts to focus on specific steps and variables. The methodology's two-level nature allows expansion with development scale data for deeper

insights into each step's dynamics. Though currently applied to yield variability, it can also address quality-related issues. Future work should include a comparison between multiblock methods like MB-PLS, SO-PLS, SMB-PLS, and process-PLS (Lauzon-Gauthier et al., 2018; van Kollenburg et al., 2021). Additionally, the optimal block order of the SO-PLS model requires further investigations (Næs et al., 2021).

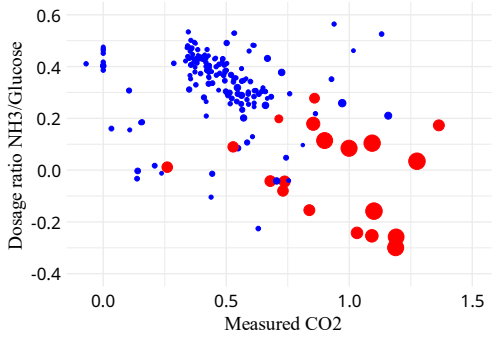

Figure 7. Scatter plot of (autoscaled) dosage ratio and measured CO_2. **Red** indicates high (log) metabolite concentrations.

Lastly, given that multiple downstream batches can share the same upstream batch, observations are inherently dependent. While PLS does not require data independence, the current approach overlooks the group structure of batches, mixing between- and within-group variances. A potential solution is to use a multi-group PLS approach (Eslami et al., 2014).

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from Novo Nordisk A/S and extend their thanks for providing the data used in the industrial application.

REFERENCES

Barton, M., Duran-Villalobos, C.A., Lennox, B. (2021). Multivariate batch to batch optimisation of fermentation processes to improve productivity. *J. Process Control*, 108, 148-156.

Bisgaard, S., Kulahci, M. (2007). Quality quandaries: using a time series model for process adjustment and control. *Qual. Eng.,* 20(1), 134–141.

Brás, L.P., Lopes, J.A., Santos, C.R., Cardoso, J.P., Menezes, J.C. (2004). Modelling and identification of individual stage contributions in an industrial pharma-ceutical process by multiblock PLS. In A. Barbosa-Póvoa, H. Matos (eds.), *Comput. Aided Chem. Eng.*, pp. 601-606. Elsevier.

Cattaldo, M., Ferrer, A., Måge, I. (2024). Dynamic multiblock regression for process modelling. *J. Chemometrics*, 38, e3618.

Diab, S., Christodoulou, C., Taylor, G., Rushworth, P. (2022). Mathematical modeling and optimization to inform impurity control in an industrial active phar-maceutical ingredient manufacturing process. *Org. Process Res. Dev.*, 26(10), 2864–2881

Eslami, A., Qannari, E.M., Kohler, A. Bougeard, S. (2014), Algorithms for multi-group PLS. *J. Chemome-trics*, 28: 192-201.

Indahl, U.G., Næs, T. (1998). Evaluation of alternative spectral feature extraction methods of textural images for multivariate modelling. *J. Chemometrics*, 12, 261-278.

Korovessi, E., Linninger, A.A. (Eds.). (2005). Batch processes. Taylor & Francis.

Lauzon-Gauthier, J., Manolescu, P., Duchesne, C. (2018). The sequential multi-block PLS algorithm (SMB-PLS): Comparison of performance and interpretability, *Chemom. Intell. Lab. Syst.,* 180, 72-83

López-Montero, E.B., Wan, J., Marjanovic, O. (2015). Trajectory tracking of batch product quality using intermittent measurements and moving window estimation. *J. Process Control*, 25, 115-128.

Marschall, L., Taylor, C., Zahel, T., Kunzelmann, M., Wiedenmann, A., Presser, B., Studts, J., Herwig, C. (2022). Specification-driven acceptance criteria for validation of biopharmaceutical processes. *Front. Bioeng. Biotechnol.*, 10, 1010583.

Næs, T., Romano, R., Tomic, O., Måge, I., Smilde, A., Liland, K.H. (2021). Sequential and orthogonalized PLS (SO-PLS) regression for path analysis: Order of blocks and relations between effects. *J. Chemometrics*, 35(10), e3243.

Rendall, R., Lu, B., Castillo, I., Chin, S.-T., Chiang, L. H., Reis, M. S. (2017). A unifying and integrated framework for feature oriented analysis of batch processes. *Ind. Eng. Chem. Res.*, 56(30), 8590-8605.

Sartori, F., Facco, P., Zuecco, F., Bezzo, F., Barolo, M. (2023). Optimal indicator-variable approach for trajec-tory synchronization in uneven-length multi-phase batch processes. *Ind. Eng. Chem. Res.*, 62 (44), 18511-18525.

Smilde, A.K., Næs, T., Liland, K.H. (2022). Multiblock data fusion in statistics and machine learning: Applica-tions in the natural and life sciences. John Wiley & Sons.

van der Voet, H. (1999). Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. *J. Chemometrics,* 13(3/4), 195-208.

van Kollenburg, G., Bouman, R., Offermans, T., Gerretzen, J., Buydens, L., van Manen, H.-J., Jansen, J. (2021). Process PLS: Incorporating substantive knowledge into the predictive modelling of multiblock, multistep, multi-dimensional and multicollinear process data. *Comput. Chem. Eng.,* 154, 107466.

Westerhuis, J.A., Kourti, T., MacGregor, J.F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemometrics,* 12, 301-321.

Westerhuis, J.A., Kourti, T., MacGregor, J.F. (1999). Comparing alternative approaches for multivariate statistical analysis of batch process data. *J. Chemometrics*, 13(3-4), 397-413.

Wold, S., Sjöström, M., Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.,* 58(2), 109-130.

Zahel, T., Hauer, S., Mueller, E. M., Murphy, P., Abad, S., Vasilieva, E., Maurer, D., Brocard, C., Reinisch D., Sagmeister P., Herwig, C. (2017). Integrated process modeling – a process validation life cycle companion. *Bioeng.*, 4, 86.

Zhu, Q., Facco, P., Zhao, Z., Barolo, M. (2024). Capturing connectivity information from process flow diagrams by sequential-orthogonalized PLS to improve soft-sensor performance. *Chemom. Intell. Lab. Syst.,* 252, 105192.