Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process

DOI:
10.1016/j.compchemeng.2019.05.037

Document Version
Accepted author manuscript

Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process

Stephen Goldrick[a], Carlos A. Duran-Villalobos[a], Karolis Jankauskas[a], David Lovett[c], Suzanne S. Farid[a] and Barry Lennox[b]

[a] The Advanced Centre of Biochemical Engineering, Department of Biochemical Engineering, University College London, Gordon Street, London, United Kingdom

[b] School of Electrical and Electronic Engineering, University of Manchester, Manchester, United Kingdom

[c] Perceptive Engineering Limited, Cheshire, United Kingdom

Abstract

This paper outlines real-world control challenges faced by modern-day biopharmaceutical facilities through the extension of a previously developed industrial-scale penicillin fermentation simulation (*IndPenSim*). The extensions include the addition of a simulated Raman spectroscopy device for the purpose of developing, evaluating and implementation of advanced and innovative control solutions applicable to biotechnology facilities. *IndPenSim* can be operated in fixed or operator controlled mode and generates all the available on-line, off-line and Raman spectra for each batch. The capabilities of *IndPenSim* were initially demonstrated through the implementation of a QbD methodology utilising the three stages of the PAT framework. Furthermore, *IndPenSim* evaluated a fault detection algorithm to detect process faults occurring on different batches recorded throughout a yearly campaign. The simulator and all data presented here are available to download at www.industrialpenicillinsimulation.com and acts as a benchmark for researchers to analyse, improve and optimise the current control strategy implemented on this facility. Additionally, a highly valuable data resource containing 100 batches with all available process and Raman spectroscopy measurements is freely available to download. This data is highly suitable for the development of big data analytics, machine learning (ML) or artificial intelligence (AI) algorithms applicable to the biopharmaceutical industry.

Highlights (Max 85 Characters per bullet)

1. Benchmark simulator for researchers to compare and validate novel controllers
2. Outline of control challenges applicable to modern day biopharmaceutical facilities
3. Development of a Raman spectroscopy simulation applicable for advanced controller design
4. Implementation of a Quality by Design approach enabling process optimisation
5. Comparison of fault detection algorithms for process fault identification

Keywords: Modelling, Control, Process Analytic Technology (PAT), Quality by Design (QbD), biopharmaceutical, Raman spectroscopy, fault detection

1. Introduction

Penicillin fermentation monitoring and control has been carried for the last three decades (Mou and Cooney, 1983, Min et al. 1995 and Lee et al. 2004, Luo and Bao, 2018). However the biopharmaceutical sector as a whole is still significantly lagging behind other sectors in their adoption of advanced process control (APC), particularly in their use of innovative process analytical technology (PAT) solutions (Tomba et al. 2013). This is more evident in comparison to other highly sophisticated industries such as oil & gas, semi-conductor and automotive where automation and lean manufacturing are better engrained into company practice and culture. A major push from industrial regulators to rectify this has been the implementation of the Quality by Design (QbD) and PAT initiatives set out by the FDA in 2004 and 2009, respectively (FDA 2004, FDA 2009). However, a major challenge remaining is the expertise and confidence required to adopt and implement these novel control solutions throughout industrial biopharmaceutical processes. Over the last 25 years the development of first principles mathematical models mimicking complex industrial processes have aided in the development and deployment of APC solutions (Downs and Vogel, 1993; Lyman and Georgakis, 1995; Birol et al. 2002; Jeppsson et al. 2007; Kontoravdi et al. 2010; Kiparissides et al. 2011; Benyahia et al. 2012; Gernaey and Gani, 2010; Goldrick et al. 2014; Papadakis et al 2018). The ability to test and validate a novel control strategy on a simulation subsequent to implementation on a real process has the potential to revolutionise control theory and applications of advanced controllers throughout the biopharmaceutical sector (Randek, J., & Mandenius 2018). A limitation of current biopharmaceutical mathematical models is their inability to address the current control challenges of a modern-day biopharmaceutical facility. In the future era of Industry 4.0, which envisions a highly intelligent data-driven manufacturing environment incorporating a multitude of advanced on-line process analytics (Sami Sivri M., Oztaysi 2018), the need for a modern-day biopharmaceutical simulation is paramount.

The simulation described in this paper aims to address the current and future challenges of biopharmaceutical process manufacturing through the extension of a highly complex industrial-scale penicillin fermentation, referred to as *IndPenSim*. The simulation was developed using the historical batch records of a 100,000 litre penicillin fermentation utilising a high-yielding industrial strain of *Penicillium chrysogenum* and accurately simulates all the available process inputs and outputs (Goldrick et al. 2015). *IndPenSim* can be operated in multiple modes enabling the generation of large volumes of realistic fermentation data. The simulation mimics a real process through its ability to include delays in off-line assay measurements, manual operator intervention of feeding strategies, inaccurate sensor readings and random deviations in growth and production levels. Furthermore, a realistic Raman spectroscopy device has been integrated within *IndPenSim*. The inclusion of this device aims to support the current and future development of innovative and advanced control strategies on biopharmaceutical facilities. Furthermore, a data set containing 100 batches (~ 2.5 GB) is available to

download at www.industrialpenicillinsimulation.com which aims to act as a valuable resource for big-data analytics, machine learning (ML) and artificial intelligence (AI) approaches.

Overview of *IndPenSim*

IndPenSim acts as a standalone application (freely available to download at www.industrialpenicillinsimulation.com). A summary of all the process inputs and outputs recorded by *IndPenSim* are shown in **Fig**. 1. Table 1 outlines the measurement frequency and primary control strategy of these main process variables in addition to the functional relationship between each variable. Automatically controlled variables; i.e. temperature (T) and pH (pH), are regulated using a feed-back proportional integral derivative (PID) loop. Manually controlled variables; i.e. substrate flowrate (F_s) and phenylacetic acid flowrate (F_{PAA}), are manipulated using a recipe driven approach which follows a fixed profile throughout the batch (Recipe driven) or are controlled by an operator that manipulates this fixed profile throughout the batch (Operator dependant). This mode of control replicates the observed control actions of the operators manually adjusting F_s and F_{paa} throughout the batch as described in Goldrick et al. (2015). The batch length can be fixed to a constant value (Fixed), typically 230 hours or dependent on delays in downstream process operation (Variable). A summary of a five-year campaign outlining the annual production metrics generated by *IndPenSim* is summarised in Table 2. Each campaign was operated in a different mode and no advanced control algorithms were implemented during any campaign. *IndPenSim* calculates the annual production metrics using the assumption that the facility has a 24-hour operating period and operates 336 days per year. The remaining 29 days are used for an annual shut-down period, allowing for routine maintenance activities to be carried out. A three-day turn around period for bioreactor cleaning and re-inoculation is required following each batch. A target production yield of 2000 kg of penicillin is required in each batch. Any batches achieving yields below this specification are considered below target batches and an investigation into their poor performance is required.

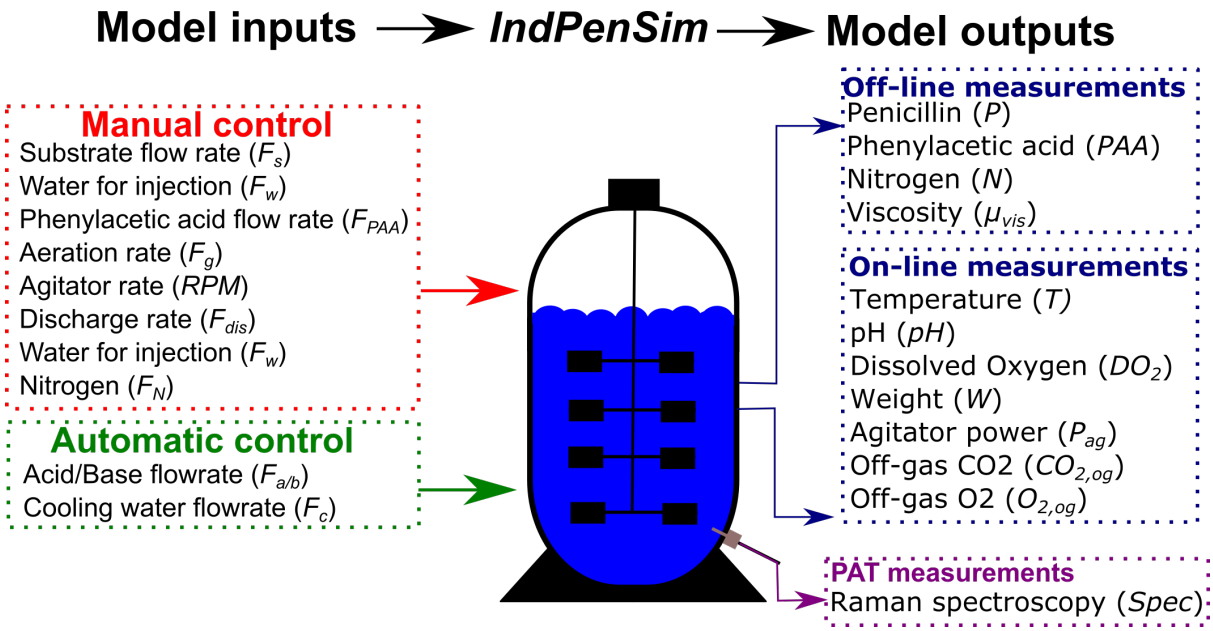

Figure 1. Summary of all model inputs and outputs recorded by *IndPenSim*. Automatic control is dependent on PID control loops whereas manual control is a recipe-driven approach maintaining a fixed profile throughout the batch which can be manually adjusted by operator intervention.

Variable reference	Measurement frequency	Primary control variables	Functional relationship	Control strategy
Dissolved oxygen (DO_2 - $mg\ L^{-1}$)	12 min	F_g,RPM	$Pressure$,$O_{2,og}$,$Viscosity$,T,V,F_{oil}	>10% of saturation
Weight (W - kg)	12 min	F_{water},F_s,F_{alb},F_{PAA},F_{dis}	P,X,V	Maintain between 7×10^{4} and 11×10^{4} kg
pH (pH)	12 min	F_{alb}	P;X;V	PID control algorithm
Temperature (T - K)	12 min	F_c	P,X,V	PID control algorithm
Off-gas measurements ($CO_{2,og}$ & $O_{2,og}$ - %)	12 min	F_g,RPM	O_2, $CO2$	Not controlled
Penicillin (P - $g\ L^{-1}$)	12 h (+ 4 delay)	F_S,F_{oil},F_{PAA},F_N	X,PAA,DO_2,S	Maximise production
Biomass (X - $g\ L^{-1}$)	12 h (+ 4 delay)	F_S,F_{oil},F_{PAA},F_N	P,PAA,N,S,pH,T,CO_2	Maximise production
Phenylacetic acid (PAA - $mg\ L^{-1}$)	12 h (+ 4 delay)	F_{PAA}	P,X,V	Maintain between 600 and 1800 mg L^{-1}
Nitrogen (N - $mg\ L^{-1}$)	12 h (+ 4 delay)	N_{shots},F_{oil},F_{PAA}	P,X,V	Maintain above 300 mg L^{-1}
Viscosity (μ - cP)	12 h (+ 4 delay)	F_{water}	P,X,V	Maintain below 100 cP
Substrate (S - $g\ L^{-1}$)	No off-line measurements available	F_s,F_{oil}	P,X,V	Maintain between 5×10^{-3} and 1×10^{-3} g L^{-1}

Table 1. Summary of measurement frequency, primary control variables, functional relationships and control strategies for recorded process variables.

Campaign summary	Campaign 1 (Year 1)	Campaign 2 (Year 2)	Campaign 3 (Year 3)	Campaign 4 (Year 4)	Campaign 5 (Year 5)
Control strategy	Operator dependant	Recipe driven	Operator dependant	Recipe driven	Operator dependant
Fixed or variable batch length	Fixed	Variable	Variable	Fixed	Fixed
Average batch length (hours)	230 ± 0	239 ± 27	239 ± 32	230 ± 0	230 ± 0
Number of batches	26	25	26	26	26
Number of below target batches	2	8	6	2	5
Average Penicillin yield per batch (kg)	2882 ± 745	2578 ± 769	2950 ± 888	2912 ± 786	2816 ± 796
Annual production ($kg \times 10^3$)	74939	64458	76690	75716	73228

Table 2. A summary of the annual production metrics recorded by *IndPenSim* operated using different control strategies throughout a five-year production period.

1.2 *IndPenSim* control objectives:

IndPenSim considers the growth, morphology, metabolic production and degeneration of a large-scale *Penicillium chrysogenum* fermentation in addition to modelling all the required on-line and off-line variables. The details regarding the mathematical structure of the model have been previous described in Goldrick et al. (2015). The primary focus of this paper is to demonstrate the ability of this simulation to act as a benchmark for the development and validation of novel control solutions applicable to biopharmaceutical processes. Currently this fermentation process has no advanced process control strategies in place and therefore presents significant process improvement opportunities. The primary goal of any control strategy is to ensure an economically viable process through increased product yields and reduced operating costs (Montague et al. 1989), therefore the following control objectives have been defined:

- Develop a control strategy to maximise annual penicillin production and reduce variation in batch yields in comparison to the five campaigns outlined in Table 2.
- Identify the critical process parameters (CPPs) and critical quality attributes (CQAs) influencing penicillin production.
- Develop an enhanced control strategy for pH and temperature variables to minimise their fluctuations in comparison to the existing PID control loops.
- Develop a control strategy that manipulates one or more of the following flowrates: substrate, nitrogen or phenylacetic acid, to maintain these variables within their acceptable ranges defined in Table 1.
- Utilise the spectra recorded by the Raman spectroscopy device to develop a soft-sensor enabling an on-line prediction of phenylacetic acid, biomass or penicillin concentration in real-time.
- Develop a control strategy that calculates the optimum harvest time for each batch to maximise annual penicillin yields generated throughout a yearly campaign.

2. Material and Materials

2.1 Simulation software

IndPenSim was written in Matlab R2018b and is freely available to download at www.industrialpenicillinsimulation.com where the historical batch records of campaigns 1-5 outlined in Table 2 are also available. *IndPenSim* has the following capabilities and functionality:

- Batch to batch variation of both the biomass and penicillin concentration as well as in-batch fluctuations

- Option to add disturbances on inlet concentrations of the substrate (c_s), oil (c_{oil}), acid/base molar concentration ($c_{a/b}$) and Phenylacetic acid concentration (c_{PAA}).
- Ability to adjust the current sequential batch control strategy for F_s, F_{oil}, F_g, RPM, F_{dis} and F_{PAA}.
- Option to include inhibition effects on the growth rates during DO_2, N and PAA limitation as well as during excessive PAA and CO_2 concentrations and sub-optimal T and pH operation.
- Includes a pre-defined delay (4 h) in the off-line measurements of P, N, PAA and μ_{app}.
- Option to include process faults including agitator trip, aeration faults, substrate faults and sensor errors.
- Option to record Raman spectra throughout the batch, enabling real-time predictions of the critical quality attributes and critical process parameters provided an accurate calibration model is developed and the spectra is pre-processed correctly.

2.2 Raman Spectroscopy simulation development

This section describes the development of an empirical mathematical model to simulate a realistic PAT analyser, specifically a Raman spectroscopy device. The simulated spectra were generated and validated through a detailed analysis of experimental Raman spectra recorded on a 5 litre fungal fermentation producing a commercially available antibiotic. Further details describing the materials and methods of this fermentation are outlined in Goldrick et al. (2018). The Raman spectroscopy device used was a Kaiser 1000 RXN system implementing an indium gallium arsenide (InGaAs) detector array with a spectral range of 200–2400 cm^{-1} and a resolution of 3 cm^{-1}. The Raman spectroscopy analyser was set-up to record a spectrum every 30 minutes based on 9 averages using an integration time of 180 seconds. In total 540 spectra were recorded throughout the 260-hour fermentation, highlighted in **Fig. 2A**. The simulated PAT analyser described here aims to mimic the three main characteristics that define this experimentally recorded Raman spectra. These are outlined by Bocklitz et al. (2011) as fluorescence baseline increase, Raman spectrum peaks and noise. The modelling of random cosmic spikes on Raman spectroscopy was not considered in this work.

2.2.1 Non-linear spectra profile and baseline increase

Raman spectra recorded on fermentation systems contain characteristic peaks related to media components and cell culture in addition to the characteristic non-linear shape associated with the background signal of the Raman spectroscopy device. This was modelled by taking the first spectrum of the experimental Raman data set and using this as a template for all spectra generated by this simulated PAT analyser, the reference spectrum is shown in **Fig. 2B**. The fluorescence increase shown in the experimental Raman spectra is visible in **Fig. 2A** where the baseline intensity of the spectra

collected during the start (0-45 hours) can be compared with that collected at the end of the fermentation (215-240 hours). To model this fluorescence increase ($\Delta_{Fluorescence_{Exp}}$) in the experimental spectra the average change in intensity from one spectrum to the next was calculated as:

$$\Delta_{Fluorescence_{Exp}}(n) = \sum_{v=250}^{v=2250} \frac{(Spectra(n+1) - Spectra(n))}{2000} \qquad \text{Eq. 1}$$

Where $\Delta_{Fluorescence_{Exp}}$ represents the average change in baseline intensity of two consecutive spectra between the wavelengths (v) 250-2250 cm^{-1}. Taking the cumulative sum of the calculated fluorescence increase results in an average fluorescence profile of the fermentation. In these empirically simulated Raman spectra, the fluorescence increase ($\Delta_{Fluorescence_{Sim}}$) was assumed to be the result of compositional changes to the fermentation broth. The compositional changes assumed to have the largest influences were the biomass (X), penicillin (P), viscosity (μ) and batch time (t), which are defined as:

$$\sum_{t=0}^{t=240} \Delta_{Fluorescence_{Sim}} = \alpha_1 X + \alpha_2 P + \alpha_3 \mu + \alpha_4 t \qquad \text{Eq. 2}$$

The coefficients ($\alpha_{1,2,3,4}$) were calculated using a step-wise linear regression function that minimised the error between the calculated experimental fluorescence increase and the simulated florescence. The fluorescence increase was found to be accurately modelled by these four variables with the product concentration identified as having the largest influence on the experimentally recorded fluorescence. The finalised coefficients ($\alpha_{1,2,3,4}$) were equal to -0.002 (X), 1.05 (P), -0.07 (μ) and -0.2 (t). It was observed in **Fig. 2A** that fluorescence had a greater influence on the lower wavelengths in comparison to the higher wavelengths. To account for this nonlinearity an exponential function was multiplied by each spectrum to mimic this as shown in Eq. 5. This exponential function is defined in this work as β, further details can be found in Goldrick (2015).

2.2.2 Non-linear characteristic peak increase related to fermentation composition

The simulated Raman spectra needs to take into account the characteristic peaks related to changes in component concentrations throughout the batch. Previous work on the use of Raman spectroscopy for on-line monitoring of biological processes has simulated these characteristic peaks as Gaussian functions (Oh et al., 2012). Furthermore, Gauassian functions have also been demonstrated to represent

specific molecules in chemical analysis utilising Raman spectroscopy (Kneipp et al., 1999). Therefore, Gaussian functions were used to represent the substrate (*S*), penicillin (*P*) and phenylacetic acid (*PAA*) concentrations in this simulation. The position of the substrate and phenylacetic acid peaks were selected based on analysis of the Raman spectra containing media spiked with high concentrations of phenylacetic acid and substrate as outlined in Goldrick (2015). The penicillin peak positions were chosen based on Raman spectra of Penicillin G samples shown in Clarke et al. (2005). These peaks were represented by a Gaussian distribution function defined as:

$$f(Peak_{(P/S/PAA)}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Peak(P/S/PAA) - \mu_{P/S/PAA}\right)^2}{2\sigma^2}} \qquad \text{Eq. 3}$$

Where $Peak_{(P/S/PAA)}$ is the specific wavelength related to either changes in penicillin (*P*), substrate (*S*) or phenylacetic acid (*PAA*), σ is the standard deviation of either *Peak(P)* or *Peak(S)* or *Peak(PAA)* and μ represents the peak mean. These component peaks are shown in **Fig. 2C**.

2.2.3 Signal-to-noise ratio

Noise is an inherent disturbance to any sensor. For Raman spectroscopy noise generally results from thermal effects, instrument read-out errors or random cosmic rays. The magnitude of the noise was modelled by calculating the signal-to-noise ratio (SNR) of the spectra. The SNR assumes the Raman spectra collected in close succession to each other should be almost identical with the main difference between these two signals being the result of noise within the signals (Grimbergen et al., 2010). By calculating the mean and standard deviation of each consecutive spectra the SNR is calculated as follows:

$$\text{SNR} = \frac{\bar{S}}{\sigma_{diff}} \qquad \text{Eq. 4}$$

Where $\bar{S}$ is the mean Raman intensity and σ_{diff} is the standard deviation of spectrum divided by $\sqrt{2}$. The SNR was calculated for 10 spectra and equalled 50 counts (intensity). The magnitude of this was used to add noise to each individual spectrum based on a random walk noise generation. A typical example of the noise added to each spectra is shown in **Fig. 2D**.

The final simulated spectrum (*Sim. Spectra*) is summarised as:

$$Sim.Spectra = Reference\ Spectra + (\delta_1 \Delta_{\text{Fluorescence}} + \delta_2 Peaks(S,P,PAA) + \delta_3 Noise) \times \beta$$

Eq. 5

Where the $\delta_{1,2,3}$ are coefficients related to the intensity of each characteristic component of the simulated spectra. $\delta_1$is the fluorescence increase due compositional changes in biomass, penicillin,

viscosity and also batch time. δ_2 is related to the intensity increase based off current concentrations of penicillin (P), substrate (S) and phenylacetic acid (PAA) in the bioreactor and δ_3 is the intensity associated with the noise added to each spectrum. β relates to the exponential function to account for non-linear increase of the lower wavelengths of the spectra.

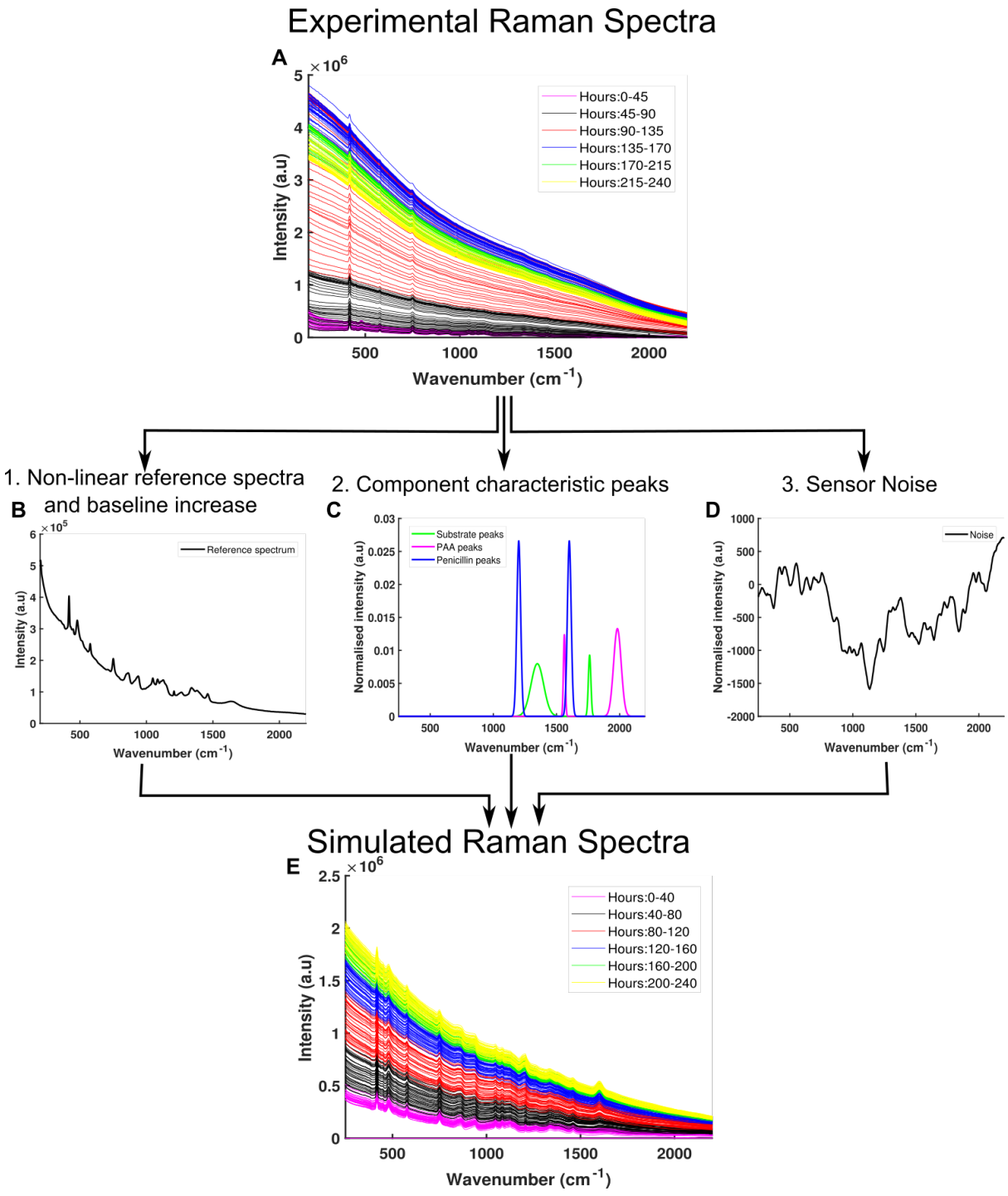

Figure 2: A summary outlining the development of the Raman spectroscopy simulation. **A)** Highlights the experimental spectra recorded by a 993 nm Raman spectroscopy. **B)** Highlights the non-linear reference spectrum implemented as the starting spectrum in this simulation. **C)** Highlights the non-linear characteristic peak increase related to fermentation compositional changes in penicillin, substrate and phenylacetic acid. **D)** Highlights an example of the typical noise added to each simulated spectrum. **E)** Shows an example of the simulated spectra developed in this work.

A demonstration of the robustness of the simulated Raman spectra to accurately predict the penicillin concentration during routine and abnormal operation is demonstrated in **Fig. 3**. In total, four batches were simulated that contained a low filtered pseudo random binary signal (PRBS) added to the substrate feed rate (F_s) to mimic realistic process deviations as shown in **Fig. 3A**. The first batch was used to build the PLS model taking the interpolated off-line penicillin concentration as the response. The spectra was pre-processed as described in Section 4. The PLS model selected four latent variables as optimum, accounting for 99% of the variance in the X-data (spectral data) and the 98% of variance in the Y-data matrix (interpolated penicillin concentration). A calibration batch was simulated and resulted in highly comparable predictions of the off-line penicillin concentration with the root mean square error (RMSE) equal to +/- 0.1 g L^{-1}. Two addition batches were simulated containing a process disturbance in the substrate flow rate (F_s) as demonstrated in **Fig 3A**. The resultant drop in penicillin concentration as a result of these process disturbances is evident from **Fig. 3B**. The PLS predictions of penicillin during these process disturbances is highly comparable with the off-line penicillin concentration measurements shown in **Fig 3B**. The ability of the spectra to be utilised as a real-time measurement of penicillin during normal and abnormal operation represents a significant opportunity to develop and implement advanced process control algorithms on this benchmark simulation. However, it must be noted that the simulated Raman spectroscopy was built using spectra collected at the 5 L scale and does not account for any potential process heterogeneities or additional process issues that may be present at the 100,000 scale.

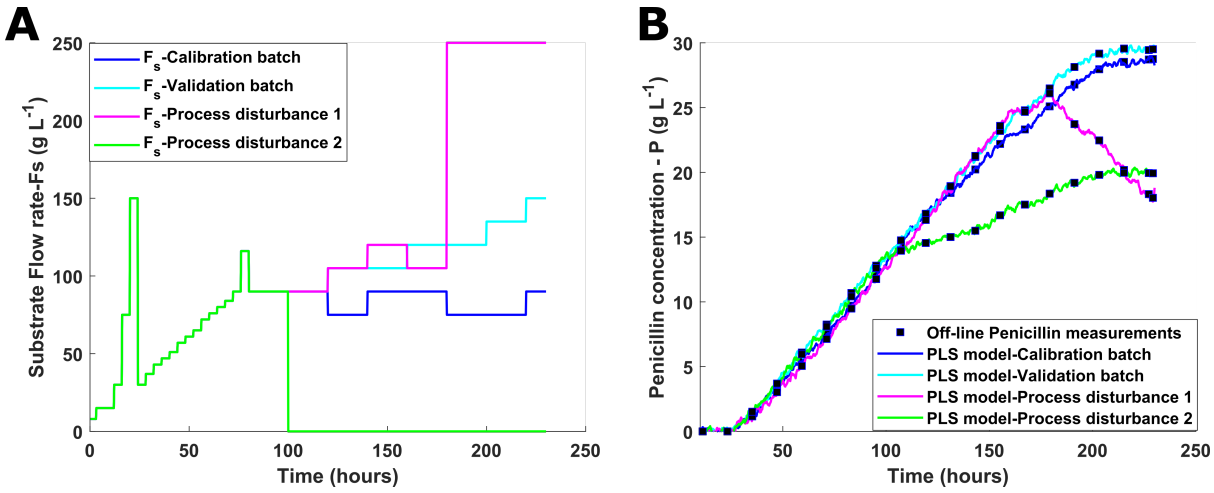

Figure 3: **A)** Outlines the substrate flow rate (F_s) for the calibration and validation batch used to generate the PLS model for predicting the penicillin in addition to two batches containing process disturbances. **B)** Represents the off-line penicillin concentration of the four batches compared to the on-line PLS penicillin predictions.

3. Theoretical section

The following section outlines the multivariate data analysis (MVDA) techniques applied in this manuscript. A batch-wise unfolding algorithm was initially implemented to decompose the data set into a structured format enabling the main sources of variation between each batch to be highlighted (Nomikos and Macgregor 1995). Both principal component analysis (PCA) and partial least squares (PLS) were implemented to reduce the high dimensionality of this large unfolded data allowing for easier data interpretation and better visualisation of hidden correlations. These two techniques have been demonstrated extensively in the monitoring and control of industrial fermentation systems (Lennox et al. 2001, Ündey et al. 2003, Kourti et al. 2005, Chiang et al. 2006, Goldrick et al. 2017).

3.1 Principal component analysis (PCA): On-line and off-line monitoring

The application of PCA for the on-line and off-line monitoring of industrial biopharmaceutical data is well described by Gunther et al. (2007). In summary, prior to applying PCA the data was mean centred and scaled to unit variance. PCA is described mathematically as:

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \, \mathbf{p}_r' + \mathbf{E} \qquad \text{Eq. 6}$$

Where $\mathbf{X}$ represents the two-dimensional data set and $\mathbf{t}_r$, $\mathbf{p}_r$ and $\mathbf{E}$ represent scores, loadings and residuals, when R principal components are retained. The scores ($\mathbf{t}$ vector) represents a single batch and can quantify the overall variability of each batch analysed by the PCA model. The loadings ($\mathbf{p}$ vector) represents the time-series variance of each variable in comparison to the average trajectory of each variable considering all batches in the PCA model. PCA is a well suited and established method to compare new batches to previously recorded normal operating conditions (NOC) batches. The

comparability is defined by calculating the new batch scores ($\mathbf{t}_{new}$) by projecting the new batch data (x_{new}) onto the previously generated PCA model generated using the NOC batches:

$$\mathbf{t}_{new} = \mathbf{x}_{new}\mathbf{P}(\mathbf{P}'\mathbf{P})^{-1} \qquad \text{Eq. 7}$$

This generated score enables off-line monitoring of the newly generated batches. To help identify any abnormal operation two statistical metrics are typically used. The first is the Hotelling T^2 statistic that captures the difference in the systematic part of the model and is calculated as:

$$T^2_{new} = \mathbf{t}_{new}\left(\frac{\mathbf{T}'\mathbf{T}}{I-1}\right)\mathbf{t}'_{new} \qquad \text{Eq. 8}$$

Where I is the number of NOC batches used to generate the PCA model. Any batches that behave abnormally can be detected through analysis of the confidence limit of the T^2_α defined by Lee et al (2004b) as:

$$T^2_\alpha = \left(\frac{R(I\ -1}{I-R}\right)F_{R,I-R,\alpha} \quad \text{Eq. 9}$$

Where $F_{R,I-R,\alpha}$ is the F-distribution assuming a confidence limit equal to α taking R principal components and using I batches to build the model. A second method to detect abnormal behaviour is to analyse the residual error of the PCA model, this is quantified by the sum of squared residuals (SPE) or Q statistic:

$$Q_{new} = \mathbf{e}_{new}\mathbf{e}'_{new} \quad \text{Eq. 10}$$

$$e_{new} = x_{new} - t_{new}P' \quad \text{Eq. 11}$$

Typically these residuals follow a chi squared distribution (χ^2) with a confidence limit approximated by Jackson and Mudholkar (1979) as:

$$Q_\alpha = \theta_1\left(\frac{z_\alpha(2\theta_2 h_0^2)^{0.5}}{\theta_1} + 1 + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2}\right)^{\frac{1}{h_0}} \quad \text{Eq. 12}$$

$$V = \frac{\mathbf{E}\mathbf{E}'}{I-1} \quad \text{Eq. 11}$$

$$\theta_i = trace(V^i) \text{ for i = 1, 2, 3 Eq. 13}$$

$$h_0 = 1 - \frac{2\theta_1\theta_2}{3\theta_2^2} \quad \text{Eq. 14}$$

With V representing the covariance matrix of $\mathbf{E}$, z_α is standardised normal variable with confidence limit equal to α. A major benefit of applying PCA to analyse biopharmaceutical data is its ability to be used for on-line monitoring. The PCA model generated from the NOC batches can be used to evaluate batch progression in real-time and utilise this information to alleviate faults and enhance control operations. This PCA projection method utilises a portion of the loading matrix corresponding to the

current lapsed time of the current batch until current sampling time k to calculate the new score vector $t_{new}(k)$ for the selected number of principal components of the model.

$$\mathbf{t}_{new}(k) = \mathbf{x}_{new_{1:Jk}}\mathbf{P}_{1:Jk}\left(\mathbf{P}'_{1:Jk}\mathbf{P}_{1:Jk}\right)^{-1} \quad \text{Eq. 15}$$

Where $\mathbf{x}_{new_{1:Jk}}$ is the available batch data up until current time point k and $\mathbf{P}_{1:JK}$ is the loadings matrix of the NOC batches calculated using data up to time point k. Both the $T^2_{new}(k)$ and $\mathbf{e}_{new}(k)$ are calculated from Eq. 8 and 11, respectively, using the time varying covariance matrix $\mathbf{S}(k)$ and the loadings matrix $\mathbf{P}_{1:Jk}$. The on-line *SPE* enables the distance between the PCA model generated by the NOC batches and the progression of the new batch and is calculated as:

$$SPE_{new}(k) = \sum_{j=1}^{J} \mathbf{e}^2_{new,jk}(k) \quad \text{Eq. 16}$$

The *SPE* and T^2 can act as an on-line indicator of overall system performance. High *SPE* or T^2 indicates that the process is behaving abnormally enabling real-time fault detection. To localise the root cause of any abnormal behaviour the variable contributions towards the *SPE* and T^2 can be evaluated at any time point k as follows:

$$C_{T_{jk}{}^2} = \sum_{a=1}^{A} \mathbf{S}^{-1}_{aa}(k)\mathbf{t}_{new,a}(k)\mathbf{x}_{new,jk}\mathbf{P}_{jk,a} \quad \text{Eq. 17}$$

$$C_{SPE_{jk}} = \mathbf{e}^2_{new,jk}(k) \quad \text{Eq. 18}$$

Where $\mathbf{S_{aa}}(k)$ is the ath digonal element of the time-varying covarance matrix at time point k.

3.2 Partial least Squares (PLS) model development

Partial least squares modelling is similar to PCA in its ability to reduce large data sets into low-dimensional vector spaces. However, this technique enables the prediction of a response variable, **Y**, using the predictor variables contained within **X**. The PLS model is generated from a set of regression vectors maximising the covariance between the **X** and **Y** data. Similar to PCA the initial step in building a PLS model was to construct the **X** data by unfolding all the available variables within each batch using a batch-wise unfolding algorithm ensuring the **X** and **Y** data have an equal number of rows. The PLS model was generated through a non-linear iterative partial least squares (NIPALS) algorithm (Wold et al., 1987). This algorithm generates an outer-relationship that identifies the main sources of variance within each of the data and links them together through an inner-relationship. The outer relationships are generated by decomposing the newly unfolded **X** and **Y** data into R latent score variables [**t**, **u**], loading vectors [**p**, **q**], weights **W** and the model residual matrices **E** and **F**. **t**, **u**, **p**, and **q** can be combined into **T**, **U**, **P**, **Q** and **W** as defined below (Wold et al. 1987):

$$\mathbf{X} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}'_r + \mathbf{E} \ldots \mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \qquad \text{Eq. 19}$$

$$\mathbf{Y} = \sum_{r=1}^{R} \mathbf{u}_r \mathbf{q}_r' + \mathbf{F} \ldots \mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F} \quad \text{Eq. 20}$$

A vector of inner-relationships (**B**) is generated that relates the scores of the **X** data to the **Y** data, which is defined as:

$$\mathbf{B} = \mathbf{U}'\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1} \quad \text{Eq. 21}$$

The PLS model implements an iterative procedure for each latent variable to reach convergence and once the procedure is complete, a matrix of regression coefficients (**β**) can be generated as follows:

$$\boldsymbol{\beta} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} diag(\mathbf{B}) \quad \text{Eq. 22}$$

Where, $\mathbf{W} = (\mathbf{U}^{-1}\mathbf{X})'$. The cumulative sum of the regression coefficients predicts the response variable ($\hat{\mathbf{Y}}$) from the **X** data taking *R* latent variables, which was equal to:

$$\hat{\mathbf{Y}} = \mathbf{X}\sum_{r=1}^{R} \boldsymbol{\beta} \quad \text{Eq. 22}$$

4: Soft-sensor development

The generation of the *PAA* soft-sensor involves generating a PLS model as described in section 3.2 taking the Raman spectra and off-line phenylacetic acid (*PAA*) concentration as the **X** and **Y** data, respectively. The Raman spectra recorded by *IndPenSim* was generated every 12 minutes and recorded data along the wavenumber 250-2250 cm^{-1} resulting in a large two dimensional matrix. The wavenumbers of interest that contain information related to the *PAA* concentration in the bioreactor were equal to 1540:1580 and 1950:2050 cm^{-1}, identified through analysis of Raman spectra recorded from fermentation media spiked with various concentrations of *PAA* (Goldrick 2015). The selected wavenumbers of the Raman spectra were pre-processed using a standard Savitzky-Golay smoothing technique using a 15-point average and taking the first derivative, this pre-processed data was taken as the **X** data. The *PAA* off-line concentrations were taken as the **Y** data in the PLS model and were interpolated using a cubic-spline function to ensure an equal number of rows as the **X** data. The selection of the optimum number of latent variables was based on a cross-validation operation employing a leave-one-out protocol (Martens and Naes 1989).

Results and Discussion

Quality by Design and PAT application

Monitoring and control of penicillin fermentation processes has been around for decades and essential to ensure the production of high yields and product quality remains within specification (Mou and Cooney, 1983, Min et al. 1995 and Lee et al. 2004, Luo and Bao, 2018). The recent Quality by Design (QbD) initiative represents a paradigm shift in biopharmaceutical manufacturing involving a systematic

approach to process optimisation enabled through enhanced process understanding and innovative control strategies. The primary focus of this approach is to ensure a predefined product quality target is confidently and consistently achieved for all batches regardless of inherent process disturbances and batch-to-batch fluctuations. To accelerate the adoption of this systematic approach the regulatory bodies have launched the process analytical technology (PAT) framework (FDA 2004) to promote the application of advanced sensors integrated through innovative control solutions. **Tab. 2** highlights the need for an improved control strategy for *IndPenSim* as both the recipe driven and operator dependent control strategies resulted in significant deviations in annual penicillin production for each of the five campaigns. To demonstrate how a QbD methodology can be correctly implemented for process improvements the three different stages of the PAT framework were implemented using *IndPenSim:*

- **Design stage**: To identify the critical process parameters (CPPs) and subsequent critical quality attributes (CQAs), all the process data recorded in campaign 5 were analysed using multivariate data analysis (MVDA). This campaign is summarised in Table 2 and resulted in 26 batches with 5 of those batches failing to meet the required target penicillin production yield of 2000 kg. All batches implemented an operator dependant control strategy and had a fixed batch length equal to 230 hours. The operator controlled flowrates of substrate (F_s) and phenylacetic acid (F_{PAA}) for this campaign are shown in **Fig. 4A** and **4B**. The significant deviations in these primary control variables results in highly varied penicillin and biomass profiles, as shown in **Fig. 4C** and **4D**, respectively. The need to improve the control strategy implemented on this process is highlighted by the five batches that failed to meet the penicillin demand at harvest shown in **Fig. 4C** and **4D**. To fully exploit the available information recorded throughout this campaign and identify the CPPs influencing the observed deviation in penicillin yields a partial least square (PLS) regression model was implemented to analyse the data. This PLS model was identified using the selected variables shown in **Fig. 5B** taking the final penicillin yield at harvest as the response variable. The development of the PLS model required the data to be restructured using a batch-wise unfolding algorithm enabling the main sources of variation between the variables to be identified. The PLS model was generated using three latent variables that captured 47.7% of the total variance in the X-data and 98.7% of the total variance in the Y-data. All 26 batches were used to build the PLS model with cross validation implemented to determine the appropriate number of latent variables to retain. **Fig. 5A** shows the first and second latent variables of this PLS model and highlights a clustering between the "below" and "above" target batches. This clustering indicates that the below target batches have similar characteristics in the data. To investigate the primary variables influencing these differences in penicillin yields the summed contribution ($\sum_{k=1}^{K} \boldsymbol{\beta}_{jk}^2$) of each process variable is shown in **Fig. 5B** for the first latent variable. The large contribution of the off-line concentrations of phenylacetic acid ($PAA_{offline}$) indicates this variable is highly influential in the final penicillin yields. Therefore, this variable was selected as the primary CPP to be considered for the **Analyse** Stage.

Analyse stage: The current control strategy for *PAA* concentration is to maintain this variable between 600 and 1800 mg L^{-1} through manipulation of the phenylacetic flow rate (F_{PAA}). However, due to the infrequent nature of the off-line measurements of *PAA* combined with a timely 4-hour delay period for this assay, the control of this CPP remains suboptimal. The challenge of controlling this variable within these limits is highlighted through analysis of the annual production records recorded for each campaign. The Analyse stage therefore confirmed a real-time measurement could significantly improve the control of this key process variable. To address this, the inclusion of a Raman spectroscopy analyser within *IndPenSim* was implemented to investigate whether a soft-sensor could be developed to enable real-time predictions of *PAA*. To facilitate the Analyse stage a calibration batch was performed on *IndPenSim* that included the simulated PAT analyser recording a Raman spectrum every 12 minutes as described in section 2.2. The routinely measured off-line *PAA* concentrations were also recorded every 12 hours and used to develop the soft-sensor. The soft-sensor was built using a PLS model as described in section 3.2. The subsequent predictions of *PAA* generated by the soft-sensor are highly comparable to the off-line concentrations of *PAA* for the calibration batch shown in **Fig. 6A**. To demonstrate these predictions in real-time a validation batch was ran using the soft-sensor built from data recorded in the calibration batch. The validation batch enabled on-line predictions of the *PAA* concentration and was shown to be comparable to the off-line *PAA* samples as shown in **Fig. 6A**. The ability to measure the *PAA* in real-time on *IndPenSim* therefore enables the Control stage to be implemented which is the final and most important step in the PAT framework.

Control stage: The final stage of the PAT framework involved the implementation of a proportional integral (PI) control loop that manipulated the F_{PAA} to maintain *PAA* at its set-point. The raw soft-sensor signal, shown in **Fig. 6A**, contains some high frequency fluctuations that may be problematic for the controller. To account for this, the signal was initially filtered using a three point moving average thus minimising any unnecessary control actions. **Figs. 6B** and **6C** highlights this APC solution in operation, where the PI controller was switched on after 25 hours and manipulates the F_{PAA} to maintain the *PAA* concentration at its set-point of 1250 mg L^{-1}. This APC solution was implemented on the *IndPenSim* for a year and the annual penicillin yield was compared against the previous campaigns, which implemented recipe driven and operator dependant control strategies. Implementing this APC strategy resulted in significant improvements in the annual production yields of penicillin. In total 26 bathes were operated through the year and there were no batches that failed to meet the production targets of 2000 kg. The average penicillin yield per batch was 3517 ± 315 kg which represents a 20% overall increase in annual penicillin yields compared to the average of the previous five campaigns. The significant increase in penicillin production demonstrates the benefits of following the QbD methodology and implementing an APC solution utilising the Raman spectroscopy analyser.

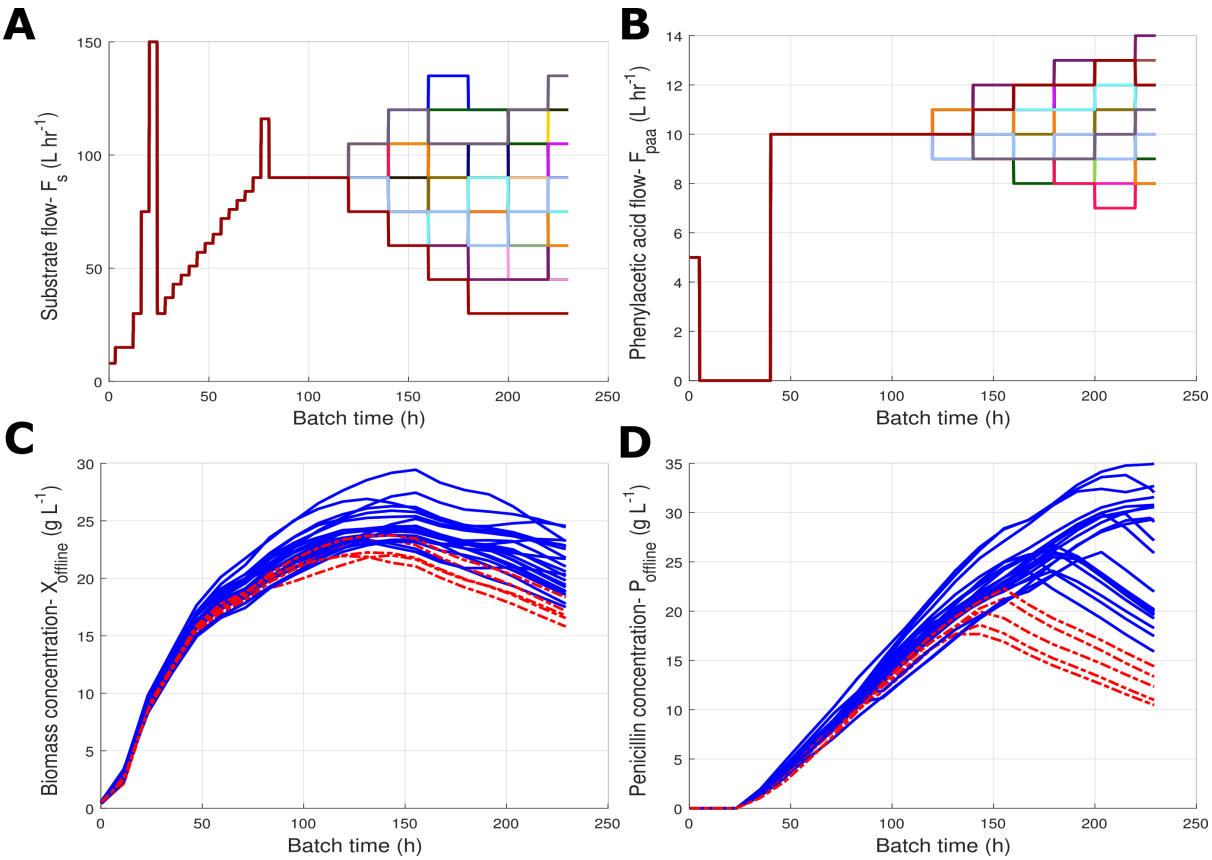

Figure 4: Summary of variable profiles for campaign 5 with **A** highlighting the substrate flow rate (F_s), **B:** Phenylacetic acid flow (F_{PAA}), **C**: Biomass (X) and **D**: Penicillin (P). The failed batches shown in C and D are highlighted by red dashed lines.

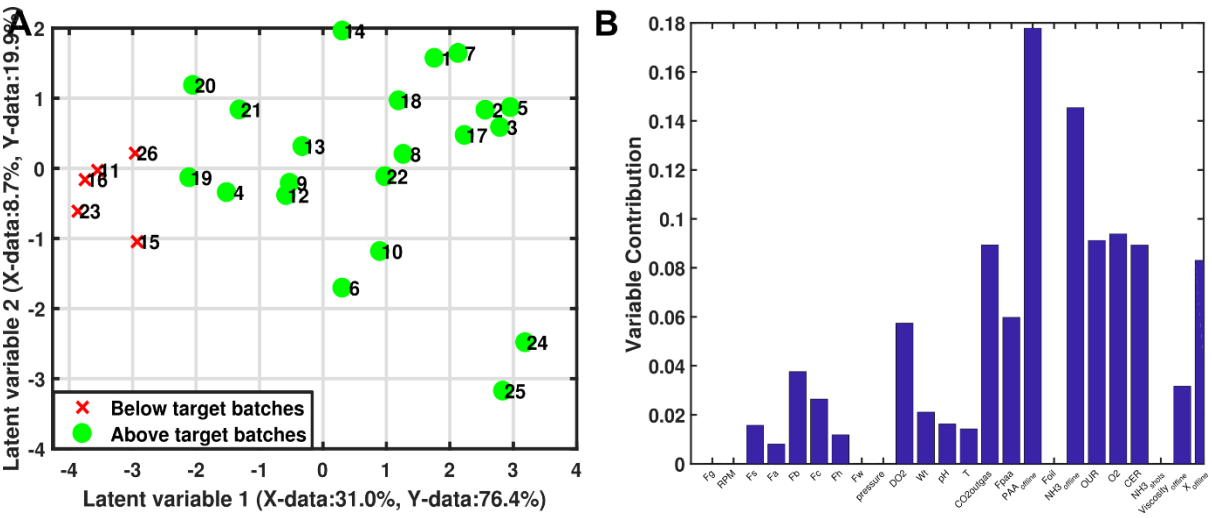

Figure 5: A The scores generated from a PLS model of the above target batches (Penicillin yield > 2000 kg) are represented by green circles and the below target batches (Penicillin yield < 2000 kg) are

represented by the red crosses. The 1st latent variable represents 31.0% and 76.4% of the variance of the **X** and **Y** data, respectively, similarly the 2nd latent represents 8.7% and 19.9% of these data. **B** represents the variable contribution plot showing the normalised weight of each variable calculated using the 1st latent variable from the PLS model.

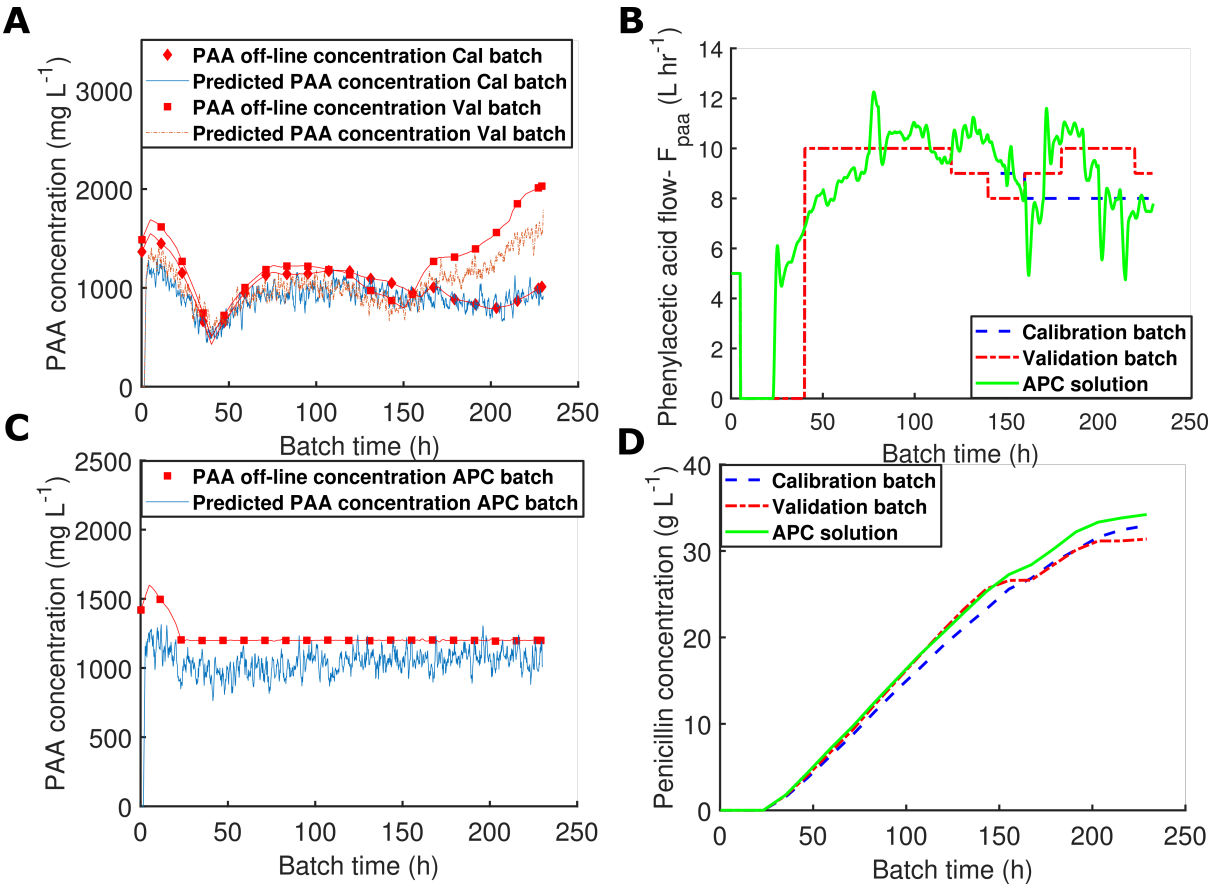

Figure 6: A Calibration and validation batches of the off-line *PAA* samples and the corresponding predictions using a PLS model combined with the Raman spectroscopy analyser. **B** Summary of F_{PAA} for the calibration and validation batches and the APC batch with F_{PAA} controlled using the soft-sensor developed here. **C** Outline of *PAA* controlled using the APC strategy implemented here where the set-point for *PAA* was equal to 1250 mg L^{-1}. **D** Profile of Penicillin concentrations during the calibration, validation and APC controlled batches.

Fault detection

Faults are an inherent hindrance to every manufacturing facility with early detection and subsequent isolation essential to minimise any significant process deviations (Venkatasubramanian et al. 2003). Early detection of faults during biopharmaceutical processes are necessary to ensure all process variables remain within a tight operating window ensuring strict target product requirements are maintained. Monitoring all available measurements is significantly challenging due to the increasing number of on-line and off-line variables recorded on industrial manufacturing facilities. Many biopharmaceutical companies rely on MVDA to help efficiently monitor the multitude of available

process measurements enabling faster detection of process faults (Nomikos and Macgregor 1995). This approach was applied here to provide a benchmark for detecting abnormal processing conditions within *IndPenSim*.

This section demonstrates the application of *IndPenSim* to generate known faults during batches. Two standardised MVDA based fault detection algorithms were implemented to identify these faults. The data set generated from Campaign 4 excluding the 2 below target batches and an additional 5 batches that were considered to be sub-optimal. Campaign 4 represents a typical campaign controlled through a recipe driven control strategy with a fixed batch length and yielded a highly diverse data set. In total there were 17 batches taken as normal operating conditions (NOC) batches with batches 18-21 containing known faults. A comparison between the nominal trajectories and the batches with faults are shown in **Fig. 7** with **A** highlighting the aeration fault, **B** the pH sensor drift fault, **C** the substrate fault and **D** the coolant fault. The nominal biomass (X) and penicillin (P) profiles calculated by averaging all 17 batches are shown in **Figs**. **7E** and **7F**, respectively in addition to highlighting the effect of the process faults on these two CPPs. PCA was selected here based on its ability to compress the large volume of data to a much smaller set of linearly uncorrelated principal components (PCs) enabling direct visualisation of all variables suitable for process monitoring fault detection (Lee et al. 2005b). The 17 NOC batches from Campaign 4 were unfolded to form the **X** data structure and generate the PCA model retaining three principal components as defined in section 3.1. All 22 of the on-line variables recorded by *IndPenSim* were used in the PCA model, the utilisation of only on-line variables enables faults to be detected in real-time.

To evaluate the comparability of the NOC batches with those containing faults the T^2 (Eq. 9) and Q (Eq. 10) statistic were calculated. The Q statistic is shown in **Fig. 8A** and highlights a clear distinction between the NOC batches and batches with faults. Each of the batches containing faults are above the Q statistic 95% confidence limit calculated from Eq. 12-14 indicating abnormal behaviour. The highest Q statistic is batch 17 which contains the aeration fault shown in **Fig. 7E** and **F** to have largest deviation in penicillin and biomass concentrations in comparison to the nominal trajectories. In contrast the T^2 statistic, shown in **Fig. 9A**, indicates all batches to be within the 95% confidence limit calculated using Eq. 8. Gunther et al. (2007) described similar results with the off-line Q statistic outperforming the T^2 statistic in its ability to successfully identify the batches with faults in comparison to NOC batches. Typically, the T^2 is better at identifying systematic errors between batches whereas the Q statistic is better at identifying a new event which that the previous PCA model has not seen which is the case for the faults described in this work. However, both the T^2 and Q statistics have successfully identified abnormal process behaviour on various different industrial processes (Westerhuis et al. 2000; Gülnur et al. 2002; Chio et al. 2008).

A second major advantage of generating these PCA models is their ability to monitor and detect the root cause of any abnormal process behaviour in real-time by analysing the *SPE* and T^2 using equations 8 and 16. These are shown for the NOC batches in **Fig. 8B** and **9B** respectively with the 95% upper control limit (UCL) shown. Monitoring both the *SPE* and T^2 chart in real-time enables any process deviations from nominal trajectories to be subsequently identified. The current monitoring system signals an alarm after the *SPE* or T^2 exceeds an upper control limit (UCL). The *SPE* UCL assumes a χ^2 distribution calculated using equations 12-14 taking the confidence limit α equal to 95%. The χ^2 distribution is the most widely implemented for monitoring the mean vector of a process (Rakitzis and Antzoulakos, 2011). The *SPE* of the four batches with faults are shown in **Figs. 8C** to **8F**. These figures highlight the ability of *SPE* to quickly identify abnormal process behaviour for the aeration faults which occur at hours 20-24 and 100-110. Calculating the variable contribution to the *SPE* at time 20.2 using Eq. 18 highlights a significant contribution from the aeration rate (F_g), as shown in **Fig. 10A**. Additional variable contributions are shown for the carbon dioxide off-gas ($CO2_{offgas}$), the dissolved oxygen (O_2) and the carbon evolution rate (*CER*). The drop in the aeration during this time period shown in **Fig. 7A**, results in a significant drop in the dissolved oxygen and effects the mass balance recoded by the $CO2_{offgas}$ and *CER* measurements explaining their contribution to the *SPE* during this fault. The *pH* sensor fault occurs on batch 19 at approximate hour 50, however the on-line *SPE* only violates the UCL at hour 104. The variables contributions at this time are shown in **Fig. 10B** indicating the error is primarily due to deviations in *pH*. The relative delay in detecting this error is most likely due to the high frequency noise associated to the *pH* process variables highlighted in **Fig. 7B**. Furthermore, the penicillin and biomass concentrations were not directly influenced by the *pH* sensor drift as shown in **Fig. 7E and 7F**. The substrate fault behaves in a similar fashion to the aeration fault and is easily detected by the *SPE* in **Fig. 8C**. The subsequent analysis of the contributions shown in **Fig. 10C** indicates a problem with substrate flow rate (F_s). The coolant fault results in a temperature shift highlighted in **Fig. 7D** and behaves similarly to the *pH* fault with a delay in the UCL violation as shown in **Fig. 8D.** The variable contributions for this time point are shown in **Fig. 10D** and clearly highlight an error with the temperature. This UCL violation occurs approximately when the temperature is 298.25 K which is 0.25 K above its set-point. This enables significant time for corrective action as it is only when the temperature increases to 298.5 K that a drop in penicillin production is observed as shown in **Fig. 7F**. The on-line T^2 are shown in **Fig. 9 B-F** and do not the highlight any process deviations with all the T^2 remaining below 95% confidence limit. The process faults in this work are better captured through the analysis of the *SPE* which summarised the variation not captured by PCA in contrast to the T^2 statistic which is better suited to describing deviations described by the PCA model.

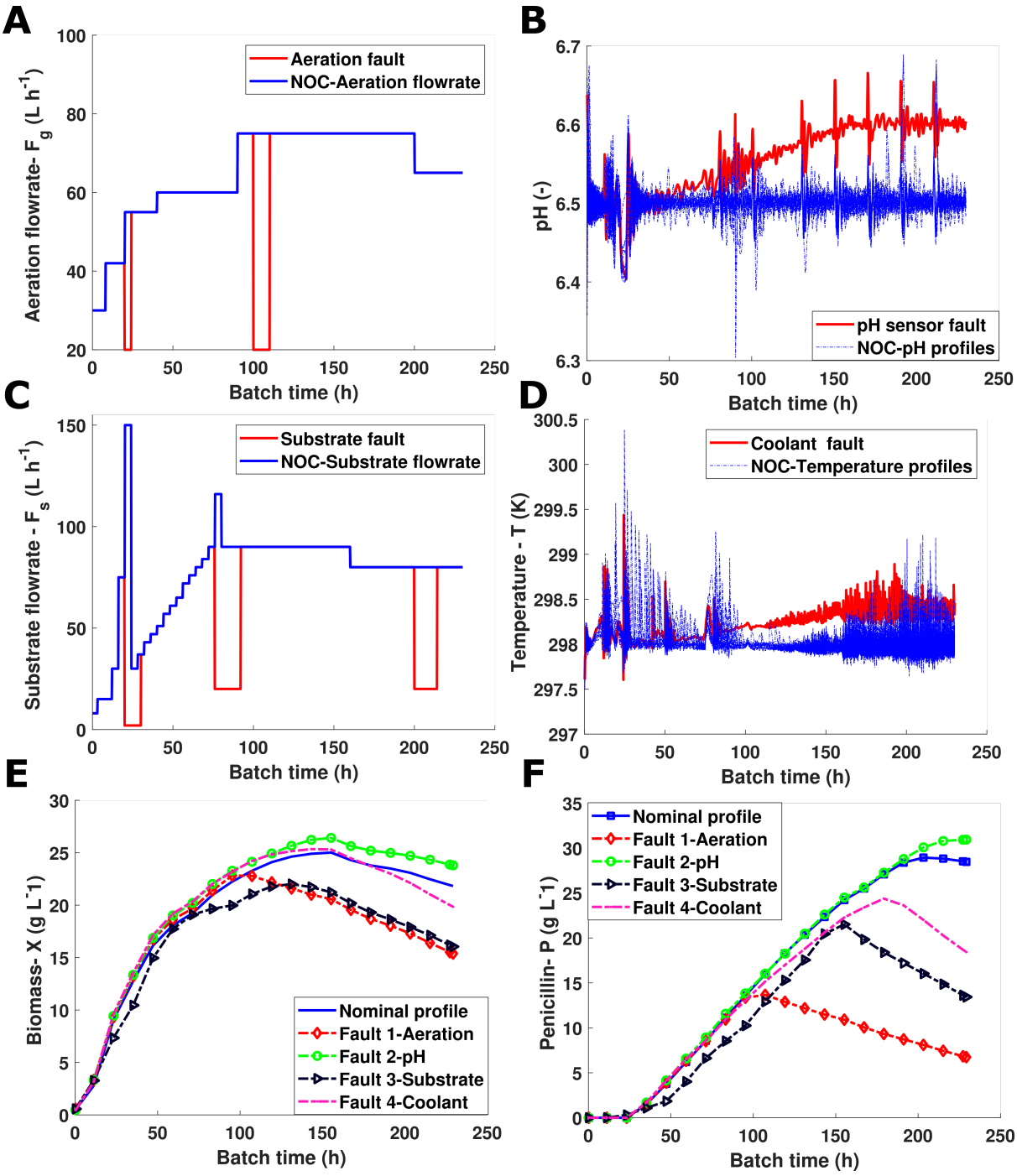

Figure 7: Outline of the nominal trajectories of representative batches (Batches 1-17) from Campaign 4 in additional to time-series profiles of batches 18-21 containing known faults. **A:** Aeration fault (Batch 18), **B:** pH sensor drift fault (Batch 19), **C**: Substrate fault (Batch 20) and **D**: Coolant fault (Batch 21). The nominal Biomass (*X*) and Penicillin (*P*) profiles are shown in **E** and **D**, respectively with the profiles shown for each of the four batches containing faults.

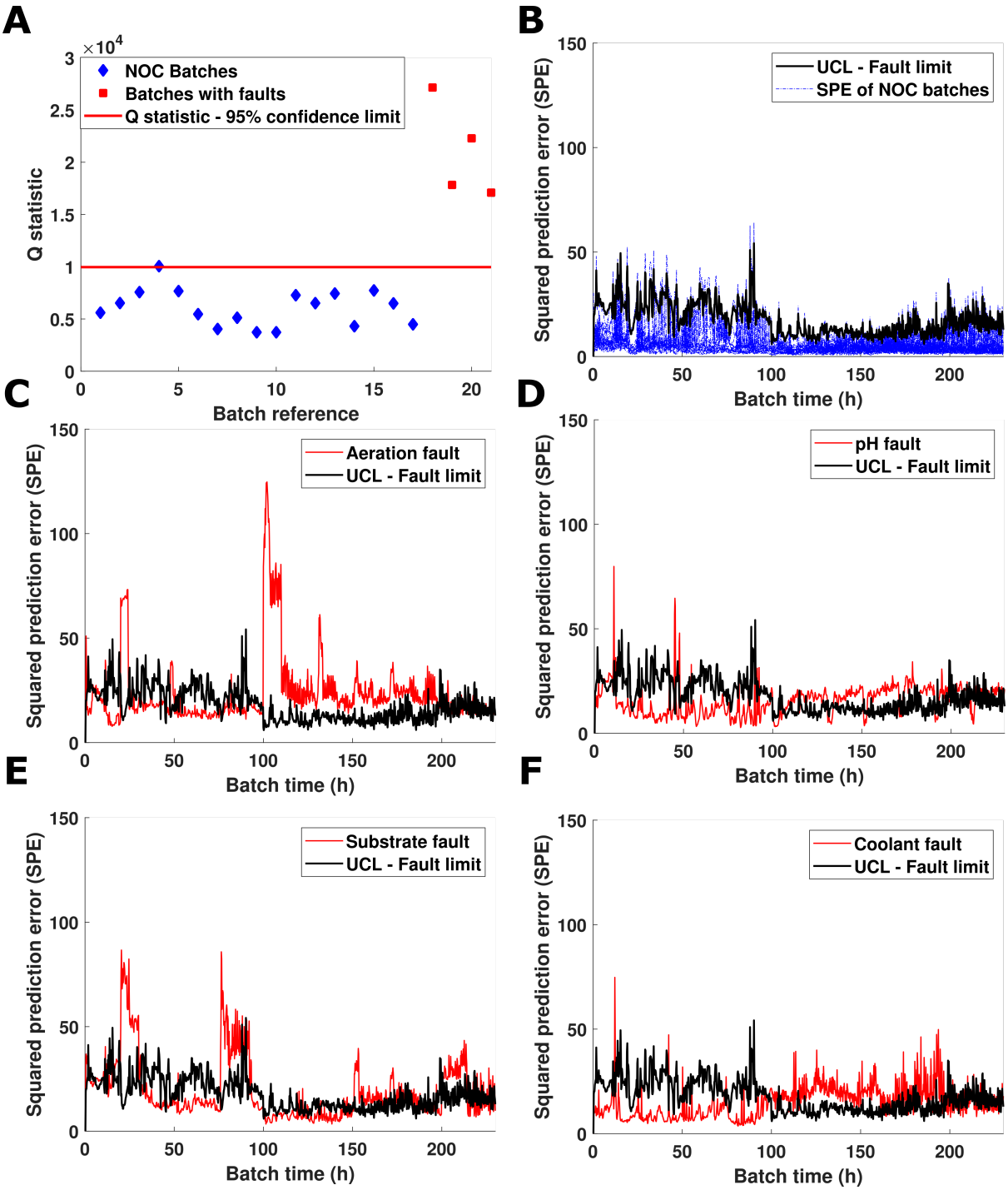

Figure 8: A Plot of Q statistic for each of the 21 batches with the NOC batches represented by diamonds and the batches with faults represented by squares **B** A summary of the SPE recorded for each of the 17 nominal batches with the UCL highlighted. **C-F** A summary of the SPE_k for each of the four batches with faults with the UCL highlighted.

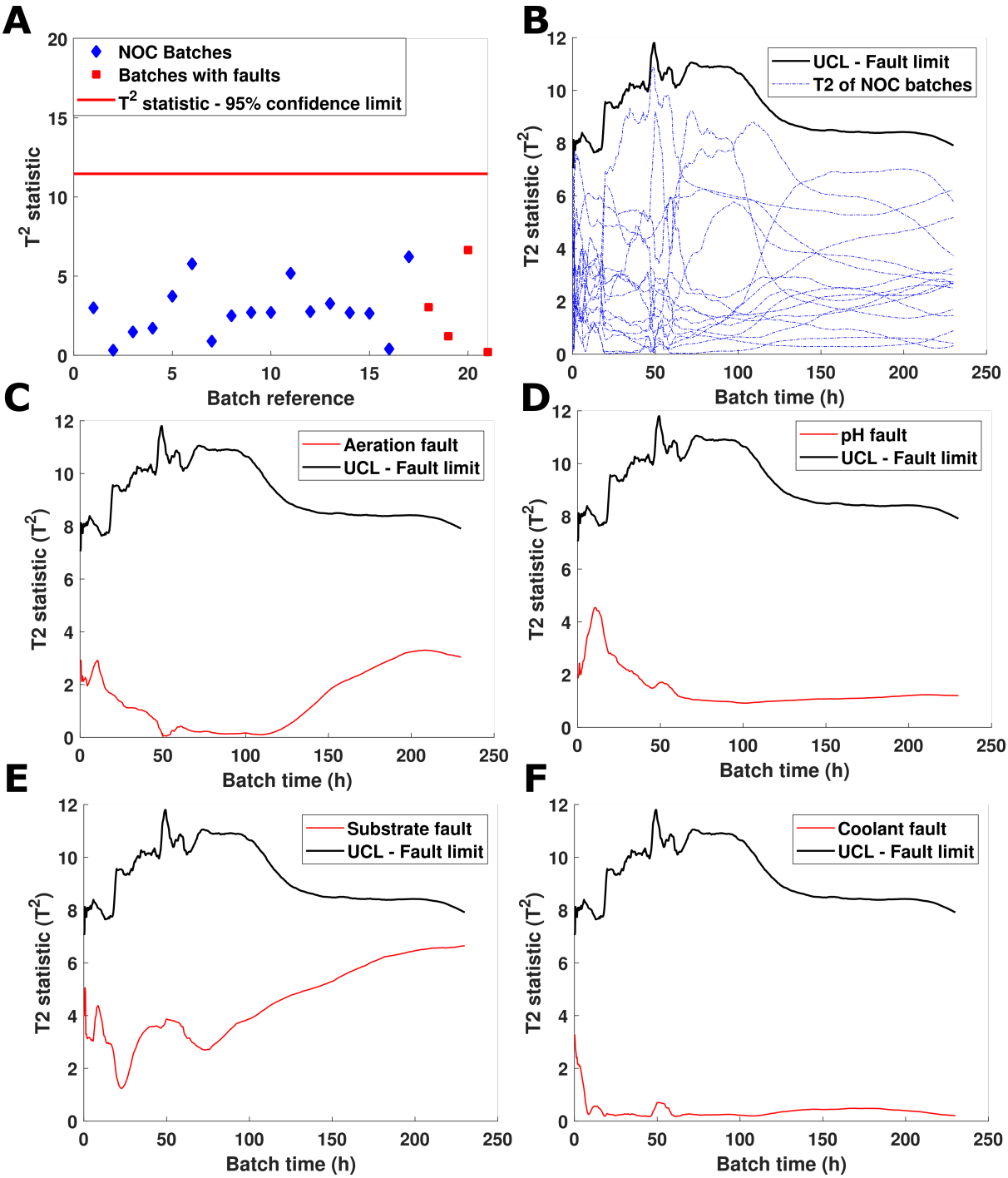

Figure 9: A Plot of T^2 statistic for each of the 21 batches with the NOC batches represented by diamonds and the batches with faults represented by squares **B** A summary of the T^2 recorded for each of the 17 nominal batches with the UCL highlighted. **C-F** A summary of the time-series T^2 for each of the four batches with faults with the UCL highlighted.

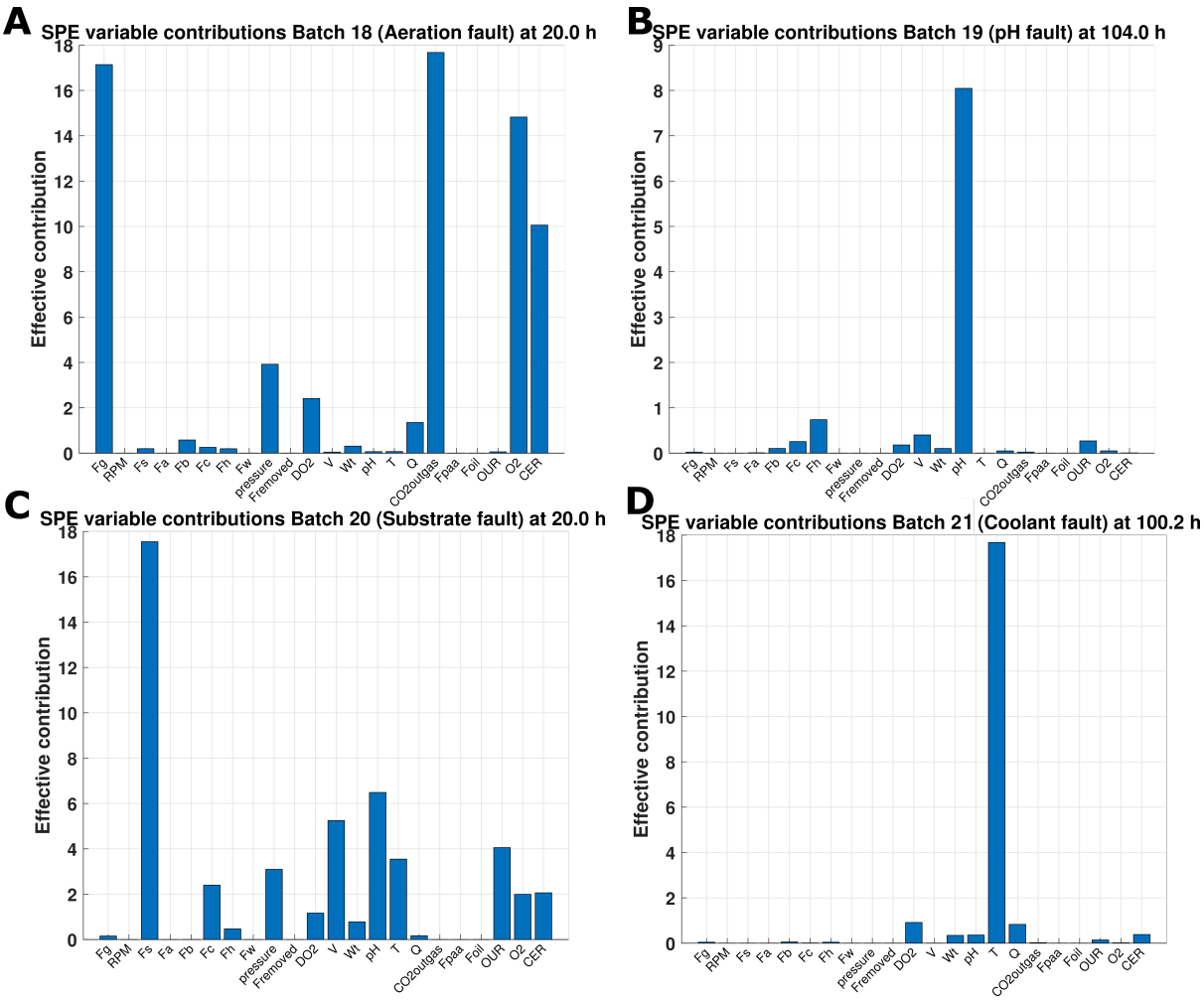

Figure 10: Variable contribution plot of SPE statistic for **A** recorded at time point 20 hours for batch 18 (aeration fault), **B** recorded at time point 104 hours for batch 19 (pH fault), **C** recorded at time point 20 hours for batch 20 (substrate fault), **D** recorded at time point 100.2 hours for batch 21 (coolant fault)

Conclusion

The industrial-scale penicillin simulation (*IndPenSim*) developed in this paper aims to act as a benchmark simulator to develop, evaluate and validate novel and advanced control strategies, applicable to real-world biopharmaceutical manufacturing facilities. The paper outlines a number of highly challenging control objectives to enhance overall yield and productivity requiring the development of adaptive and innovative control solutions. Furthermore, using the simulator all process improvements or modifications can be effectively compared and evaluated against the annual production yields generated by the previous five campaigns implementing operator dependant and recipe driven control strategies. The modifications to *IndPenSim* that are introduced in this paper represent the first bioprocess simulation to include a PAT device that accurately mimics the spectra recorded by a Raman spectroscopy device. The inclusion of this device represents a significant opportunity to help drive FDA's goal of enhancing process understanding and supporting innovative control solutions utilising

real-time sensors. The capabilities and functionality of *IndPenSim* were demonstrated through two case studies. The first involves implementing all three stages of the PAT initiative using the Raman spectroscopy probe to enhance control of phenylacetic acid, previously identified as a CPP. The application of this control strategy resulted in a significant increase in yield improvements, increasing the annual penicillin yields to 3517 kg representing a 20% increase when compared to the previous five campaigns. Furthermore, this control strategy reduced the number of below target batches to zero emphasising the importance of implementing advanced controllers on biopharmaceutical processes. The second case study involved the evaluation of a benchmark fault detection algorithm to identify the occurrence of known faults. The *SPE* statistic significantly outperformed the T^2 statistic in it's ability to identify and locate the root cause of process faults during abnormal process operation. *IndPenSim* and all data presented here are available to download at www.industrialpenicillinsimulation.com and acts as an open resource for researchers to analyse, improve and optimise the current control strategy implemented on this facility.

Acknowledgements

This work was supported by the EPSRC grant (EP/G037620/1) as part of an Engineering Doctorate for SG in Biopharmaceutical Process Development at Newcastle University with financial support and assistance from Perceptive Engineering Ltd.

References

Benyahia, B., Lakerveld, R., Barton, P.I., 2012. A plant-wide dynamic model of a continuous pharmaceutical process. Ind. Eng. Chem. Res., 51(47), 15393-15412.

Birol, G., Ündey, C., Çinar, A., 2002. A modular simulation package for fed-batch fermentation: penicillin production. Comput. Chem. Eng. 26, 1553–1565.

Bocklitz, T., Walter, A., Hartmann, K., Rösch, P. and Popp, J., 2011. How to pre-process Raman spectra for reliable and stable models?. Analytica chimica acta, 704(1-2),47-56.

Chiang, L.H., Leardi, R., Pell, R.J., Seasholtz, M.B., 2006. Industrial experiences with multivariate statistical analysis of batch process data. Chemom. Intell. Lab. Syst., 81(2), 109-119.

Clarke, S.J., Littleford, R.E., Smith, W.E., Goodacre, R., 2005. Rapid monitoring of antibiotics using Raman and surface enhanced Raman spectroscopy. Analyst, 130, 1019-26.

Downs, J.J., Vogel, E.F., 1993. A plant-wide industrial process control problem.Comput. Chem. Eng., 17(3), 245-255.

Food and Drug Administration, 2004. Guidance for industry: PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance. DHHS, Rockville, MD.

Gernaey, K.V., Gani, R., 2010. A model-based systems approach to pharmaceutical product-process design and analysis. Chem. Eng. Sci, 65(21), 5757-5769.

Gunther, J.C., Conner, J.S. and Seborg, D.E., 2007. Fault detection and diagnosis in an industrial fed-batch cell culture process. Biotechnol. Prog., 23(4), 851-857.

Goldrick, S., Mercer, E., Montague, G., Lovett, D., Lennox, B., 2014. Control of an Industrial Scale Bioreactor using a PAT Analyser. IFAC Proceedings Volumes, 47(3), 6222-6227.

Goldrick S., 2015. Application of Multivariate Data Analysis and First Principle Mathematical Modelling to the Biotechnology Industry, Newcastle University, EngD thesis.

Goldrick, S., Holmes, W., Bond, N.J., Lewis, G., Kuiper, M., Turner, R., Farid, S.S., 2017. Advanced multivariate data analysis to determine the root cause of trisulfide bond formation in a novel antibody–peptide fusion. Biotechnol. Bioeng., 114(10), 2222-2234.

Goldrick, S., Lovett, D., Montague, G., Lennox, B., 2018. Influence of Incident Wavelength and Detector Material Selection on Fluorescence in the Application of Raman Spectroscopy to a Fungal Fermentation Process. Bioeng., 5(4), 79-83.

Grimbergen, M.C.M., Van Swol, C.F.P., Kendall, C., Verdaasdonk, R.M., Stone, N., Bosch, J.L.H.R., 2010. Signal-to-noise contribution of principal component loads in reconstructed near-infrared Raman tissue spectra. Appl. Spectrosc., 64(1),8-14.

Jeppsson, U., Pons, M.N., Nopens, I., Alex, J., Copp, J.B., Gernaey, K.V., Rosén, C., Steyer, J.P., Vanrolleghem, P.A., 2007. Benchmark simulation model no 2: general protocol and exploratory case studies. Water Sci. Technol., 56(8), 67-78.

Jackson, J.E. and Mudholkar, G.S., 1979. Control procedures for residuals associated with principal component analysis. Technometrics, 21(3), 341-349.

Kneipp, K., Kneipp, H., Itzkan, I., Dasari, R.R., Feld, M.S., 1999. Ultrasensitive chemical analysis by Raman spectroscopy. Chem. Rev., 99, 2957-76.

Kontoravdi, C., Pistikopoulos, E. N., Mantalaris, A., 2010. Systematic development of predictive mathematical models for animal cell cultures. Comput. Chem. Eng., 34(8), 1192-1198.

Kiparissides, A., Koutinas, M., Kontoravdi, C., Mantalaris, A., & Pistikopoulos, E. N. (2011). 'Closing the loop'in biological systems modeling—From the in silico to the in vitro. Automatica, 47(6), 1147-1155.Kourti, T., 2005. Application of latent variable methods to process control and multivariate statistical process control in industry. Int. J. Adapt. Control Signal Process , 19(4), 213-246.

Lee, J.M., Yoo, C.K., Lee, I.B., 2004. Enhanced process monitoring of fed-batch penicillin cultivation using time-varying and multivariate statistical analysis. J. Biotechnology, 110(2), 119-136.

Lee, J.M., Yoo, C. and Lee, I.B., 2004b. Statistical process monitoring with independent component analysis. J. Process Contr., 14(5), pp.467-485.

Lennox, B., Montague, G. A., Hiden, H. G., Kornfeld, G., Goulding, P. R., 2001. Process monitoring of an industrial fed-batch fermentation. Biotechnol. Bioeng., 74(2), 125-135.

Luo, L., Bao, S., 2018. Knowledge-data-integrated sparse modeling for batch process monitoring. Chem. Engin. Sci., 189, 221-232.

Lyman, P. R., Georgakis, C., 1995. Plant-wide control of the Tennessee Eastman problem. Comput. Chem. Eng., 19(3), 321-331.

Min, R.W., Nielsen, J., Villadsen, J., 1995. Simultaneous monitoring of glucose, lactic acid and penicillin by sequential injection analysis. Anal. Chim. Acta., 312(2),149-156.

Montague, G. A., Morris, A. J., Ward, A. C., 1989. Fermentation monitoring and control: a perspective. Biotechnol. Genet. Eng. Rev., 7(1), 147-188.

Mou, D.G., Cooney, C.L., 1983. Growth monitoring and control through computer-aided on-line mass balancing in a fed-batch penicillin fermentation. Biotechnol. Bioeng., 25(1), 225-255.

Nomikos, P., MacGregor, J.F., 1995. Multivariate SPC charts for monitoring batch processes. Technometrics, 37(1), 41-59.

Oh, S.K., Yoo, S.J., Lee, J.M., 2012. Predicting concentrations of a mixture in bioreactor for on-line monitoring using Raman spectroscopy. IFAC Proceedings Volumes, 45(15), pp.822-827.

Papadakis, E., Woodley, J.M., Gani, R., 2018. Perspective on PSE in pharmaceutical process development and innovation. Comput.-Aided Chem. Engin., 41, 597-656.

Rakitzis, A.C., Antzoulakos, D.L., 2011. Chi-square control charts with runs rules. Methodology and Computing in Applied Probability, 13(4), pp.657-669.

Randek, J., Mandenius, C.F., 2018. On-line soft sensing in upstream bioprocessing. Crit. Rev. Biotechnol., 38(1), 106-121.

Sami Sivri M., Oztaysi B., 2018. Data Analytics in Manufacturing. In: Industry 4.0: Managing The Digital Transformation. Springer, Cham.

Tomba, E., Facco, P., Bezzo, F., Barolo, M., 2013. Latent variable modeling to assist the implementation of Quality-by-Design paradigms in pharmaceutical development and manufacturing: a review. Int. J. Pharm, 457(1), 283-297.

Ündey, C., Ertunç, S., Çınar, A., 2003. Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. Ind. Eng. Chem. Res. 42(20), 4645-4658.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N., Yin, K., 2003. A review of process fault detection and diagnosis: Part III: Process history based methods. Comput. Chem. Eng., 27(3), 327-346.

Wold, S., Geladi, P., Esbensen, K. and Öhman, J., 1987. Multi-way principal components-and PLS-analysis. J. Chemom., 1(1), pp.41-56.Yuan, Q., Lennox, B., McEwan, M., 2009. Analysis of multivariable control performance assessment technique. J. Process Control, 19(5), 751-760.

Westerhuis, J.A., Gurden, S.P. and Smilde, A.K., 2000. Generalized contribution plots in multivariate statistical process monitoring. Chemometrics and intelligent laboratory systems, 51(1), pp.95-114.