Monitoring Recombinant Protein Titer in Escherichia Coli Fermentations Using Advanced Filtering Techniques

Abstract:
Recombinant protein production in *Escherichia coli* (*E. coli*) is crucial for biotechnological applications, but optimizing protein titer is challenging and often requires iterative experimentation and expert guidance. This study utilizes advanced filtering techniques—Kalman, particle, and spatial (bilateral) filters—to enhance titer estimation accuracy using data from 7 fermentation experiments provided by Cytovance Biologics. These filters effectively impute missing titer values and are validated on a secondary dataset from Chinese Hamster Ovary (CHO) cell lines, demonstrating robustness and accuracy. Their integration promises to improve titer estimation in *E. coli* production, streamline optimization, and reduce costs.

Keywords: Biomanufacturing, Filtering Techniques, Recombinant Protein Production, Escherichia coli.

1 Introduction

Recombinant protein production is crucial for pharmaceutical, agricultural, and industrial sectors. *E. coli* is a preferred host for this purpose due to its well-characterized genetics, rapid growth, and cost-effective cultivation [1]. About 30% of recombinant pharmaceutical proteins are produced in *E. coli*, highlighting its significance in biopharmaceutical manufacturing [2]. However, high yields of soluble, functional proteins in *E. coli* are difficult to achieve due to low expression levels and insoluble aggregates known as inclusion bodies [3].

Optimizing protein expression in *E. coli* typically involves trial and error to find suitable conditions, which is time-consuming and resource-intensive. Traditional methods rely on adjusting various parameters such as temperature, pH, induction time, and media composition [4]. However, the complex interplay of these factors necessitates more systematic, predictive approaches. Advanced filtering techniques have shown promise for real-time bioprocess monitoring and control, improving process control precision by providing accurate estimates of state variables and system parameters, even in the presence of noise and uncertainties. Kalman filters (KF) have been used successfully to predict and control fermentation dynamics. Kager et al. [5] applied KFs in *Penicillium chrysogenum* fed-batch cultivation for penicillin production, estimating biomass and substrate concentrations to enhance process control. Dewasme et al. [6] used Extended KFs (EKFs) to estimate acetate concentrations in *E. coli* cultures, integrating biomass and gas analysis sensor measurements due to the lack of reliable acetate probes. EKFs have also been applied in recombinant adeno-associated virus (rAAV) production. Iglesias et al. [7] used EKFs to estimate metabolite concentrations and rAAV production yields based on viable cell density in HEK293 cultures. Kager et al. [8] implemented a particle filter algorithm for biomass, precursor, and product concentration estimation in *Penicillium chrysogenum* fed-batch cultivation, combining online and offline measurements for improved accuracy. Particle filters (PFs) have also been used in mammalian cell cultures to monitor and control fed-batch processes. Simutis et al. [9] demonstrated PFs for biomass and specific growth rate estimation in hybridoma cell cultures, noting lower estimation errors and simpler tuning procedures compared to EKFs, though not addressing recombinant protein production in *E. coli.*

Despite these advances, gaps remain, particularly in integrating and evaluating filtering techniques for optimizing recombinant protein titer in *E. coli* fermentation. Previous studies often focus on specific filters without leveraging the strengths of multiple techniques. Many studies, such as those by

Kager et al. [5] and Simutis et al. [9], did not utilize *E. coli* or address recombinant protein titer, focusing instead on other organisms or metrics. *E. coli* is distinctive due to its rapid growth, well-characterized genetics, and cost-effectiveness, which are crucial for high-yield recombinant protein production but also present unique challenges like inclusion body formation and metabolic burden. This study aims to fill these gaps by developing a monitoring framework using KFs and spatial (bilateral) filters to enhance recombinant protein titer in *E. coli* fermentation. Applying these filters to datasets with high percentages of missing values and validating them against more complete datasets offers a novel contribution to the field, demonstrating potential to streamline optimization, reduce costs, and improve titer estimation accuracy.

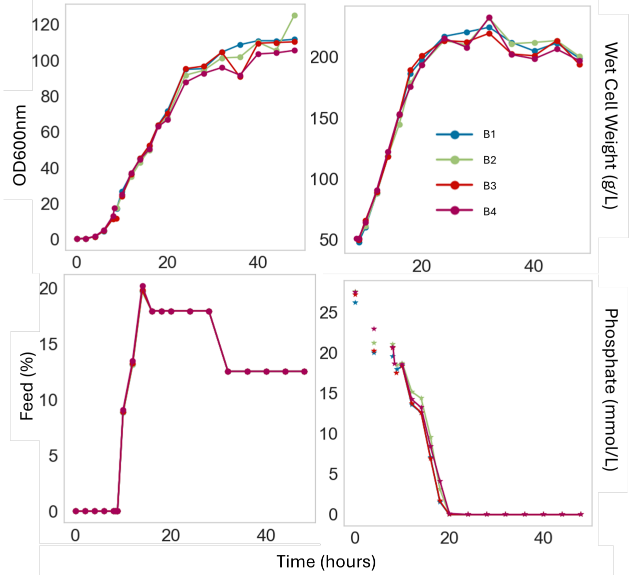


Figure 1: Time series plots of OD600nm, WCW, feed, and phosphate variables across four experiments.

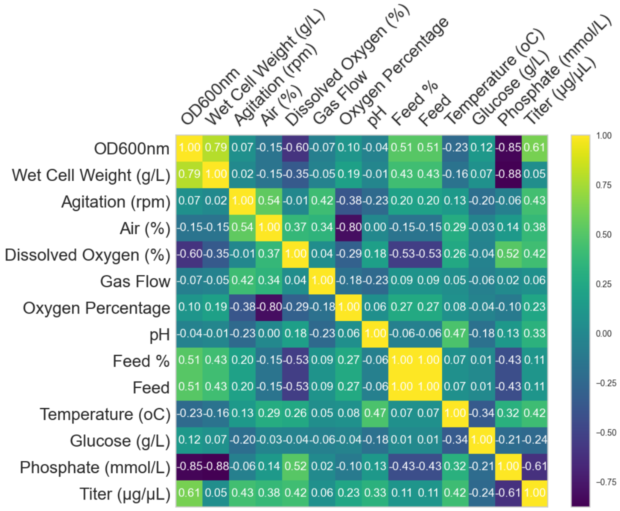
Table 1. Data characteristics of recombinant protein production in *E. coli*. Miss. represents missing data.

Variable	Miss.	Mean	Variance	Median
OD600nm	10%	53.3	1862.2	52.0
WCW	43%	170.2	3375.2	140.0
Agitation	0%	1088.6	42819.2	1185.1
Air	0%	93.4	245.9	99.6
DO	0%	53.4	511.4	41.6
Gas Flow	0%	4.4	1.6	5.0
Oxygen	0%	5.7	168.8	0.4
pH	0%	6.6	0.05	6.7
Feed	0%	10.8	76.6	12.5
Temperature	0%	29.6	1.6	30.0
Glycerol	75%	0.8	4.1	0.0
Glucose	24%	3.8	63.9	0.0
Acetate	24%	25.8	8084.4	1.9
Phosphate	29%	10.2	113.8	7.0
Titer	97%	0.6	0.6	0.4

2 Data

The data for this study, obtained from Cytovance Biologics, includes 7 fermentation experiments aimed at optimizing recombinant protein production in *E. coli*. Project G consists of 4 experiments using 4 fermenters each, totaling 16 fermenters. Project S includes 3 experiments: two with 6 fermenters each and one demo with 4 fermenters, also totaling 16 fermenters. Each experiment is conducted over a 48-hour period, systematically recording both controlled input variables and generated outputs. The input variables include vessel type (5L) and fermenter volume (5L), agitation speed, dissolved oxygen (DO) levels, pH, gas flow rate, and media components. Measured outputs are optical density measured at a wavelength of 600 nm (OD600nm), wet cell weight (WCW), protein titer, and metabolite concentration. These measured outputs are crucial for assessing fermentation process efficiency and productivity. Table 1 provides a summary of the data. For the rest of the paper, we will refer to this dataset as the E. coli data for convenience.

Figure 1 show the dynamics of four key variables over the 48-hour fermentation period across four experiments (*B*1-*B*4 from project S). OD600nm and WCW (mass of cells produced based on the OD600nm) indicate continuous cell growth, which are crucial for protein production as supported by literature [2, 10]. The feed percentage rises sharply initially, then stabilizes, reflecting controlled feeding for optimal growth. The feed, consisting of a 50% glucose solution, was added whenever acetate levels fell below 5 mmol/L. The feed percentage rises sharply initially, then stabilizes, reflecting controlled feeding for optimal growth. The feed, consisting of a 50% glucose solution, was added whenever acetate levels fell below 5 mmol/L. The downward trend in phosphate concentration, especially



	OD600nm	Wet Cell Weight (g/L)	Agitation (rpm)	Air (%)	Dissolved Oxygen (%)	Gas Flow	Oxygen Percentage	pH	Feed %	Feed	Temperature (oC)	Glucose (g/L)	Phosphate (mmol/L)	Titer (µg/µL)
OD600nm	1.00	0.79	0.07	-0.15	-0.60	-0.07	0.10	-0.04	0.51	0.51	-0.23	0.12	-0.85	0.61
Wet Cell Weight (g/L)	0.79	1.00	0.02	-0.15	-0.35	-0.05	0.19	-0.01	0.43	0.43	-0.16	0.07	-0.88	0.05
Agitation (rpm)	0.07	0.02	1.00	0.54	-0.01	0.42	-0.38	-0.23	0.20	0.20	0.13	-0.20	-0.06	0.43
Air (%)	-0.15	-0.15	0.54	1.00	0.37	0.34	-0.80	0.00	-0.15	-0.15	0.29	-0.03	0.14	0.38
Dissolved Oxygen (%)	-0.60	-0.35	-0.01	0.37	1.00	0.04	-0.29	0.18	-0.53	-0.53	0.26	-0.04	0.52	0.42
Gas Flow	-0.07	-0.05	0.42	0.34	0.04	1.00	-0.18	-0.23	0.09	0.09	0.05	-0.06	0.02	0.06
Oxygen Percentage	0.10	0.19	-0.38	-0.80	-0.29	-0.18	1.00	0.06	0.27	0.27	0.08	-0.04	-0.10	0.23
pH	-0.04	-0.01	-0.23	0.00	0.18	-0.23	0.06	1.00	-0.06	-0.06	0.47	-0.18	0.13	0.33
Feed %	0.51	0.43	0.20	-0.15	-0.53	0.09	0.27	-0.06	1.00	1.00	0.07	0.01	-0.43	0.11
Feed	0.51	0.43	0.20	-0.15	-0.53	0.09	0.27	-0.06	1.00	1.00	0.07	0.01	-0.43	0.11
Temperature (oC)	-0.23	-0.16	0.13	0.29	0.26	0.05	0.08	0.47	0.07	0.07	1.00	-0.34	0.32	0.42
Glucose (g/L)	0.12	0.07	-0.20	-0.03	-0.04	-0.06	-0.04	-0.18	0.01	0.01	-0.34	1.00	-0.21	-0.24
Phosphate (mmol/L)	-0.85	-0.88	-0.06	0.14	0.52	0.02	-0.10	0.13	-0.43	-0.43	0.32	-0.21	1.00	-0.61
Titer (µg/µL)	0.61	0.05	0.43	0.38	0.42	0.06	0.23	0.33	0.11	0.11	0.42	-0.24	-0.61	1.00

Figure 2: Correlation matrix in E. coli data

when levels drop below 5 mmol/L, indicates the onset of protein production. Phosphate levels are monitored every four hours from the beginning of the run until depletion. At around 8 hours, once glycerol is depleted in the batch media and glucose is introduced, phosphate levels are checked hourly. The phoA induction process is initiated at 20 hours when the phosphate level reaches zero. Figure 2 shows strong positive correlations between OD600nm and wet cell weight (0.79), and negative correlations between phosphate and OD600nm (-0.85) and phosphate and wet cell weight (-0.88), confirming phosphate as a limiting nutrient in protein synthesis. Effective imputation is crucial for optimizing recombinant protein titer production in *E. coli* fermentation due to substantial missing titer data (>90% missing). Unlike other variables tracked bi-hourly, titer data are recorded only at the end of the 48-hour experiment period for 16 of the 32 fermenters, mainly in project S. The infrequency of titer measurements is due to the labor-intensive and costly nature of the required assays [2], which often involve complex biochemical processing that is not feasible to perform hourly. The sensitivity of titer assays and the need for specialized equipment further complicate frequent data collection [4], hindering real-time monitoring.

2.1 CHO Dataset

To validate the proposed filtering techniques for the *E. coli* dataset, a secondary dataset with more complete titer values is used. This dataset, also from Cytovance Biologics, includes data from Chinese Hamster Ovary (CHO) cell lines, specifically CHO-S strain. Unlike the *E. coli* dataset, with over 90% missing titer values, the CHO dataset has significantly fewer missing values, providing a robust validation basis. The CHO-S dataset has 640 records with 50% missing values across four experiments (E_1-E_4).

3 Methods

In this section, advanced filtering techniques are employed to impute missing values for data analysis. Given that titer values are recorded only at the end of the 48-hour experiment, several assumptions facilitate effective imputation. Initial titer values are assumed to be zero until significant phosphate depletion ($\leq$ 5 mg/ml) indicates the commencement of protein synthesis. End-point titer values are assumed within 0.3 to 2.2 $\mu g/\mu L$, based on empirical data and expert knowledge, ensuring realistic bounds. Then, we apply advanced filtering techniques for imputation.

3.1 Kalman Filters

Kalman filters (KFs), known for efficient real-time state estimation [9], are employed to impute missing protein titer data in *E. coli* [11, 7]. The standard Kalman filter algorithm identifies optimal estimates of both measured and unmeasured state variables by integrating data from a linear stochastic difference model with real-time measurements [12]. The state vector $\mathbf{x}_t$ is the protein titer at time t. The system model describes the evolution of the state from time t to time $t-1$, formulated as

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \mathbf{w}_t$$

where A is the state transition matrix, B is the control input matrix, $\mathbf{u}_t$ represents control inputs (e.g., pH, temperature, dissolved oxygen) at time t, and $\mathbf{w}_t$ is the process noise at time t with covariance Q. The system model is combined with the measurement model, which explains the relationship between the state and the measurement at time t as follows,

$$\mathbf{z}_t = H\mathbf{x}_t + \mathbf{v}_t$$

where $\mathbf{z}_t$ is the measurement variable at time t, H is the observation matrix, and $\mathbf{v}_t$ is the measurement noise at time t with covariance R. The KF algorithm consists of two main steps: [13]: (1) The system model predicts the state variables and the estimation error covariance until the initial measurement is obtained (prediction step), (2) Predicted model estimates are integrated with measured values to produce corrected estimates (correction step).

3.2 *Particle Filters*

Particle filters (PF), also known as Sequential Monte Carlo methods, aestimate the state of non-linear and non-Gaussian systems by approximating the posterior distribution of state variables using a set of weighted particles [14, 9]. These methods have been utilized to impute missing protein titer values during *E. coli* fermentation [15]. The PF algorithm involves the following steps [9]: (1) **Initialization:** Generate an initial population of state vectors (particles) $\{x_0^i\}_{i=1}^N$, and set the initial weight of the i-th particle as $w_0^i = \frac{1}{N}$. We assume $x_0^i \sim \mathcal{N}(0, Q)$.
(2) **Prediction:** predict state of the i-th particle (x_t^i) at time t through the process model $x_t^i = f(x_{t-1}^i, u_{t-1}) + w_{t-1}^i$, where $w_{t-1}^i \sim \mathcal{N}(0, Q)$.
(3) **Weight Update:** Update the weight of the i-th particle at time t based on the likelihood of observations ${}_t^i = w_{t-1}^i \cdot p(y_t \mid x_t^i)$, where $p(y_t \mid x_t^i)$ is the likelihood of the observation y_t given the predicted state x_t^i. Normalize the weights so that $\sum_{i=1}^N w_t^i = 1$.
(4) **Resampling:** Resample the particles based on their weights to avoid particle degeneracy [16].
(5) **Estimation:** The state estimate ($\hat{x}_t$) at time t is given by the weighted mean of the particles $\hat{x}_t = \sum_{i=1}^N w_t^i x_t^i$. The likelihood function is modeled as $p(y_t \mid x_t^i) \sim \mathcal{N}(h(x_t^i), R)$, where y_t is the measurement at time t and R is the measurement noise covariance. $h(x_t^i)$ is the measurement function mapping the state x_t^i to the observation space. [9]. This method is ideal for estimating missing titer data in non-linear fermentation processes [9].

3.3 *Spatial (Bilateral) Filters*

Bilateral filters (BFs), a type of spatial filter, are well-known for their edge-preserving properties, making them perfect for noise reduction and signal smoothing without blurring important edges [17, 18]. In the context of optimizing recombinant protein titer in *E. coli* fermentation, BFs can effectively impute missing protein titer values while preserving the integrity of key features in the data [18]. The filtered output $\mathbf{y}_i$ for a data point i is calculated as a weighted average of neighboring data points, with the weights are determined by spatial distance and intensity difference [19], as follows:

$$\mathbf{y}_i = \frac{1}{W_i} \sum_{j \in \mathcal{N}(i)} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_s^2}\right) \exp\left(-\frac{|I_i - I_j|^2}{2\sigma_i^2}\right) \mathbf{x}_j$$

where W_i is the normalization factor:

$$W_i = \sum_{j \in \mathcal{N}(i)} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_s^2}\right) \exp\left(-\frac{|I_i - I_j|^2}{2\sigma_i^2}\right)$$

Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the input values at positions i and j, I_i and I_j are the intensity values, σ_s is the spatial standard deviation, and σ_i is the intensity standard deviation [18]. The hyperparameters σ_s and σ_i are crucial for the filter's effectiveness and are determined through grid search methods [20, 21].

4 Results and Discussion

Advanced filtering techniques are implemented to in Python using the NumPy library to estimate recombinant protein titer ($\mu g/\mu L$) in *E. coli* fermentation. The process is performed on a laptop with an AMD Ryzen 5 7535HS processor, 64 GB of RAM, and a 64-bit operating system. The optimal hyperparameters for each filtering technique are determined using grid search. For the Kalman filters, control inputs such as pH, temperature, and dissolved oxygen percentage are included. The state transition and observation matrices are set to the identity matrix, with process noise covariance $Q = 0.01I$ and measurement noise covariance $R = 0.01I$, where I is the identity matrix. In particle filters, the optimal number of particles is set as 1000, with process noise and measurement noise standard deviations of 0.1 and 0.01, respectively. In bilateral filters, a window size of 5 and standard deviations of 1.0 are employed for both spatial and intensity parameters to balance smoothing and detail preservation.

Figure 3 illustrates the filtered titer values of two experimental batches (B_1 and B_2) in E. coli data, utilizing KFs, PFs, and BFs. The KF applied to these

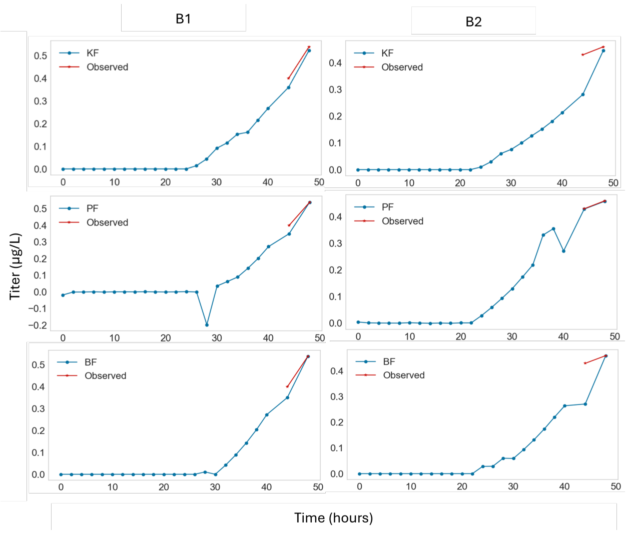


Figure 3: Filtered titer values of two experimental batches (B_1 & B_2) in E. coli data using KF (first row), PF (second row), and BF (third row).

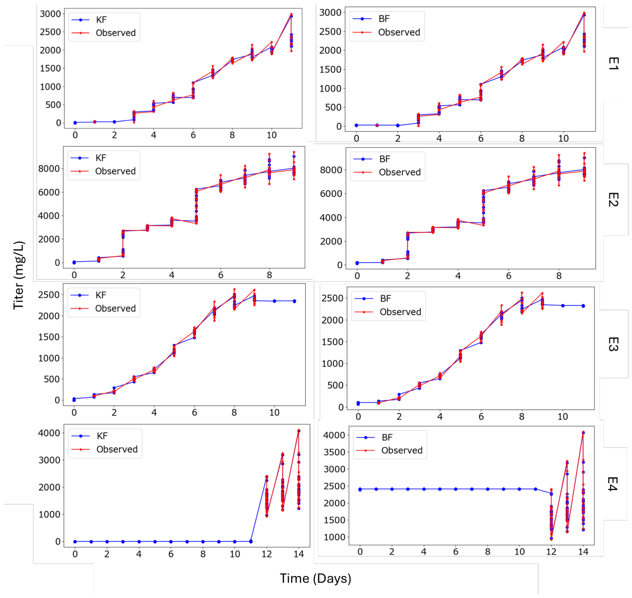


Figure 4: Filtered titer values for four experiments E_1-E_4. Left (right) column shows KF (BF) results.

batches shows a smooth growth of titer values over time, with fewer fluctuations compared to the observed data. PFs capture the non-linear dynamics of the fermentation process, with titer values showing more variability and adapting to sudden changes, providing a more accurate estimate compared to KFs. BFs balance smoothing and preserving key data features, effectively maintaining important trends and variations in titer values.

4.1 *Validation Using CHO Data*

Further, we evaluated the filtering techniques using the CHO dataset, which had fewer missing values. Figure 4 shows the filtered titer values for four experiments E_1 - E_4 using KF and BF. The figure is divided into two columns: the left column shows the results of KF, while the right column displays the results of BF. Both KF and BF plots show a close fit to the observed data. The results of the PF were inferior compared to the other two techniques, and we decided not to include it due to page limitations. We note that the BF estimated values maintain important data features while effectively reducing noise.

Table 2. RMSE, MAE, and MAPE results for KF and BF for the CHO dataset for four experiments.

E.	**KF**			**BF**		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
E1	79.5	64.4	14.8%	79.4	64.3	13.9%
E2	203.4	159.0	5.1%	203.3	158.5	4.8%
E3	76.1	57.4	6.7%	75.7	56.7	6.2%
E4	64.5	53.0	3.5%	61.7	49.7	3.3%

5 Conclusion

This paper have implemented the advanced filtering techniques—Kalman, particle, and spatial (bilateral) filters—to impute missing recombinant protein titer in *E. coli* fermentations. Using Cytovance Biologics data, these filters enhanced imputation accuracy and reliability. KF provided robust smoothing, PF adapted to non-linear dynamics, and BF preserved data features while reducing noise. Validation with CHO data confirmed these techniques' robustness and applicability. This study highlights integrating computational techniques with experimental bioprocessing to enhance fermentation stability and efficiency.

6 Acknowledgement

The authors would like to thank Cytovance Biologics for datasets and collaboration, and to OKC ED Fd, OU Biotech Core Facility, and U.S. Economic Development Administration (Award #: 08-79-05677) for the support.

References

[1] Thomas Gundinger, Stefan Kittler, Sabine Kubicek, Julian Kopp, and Oliver Spadiut. Recombinant protein production in e. coli using the phoa expression system. *Fermentation*, 8(4):181, 2022.

[2] Narjeskhatoon Habibi, Alireza Norouzi, Siti Z Mohd Hashim, Mohd Shahir Shamsir, and Razip Samian. Prediction of recombinant protein overexpression in escherichia coli using a machine learning based model (rpolp). *Computers in biology and medicine*, 66:330–336, 2015.

[3] Christos P Papaneophytou and George Kontopidis. Statistical approaches to maximize recombinant protein expression in escherichia coli: a general review. *Protein expression and purification*, 94:22–32, 2014.

[4] Catherine Ching Han Chang, Chen Li, Geoffrey I Webb, BengTi Tey, Jiangning Song, and Ramakrishnan Nagasundara Ramanan. Periscope: quantitative prediction of soluble protein expression in the periplasm of escherichia coli. *Scientific reports*, 6(1):21844, 2016.

[5] Julian Kager, Johanna Bartlechner, Christoph Herwig, and Stefan Jakubek. Direct control of recombinant protein production rates in e. coli fed-batch processes by nonlinear feedback linearization. *Chemical Engineering Research and Design*, 182:290–304, 2022.

[6] Laurent Dewasme, Guillaume Goffaux, A-L Hantson, and A Vande Wouwer. Experimental validation of an extended kalman filter estimating acetate concentration in e. coli cultures. *Journal of Process Control*, 23(2):148–157, 2013.

[7] Cristovão Freitas Iglesias Jr, Xingge Xu, Varun Mehta, Mounia Akassou, Alina Venereo-Sanchez, Nabil Belacel, Amine Kamen, and Miodrag Bolic. Monitoring the recombinant adeno-associated virus production using extended kalman filter. *Processes*, 10(11):2180, 2022.

[8] Julian Kager, Christoph Herwig, and Ines Viktoria Stelzer. State estimation for a penicillin fed-batch process combining particle filtering methods with online and time delayed offline measurements. *Chemical Engineering Science*, 177:234–244, 2018.

[9] R Simutis, V Galvanauskas, D Levisauskas, J Repsyte, and V Grincas. State estimation of a biotechnological process using extended kalman filter and particle filter. *Veterinary and Agricultural Engineering*, pages 920–924, 2014.

[10] Domenico Bonanni, Mattia Litrico, Waqar Ahmed, Pietro Morerio, Tiziano Cazzorla, Elisa Spaccapaniccia, Franca Cattani, Marcello Allegretti, Andrea Rosario Beccari, Alessio Del Bue, et al. A deep learning approach to optimize recombinant protein production in escherichia coli fermentations. *Fermentation*, 9(6):503, 2023.

[11] PR Patnaik. An integrated hybrid neural system for noise filtering, simulation and control of a fed-batch recombinant fermentation. *Biochemical Engineering Journal*, 15(3):165–175, 2003.

[12] Andrew H Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[13] Greg Welch, Gary Bishop, et al. An introduction to the kalman filter. 1995.

[14] Petar M Djuric, Jayesh H Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F Bugallo, and Joaquin Miguez. Particle filtering. *IEEE signal processing magazine*, 20(5):19–38, 2003.

[15] James Carpenter, Peter Clifford, and Paul Fearnhead. Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, 146(1):2–7, 1999.

[16] Yvo Boers and Johannes N Driessen. Particle filter based detection for tracking. In *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, volume 6, pages 4393–4397. IEEE, 2001.

[17] Qingxiong Yang. Recursive bilateral filtering. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 399–413. Springer, 2012.

[18] Sylvain Paris and Frédo Durand. A fast approximation of the bilateral filter using a signal processing approach. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*, pages 568–580. Springer, 2006.

[19] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998.

[20] Ruturaj G Gavaskar and Kunal N Chaudhury. Fast adaptive bilateral filtering. *IEEE transactions on Image Processing*, 28(2):779–790, 2018.

[21] Michael Elad. On the origin of the bilateral filter and ways to improve it. *IEEE Transactions on image processing*, 11(10):1141–1151, 2002.