Full-state monitoring of protein refolding reactions using particle filters and delayed measurements

Julian Karger¹, Jan Niklas Pauk², Chika Linda Igwe², and Christoph Herwig³

¹Danmarks Tekniske Universitet Institut for Kemiteknik
²Competence Center CHASE GmbH
³Lisalis GmbH

August 19, 2023

Abstract

Proper monitoring as basis for process optimization and control of protein refolding reactions in real-time is difficult and currently available techniques are either expensive, not applicable in real-time or give only limited information about the ongoing process. Model-based methods such as particle filters (PFs) have been used in different biological systems for state-estimation to overcome difficulties arising from states that are hard or impossible to measure, often low measurement frequencies and high measurement delay. Since recent approaches had difficulties to overcome all these problems, a novel approach via a PF including a mechanistic model is used. The PF is calibrated and tuned with experimental data and its applicability validated with two additional experiments. It is shown how augmentation of model parameters can be used for state-estimation in real-time to better adapt to model inaccuracies, poor model calibration or application of the calibrated model to a new process. Furthermore, it is shown that the PF can deal with low measurement frequencies and high measurement delay, resulting in reliable tracking of the process with normalized root mean squared errors (NRMSE) of the native protein and folding intermediates between 3.44 and 6.62 %, values in the range of 18 to 93 % less compared to simple feed-forward simulation.
ARTICLE

Full-state monitoring of protein refolding reactions using particle filters and delayed measurements

Jan Niklas Pauk1,2 | Chika Linda Igwe1,2 | Christoph Herwig3 | Julian Kager*4

Abstract

Proper monitoring as basis for process optimization and control of protein refolding reactions in real-time is difficult and currently available techniques are either expensive, not applicable in real-time or give only limited information about the ongoing process. Model-based methods such as particle filters (PFs) have been used in different biological systems for state-estimation to overcome difficulties arising from states that are hard or impossible to measure, often low measurement frequencies and high measurement delay. Since recent approaches had difficulties to overcome all these problems, a novel approach via a PF including a mechanistic model is used. The PF is calibrated and tuned with experimental data and its applicability validated with two additional experiments. It is shown how augmentation of model parameters can be used for state-estimation in real-time to better adapt to model inaccuracies, poor model calibration or application of the calibrated model to a new process. Furthermore, it is shown that the PF can deal with low measurement frequencies and high measurement delay, resulting in reliable tracking of the process with normalized root mean squared errors (NRMSE) of the native protein and folding intermediates between 3.44 and 6.62 %, values in the range of 18 to 93 % less compared to simple feed-forward simulation.

KEYWORDS:
Process Monitoring; Particle Filter; Protein Refolding; State-Estimation; Mechanistic Model

1 | INTRODUCTION

Biological systems are by nature prone to variation, uncertainty and despite recent advances, mechanistic descriptions of these complex systems are only approximations of the true underlying process. To increase process robustness, sequential monte carlo (SMC) methods have been widely used as Bayesian state-observers (Doucet et al.). These techniques combine process analytical technologies (PAT) with mathematical models, thus enabling monitoring of process states that are otherwise not measurable and can lead to a reduced number of necessary measurements. Particle filters (PFs), a subclass of SMC methods, are often used for biological systems, because they can handle non-linear process and measurement models as well as non-gaussian noise (Doucet et al.). Good convergence to process data in biological upstream processes was shown (Kager et al.; Sinner et al.; Müller et al.).

The PF estimates all process states by approximating its unknown distribution by drawing a number of particles from a known distribution - a technique called importance sampling - with increasing accuracy as the number of particles increases. Repeating this process sequentially using past process information and updating the a priori belief with measurements - through scaling
of each particles individual likelihood by a weight - yields the recursive sequential importance sampling (SIS) (Doucet et al.).
To counteract particles with poor representation of the process dynamics - so called sample impoverishment - Gordon et al. 
introduced a resampling step to eliminate particles with low weight by redrawing all based on the a posteriori weight distribution.
This filter is called the bootstrap or sequential importance resampling (SIR) filter.

The underlying mechanistic models contain model parameters estimated by fitting the model onto experimental data. As such,
calibrated models usually include time invariant model parameters specific to data-sets and generalization can be challenging
when conditions are changed that are not reflected by the model. Thus, approaches have been made to incorporate parameter
estimation into filtering applications to be able to adapt to changing process conditions and to counteract model mismatch
(Kantas et al.). Kitagawa first extended the process model by the model parameters which are then propagated by a random
walk model with white noise. This approach has shown enhanced estimation of the process compared to static model parameters
for biological upstream processes (Kager et al., Sinner et al., Müller et al.). Furthermore, if not all necessary states can be
measured on-line, at-line measurements have to be included in the state-estimation to combine on-line measurements with at-line
measurements to observe an otherwise unobservable system (Kager et al.). At-line measurements often come with a rather large
time delay between sampling time and measurement evaluation and unlike on-line methods they are often drawn in irregular
intervals (Gopalakrishnan et al.). To account for the time delay $\Delta t$, the pdf of the filter is updated at the measurement point $t_{k-\Delta t}$
with the incoming measurement and subsequently recalculated using the historical input vector until $t_k$ (Kager et al.).

Applications using different types of state-observers have been shown for downstream and other operations (Hørsholt et al.,
Peidl et al., Künner and Engell, Schiemer et al., Mohd Ali et al.). However, to our knowledge, no state-observer using a mecha-nistic model has been used for full-state monitoring of protein refolding reactions in real-time (Pauk et al.). This is a relevant
downstream operation, because the formation and accumulation of aggregated and non-active proteins in inclusion bodies (IBs)
is one of the main challenges of recombinant protein production using *Escherichia coli* (Pham et al., McElwain et al.). Activation
of these misfolded proteins involves time consuming methods of solubilization and refolding in appropriate conditions. A
standard procedure is the dilution of the solubilized protein into large amounts of buffer. Different dilution modes can be applied,
e.g. a instantaneous dilution (batch) or continuous buffer addition (fed-batch) (Pauk et al.). In either case, high dilution of the
solubilized protein is necessary to achieve competitive yields, as the higher order aggregation reaction is favored with rising
protein concentrations (De Bernardes Clark et al.). To prevent unnecessary high dilutions and waiting times, on-line monitoring
of protein refolding is necessary. However, on-line and at-line measurements are scarce and often subject to low measurement
frequencies and large delays between measurement time and evaluation of the results (Pauk et al.). Furthermore, the necessary
low protein concentrations present during refolding processes can result in measurements near the limit of quantification or with
low signal-to-noise ratio (Igwe et al.). Thus, refolding often runs completely unobserved or only partially observed resulting in
low yields and varying outcomes. However, to achieve optimization of refolding processes and consequently improving yields
and reducing process times requires knowledge and process understanding, which could come in the form of advanced PAT,
process models or a combination thereof.

Several different strategies to monitor refolding processes have been described recently: Walther et al. used ATR-FTIR for
optimal process endpoint detection in a stirred-tank-reactor (STR). Although, this method proved to monitor the formation of the
native protein in a batch refolding process, changing background signals including increasing denaturant concentration due
to continuous carry-over from the reservoir of the solubilized protein make this approach challenging in fed-batch applications.
Pathak et al. used various PATs including bioassays, chromatographic and spectroscopic methods to monitor the formation of
secondary, tertiary and quaternary protein structures. This mixture of analysis tools resulted in accurate descriptions of the
refolding kinetics, but analysis times of these methods of up to 2 hours make this approach unfeasible for real-time applications.
Pizarro et al. were able to correlate product quality to dissolved oxygen (DO) and oxidation-reduction potential (ORP) levels.
By measuring DO and ORP inline they were able to choose a correct process endpoint and quenched further refolding reactions
by sparging of nitrogen. Thus, they could reliably refold with constant product quality even when using feedstocks with varying
concentrations. Hebbi et al. used conventional PAT such as pH, ORP or temperature for statistical process control and monitoring
to develop a statistical model of all unit operations starting from buffer preparation.

Although not used in real-time applications, mechanistic descriptions of refolding processes have been formulated based on
a set of differential equations describing the competitive formation of native and aggregated protein from solubilized protein with
an average aggregation number (Kiefhaber et al., Hevehan and De Bernardes Clark). Extensions included detailed descriptions
of the formation of aggregates (Cleland et al., Speed et al., Buswell and Middelberg, Freydel et al.), addition of co-factors (Cao
and Li) or misfolded proteins consisting of incorrectly connected disulfide bonds (Rys et al.). Furthermore, the equations from the
basic model have been transferred from batch to fed-batch mode (Buswell and Middelberg, Dong et al.).
Usage of mechanistic models in a state-observer for process monitoring observability of a process model needs to be proven. Observability of a model refers to the ability to infer the full internal state of a system from a measurement (sub-)set. Local observability of non-linear models can be calculated analog to linear observability by checking for full rank of the observability matrix. To construct the observability matrix of a non-linear model, Lie derivatives are used (Hermann and Krener; Villaverde). Further analysis of a mechanistic model involves sensitivity of model parameter and their identifiability. Sensitivity is a measure on how sensitive the model output reacts to changes in the model parameter value. Parameters with high sensitivity can have a strong impact on the model, whereas parameters with very low sensitivity could take many different values without changing the models output. The local sensitivity of a model can be numerically approximated by simulating the model and changing one model parameter at a time for a small amount (Daume et al.). If the model is differentiable, the sensitivity matrix can be constructed as a jacobian. The identifiability of a process model is a measure to assess if all model parameters are theoretically identifiable from the model structure without noise (structural identifiability) or from the noisy measurements (practical identifiability). Structural identifiability is proven if the sensitivity matrix has full column rank. Quantification of practical identifiability usually revolves around calculation of covariances, correlations and confidence intervals of the calibrated system based on the fisher information matrix (FIM).

Despite the availability of mechanistic descriptions of protein refolding reactions, to our knowledge, no model in combination with available PAT has been used for full-state monitoring in real-time for this application (Pauk et al.). Therefore, the goal was to capture process knowledge in the form of mechanistic models for enhanced monitoring of a refolding process. Augmentation of the model parameters into the state vector can help to reduce model mismatch due to the re-estimation of model parameters in real-time. Not only can successful state-estimation lead to decreased need of potentially expensive PAT, but also to available information of the protein folding states in real-time even if measurement frequencies are low and subject to large time delays. Thus, the availability of this information is the first step towards optimal dynamic control of protein refolding processes. This contribution is organized as follows. In the section "Material and Methods" an overview is given of experimental set-up, analytical methods, process and measurement models, model analysis tools and the observer design. In the "Results" section the model is calibrated, analysed and its applicability in a PF optimized and verified using one tuning and two validation experiments.

2 MATERIAL AND METHODS

2.1 Solubilization and refolding

Frozen IBs containing L-Lactate dehydrogenase 1 (LDH), originating from fermentation of recombinant E. coli as described in Igwe et al., were thawed at room temperature and dissolved in solubilization buffer (4 M GuHCl, 150 mM phosphate buffer, pH 6.0) to a final concentration of 100 g wet IB L⁻¹. After incubation for 2 hours at room temperature and under slight agitation, the mixture was centrifugated (20000 x g, 10 min, 4 °C) and the supernatant was stored on ice until further use in the refolding process.

The refolding was performed via fed-batch dilution in a 3.6 L Labfors 5 stirred tank reactor (Infors AG, Bottmingen, Switzerland). The vessel was filled with an initial volume of 0.8 L of refolding buffer (150 mM phosphate buffer, 1 mM EDTA, 20 µM Nicotinamide adenine dinucleotide, pH 6.0) and the solubilized protein was fed into the reactor at a constant feed-rate via a Preciflow peristaltic pump (Lambda Laboratory Instruments, Brno, Czech Republic). Exact feed-rates and concentration of the three performed experiments can be found in table I. Conditions in the reactor were kept constant by controlling the pH at 6.0 by addition of 0.2 M HCl and 0.5 M NaOH, agitation at 200 rpm with 2 Rushton impellers and head-space aeration with pressurized air at 1 vvm. The temperature was controlled to 10 °C using a Lauda Alpha R8 thermostat (Lauda, Königshofen, Germany). Samples for at-line analysis were taken every hour and centrifugated (20000 x g, 10 min, 4 °C) prior to analysis. A fraction of the prepared samples were used for off-line analysis.

2.2 Analytical methods

2.2.1 Off-line methods

Off-line measurements of the total soluble protein fraction were performed via reversed-phase high-performance-liquid-chromatography (RP-HPLC) on a Dionex Ultimate 3000 system (Thermo Fisher, Waltham, USA) using a Polyphenyl
TABLE 1  Experimental conditions and respective usage of the three experiments.

<table>
<thead>
<tr>
<th></th>
<th>$F_R$</th>
<th>$c_{SR}$</th>
<th>$c_{DR}$</th>
<th>Duration</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mL h$^{-1}$</td>
<td>g L$^{-1}$</td>
<td>mol L$^{-1}$</td>
<td>h</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>5.26</td>
<td>11.57</td>
<td>4.00</td>
<td>24</td>
<td>Tuning</td>
</tr>
<tr>
<td>P2</td>
<td>5.61</td>
<td>10.75</td>
<td>4.00</td>
<td>12</td>
<td>Verification</td>
</tr>
<tr>
<td>P3</td>
<td>6.51</td>
<td>10.75</td>
<td>4.00</td>
<td>12</td>
<td>Verification</td>
</tr>
</tbody>
</table>

Abbreviations: $F_R$, feed-rate; $c_{SR}$, concentration of solubilized protein in the reservoir; $c_{DR}$ concentration of denaturant in the reservoir; P1-3, processes 1 to 3

BioResolve-RP-mAb 2.7 μm 3.0 x 100 mm column (Waters Corporation, Milford, USA). The mobile phase consisted of two lines, one with Milli-Q® water (A), the other with acetonitrile (B) and both lines were supplemented with 0.1 % (v/v) trifluoroacetic acid (TFA). 8 μL of sample was separated by gradient elution from 25 % B to 75 % B for 10.4 minutes at a column temperature of 70 °C and the absorbance was monitored at wavelengths of 214 nm and 280 nm.

2.2.2  At-line methods

At-line measurements of the folding intermediates and native protein was performed via size-exclusion high-performance-liquid-chromatography (SE-HPLC) on a Dionex UltiMate 3000 system (Thermo Fisher, Waltham, USA) using a BEH 200A SEC 1.7 μm, 4.6 x 300 mm, 3.5 μm column (Waters Corporation, Milford, USA). 2 μL of sample were injected at a flow-rate of 0.5 mL min$^{-1}$. Isocratic elution (80 mM phosphate buffer, 250 mM KCl, pH 6.8) was performed at a runtime of 18 minutes, the column oven was controlled at 25 °C and absorbance was measured at wavelengths of 214 nm and 280 nm.

2.2.3  On-line methods

The reservoir containing the solubilized protein was placed on a balance (Sartorius, Göttingen, Germany) and weight measurements were logged every 12 seconds via connection to the process control system Lucullus® (Securecell AG, Urdorf, Switzerland). The feed-rate $F_R$ was calculated by differentiation of the smoothed balance signal. From $F_R$ the volume $V_L$, denaturant concentration $c_{DL}$ and total protein concentration $c_{PL}$ can be calculated.

2.3  Process & measurement models

2.3.1  Mechanistic refolding model

The mechanistic process model was implemented as described earlier (Dong et al.; Pauk et al.) according to the reaction scheme (see figure 2 a, equations 1 - 5). The solubilized protein $c_{SL}$ immediately starts to fold upon entering the mild conditions in the refolding reactor, thus the solubilized protein state was omitted from the model as it is directly transitioned to the folding intermediate $c_{IL}$.

$$ \frac{dc_{IL}}{dt} = -(k_n \cdot c_{IL} + k_a \cdot c_{IL}^2) + \frac{F_R}{V_L} \cdot (c_{SR} - c_{IL}) $$

with $c_{IL}$ the concentration of folding intermediates, $F_R$ the feed-rate of solubilized protein into the refolding vessel, $V_L$ the reactor volume, $c_{SR}$ the concentration of the solubilized protein in the reservoir and the two reaction rates $k_n$ and $k_a$ for the 1st order refolding and second order aggregation reaction, respectively. The folding of the protein to either native product $c_{NL}$ and aggregation $c_{AL}$ happens according to equations 2 and 3.

$$ \frac{dc_{NL}}{dt} = k_n \cdot c_{IL} - \frac{F_R}{V_L} \cdot c_{NL} $$

$$ \frac{dc_{AL}}{dt} = k_a \cdot c_{IL}^2 - \frac{F_R}{V_L} \cdot c_{AL} $$
Due to the fed-batch refolding, changing concentrations of the denaturant and the volume has to be accounted for.

\[
\frac{dc_{DL}}{dt} = \frac{F_R}{V_L} \cdot (c_{DR} - c_{DL}) \tag{4}
\]

\[
\frac{dV_L}{dt} = F_R \tag{5}
\]

with \(c_{DR}\) being the denaturant concentration in the feed. Changes of the volume due to sampling was neglected in the volume calculation. The reaction rates \(k_n\) for refolding and \(k_a\) for aggregation as proposed by Hevehan and De Bernardez Clark have been adopted, describing both reaction rates as functions of the denaturant concentration,

\[
k_n(t) = a_n \cdot (1 + c_{DL}(t))^{b_n} \tag{6}
\]

\[
k_a(t) = a_a \cdot (1 + c_{DL}(t))^{b_a} \tag{7}
\]

with model parameters \(a_n, b_n, a_a\) and \(b_a\). Values for both rates decrease as \(c_{DL}\) increases, due to negative parameters in the exponents (\(b_n, b_a\), see table 3).

### 2.3.2 Augmented observer model

The state vector \(x\) (equations 1-5) is augmented by the model parameters to include parameter estimation into the PF (Kitagawa, Kager et al., Sinner et al., a).

\[
\tilde{x} = (x, \theta)^T \tag{8}
\]

The augmented model parameters have zero dynamics and propagate via a random walk model with independent white noise \(\nu\).

\[
\frac{d\theta}{dt} = 0 + \nu \tag{9}
\]

The extended state vector is composed as,

\[
\tilde{x} = (c_{IL}, c_{NL}, c_{AL}, c_{DL}, V_L, a_n, a_a, b_n, b_a)^T \tag{10}
\]

### 2.3.3 Measurement model

The measurement model consists of at-line measurements for the states \(c_{IL}\) and \(c_{NL}\) from SE-HPLC (see 2.2.2),

\[
y_{at} = \begin{cases} c_{IL} \\ c_{NL} \end{cases} \tag{11}
\]

and on-line measurements for volume \(V_L\), the denaturant concentration \(c_{DL}\) and the total protein concentration \(c_{PL}\) with \(F_R\) derived from the balance signal of the feed reservoir,

\[
y_{on} = \begin{cases} V_L = V_L(t_{k-1}) + F_R(t_k) \cdot t_s \\ c_{DL} = c_{DL}(t_{k-1}) + \frac{F_R(t_k) \cdot c_{DR} \cdot t_s}{V_L(t_k)} \\ c_{PL} = c_{IL} + c_{NL} + c_{AL} \\ = c_{PL}(t_{k-1}) + \frac{F_R(t_k) \cdot c_{SR} \cdot t_s}{V_L(t_k)} \end{cases} \tag{12}
\]

with \(t_s = t_k - t_{k-1}\) the calculation time of the PF. The measurement vector summarizes as,

\[
y = (c_{IL}, c_{NL}, V_L, c_{DL}, c_{PL})^T \tag{13}
\]

### 2.4 Model analysis

#### 2.4.1 Model calibration

The model (equations 1-7) is calibrated on experimental process data by finding the optimal parameter set \(\hat{\theta}\) that minimizes the weighted residual sum of squares (WRSS).

\[
WRSS(\theta) = \sum_{p=1}^{m} \sum_{k=1}^{d} \frac{(y_{p,k} - x_{p,k})^2}{\sigma^2_p} \tag{14}
\]
Here, the squared differences of the measured concentration $y_{p,k}$ and the simulated concentration $x_{p,k}$, weighted by the measurement variances $\sigma_p^2$, is summed for each state $p$ at measurement time $k$. $m$ is the number of states and $d$ is the number of measurement times. Global optimization is achieved by using the genetic algorithm from the global optimization toolbox in Matlab (2020a, Mathworks, Natick, USA). The goodness of fit is evaluated by the normalized root mean squared error (NRMSE),

$$\text{NRMSE}(x_p) = \sqrt{\frac{1}{d} \sum_{k=1}^{d} \frac{(y_{p,k} - x_{p,k})^2}{\max(y_p) - \min(y_p)}}$$

(15)
giving a measure for the error of a single state of the fitted model (Daume et al.).

### 2.4.2 Sensitivity

The local sensitivity of the model states is computed as a 3D matrix $S$ with the jacobian of the model parameters. Since the model output is time dependent, the sensitivity is calculated for all measurement time steps $k$.

$$S = \begin{bmatrix} \frac{\partial c_1}{\partial \theta_1} & \frac{\partial c_1}{\partial \theta_2} & \cdots & \frac{\partial c_1}{\partial \theta_n} \\ \frac{\partial c_2}{\partial \theta_1} & \frac{\partial c_2}{\partial \theta_2} & \cdots & \frac{\partial c_2}{\partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_m}{\partial \theta_1} & \frac{\partial c_m}{\partial \theta_2} & \cdots & \frac{\partial c_m}{\partial \theta_n} \end{bmatrix}$$

(16)

To quantify the influence of each model parameter $j$ on the model output Brun et al. suggested the parameter importance index (equation [17]). By stacking the sensitivity matrix in the time-dimension we can calculate the ranking of the parameter importance as follows:

$$\delta_j^{\text{msqr}} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} S_{ij}^2}$$

(17)

with $n = m \cdot d$ the combined number of measurement times for all model states.

### 2.4.3 Identifiability and parameter errors

Structural identifiability of the full model was proven by verifying full rank of the stacked sensitivity matrix. To assess practical identifiability, first the sensitivity matrix is summed up over the measured time-points,

$$S_{\text{sum}} = \sum_{k=1}^{d} S_k$$

(18)

and multiplied by the measurement covariance matrix $W$ under the assumption of independent and normally distributed measurement errors to obtain the Fisher Information Matrix.

$$FIM = S_{\text{sum}}^T W S_{\text{sum}}$$

(19)

The minimum parameter covariance matrix is calculated based on the Cramer-Rao theorem, as the inverse of the FIM.

$$COV_{\text{min}}(\hat{\theta}) = FIM^{-1}$$

(20)

The diagonal entries of the minimum covariance matrix are used to calculate the $0.68\%$ confidence intervals for the estimated model parameters based on a $t$ distribution.

$$\hat{\theta}_{0.68} = \hat{\theta} \pm \text{diag}(COV_{\text{min}}(\hat{\theta})) \cdot t_{\text{dist}}$$

(21)
An alternative measure for the correlation of model parameters was proposed by Brun et al. analyzing the collinearity index \( \gamma_h \) of the full or subset of the model parameters to discriminate whether all model parameters or which parameter subset can be simultaneously identified.

\[
\gamma_h = \frac{1}{\sqrt{\lambda_h}}
\]  

(22)

where \( \lambda_h \) is the smallest eigenvalue of the subset \( h \) of the FIM.

2.4.4 Observability

The combined observability and identifiability of the augmented process model given the available measurement subset (equation 13) was calculated using the STRIKE-GOLDD toolbox for Matlab on the basis of Lie derivatives (Villaverde et al.).

2.5 Observer design

The implemented Bayesian state-observer is a PF based on the bootstrap or SIR filter proposed by Gordon et al. The numerical implementation of the filter is based on the detailed description of Simon.

For the filter a discretized form of a non-linear process model with additive, normal distributed and independent process \( w_k \) and measurement \( v_k \) noise was used (Wigren et al.):

\[
\begin{align*}
    x_{k+1} &= f_k(x_k, u_k, \theta_k) + w_k \\
    y_k &= g_k(x_k) + v_k
\end{align*}
\]

(23)

with the state vector \( x \) described by the time-discrete and non-linear function \( f_k \) with parameters \( \theta_k \) and time varying inputs \( u_k \) and the measurement model \( y_k \) as a function \( g_k \) of the system states. The PF algorithm consists of an initialization step (i) and four sequentially repeating steps: prediction (ii), update (iii), resampling (iv) and estimation (v).

i) Initialization: During the initialization of the filter, the initial pdf \( p(x_0) \) of the particles

\[
x_i^0, \quad i = 1, \ldots, N
\]

(24)

with \( N \) particles, is drawn from a known distribution to define the initial state.

ii) Prediction: All particles are propagated based on the non-linear process model \( f_k \) and process noise \( w_k \).

\[
\bar{x}_{k,i} = f_{k-1} (x_{k-1,i}, u_k, \theta_k) + w_{k-1},
\]

\[
i = 1, \ldots, N
\]

(25)

In terms of bayesian estimation \( \bar{x}_{k,i} \) is called the prior.

iii) Update: The relative likelihood of each particle defined as an importance weight \( q_{k,i} \) is calculated by evaluation of the conditional \( p(y_k | \bar{x}_{k,i}) \) using the measurement model \( g_k \) and measurement noise \( v_k \). Since \( p(y_k | \bar{x}_{k,i}) \) is only proportionally estimated given the distribution - usually a normal distribution - of the measurement noise, all relative likelihoods are subsequently normalized.

\[
q_{k,i} = \frac{\tilde{q}_{k,i}}{\sum_{j=1}^{N} \tilde{q}_{k,j}}, \quad i = 1, \ldots, N
\]

(26)

iv) Resampling: To prevent filter degeneracy, the particles are resampled based on their computed likelihoods \( q_{k,i} \), yielding the posterior \( x_{k,i} \). Resampling can be performed at each filter iteration or after a certain threshold of effective particles \( N_{\text{eff}} \) is reached (Arulampalam et al.). Multiple resampling strategies are available but systematic resampling was shown to be superior (Guo et al.). Thus it was chosen as resampling strategy in this study when surpassing a threshold of \( N_{\text{eff}} < 0.5N \) with

\[
N_{\text{eff}} = \frac{1}{\sum_{j=1}^{N} q_{k,j}^2}
\]

(27)

v) Estimation: Finally, the posterior \( x_{k,i} \) and corresponding normalized weights \( q_{k,i} \) can be used to calculate measures of the pdf such as a weighted mean and weighted variance, assuming a unimodal distribution.
2.6 Process, parameter and measurement noise

The gaussian process noise \( w \) with zero mean and covariance matrix \( \Sigma_{w,k} \) is added to the augmented model parameters after each update step, resulting in a random walk of the model parameters (Kitagawa). The covariance matrix consists of zeros off the diagonal and variances for the states of the augmented process model on the diagonal.

\[
\Sigma_{w,k} = \text{diag} \left( 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, \sigma_a^2, \sigma_b^2, \sigma_a^2, \sigma_b^2 \right)
\]

No process noise was added to the model states to prevent cross-correlation and because the propagation of the noise through the augmented model parameters and reaction kinetics was shown to be sufficient (Müller et al.). The process noise of the model parameters is calculated as the variance from a 68 % confidence interval (see equation 21). Furthermore, additional noise is introduced to the system by the on-line calculation of the feed-rate from noisy balance signals.

Exact values for the model parameters and model parameter noise as well as mean and covariance of the initial distribution can be found in section 3.2. Changes to the sign of the model parameters are physically not plausible and lead to numerical instabilities. Thus, the model parameters are constrained after the noise addition and set to zero if the constrain is violated:

\[
a_{n}, a_{a} \geq 0, b_{n}, b_{a} \leq 0
\]

The measurement noise \( v \) is assumed to be independent and gaussian with zero mean and diagonal covariance matrix \( \Sigma_{v,k} \), calculated from the propagated error of the measurement equipment as described by Igwe et al.. The noise of the measurements is time variant, as it is dependent on the measured value (equations 28 to 32). Exponentially decaying variance is assumed for the at-line measurements of the folding intermediates \( c_{IL} \) and native protein \( c_{NL} \) and linear dependencies are assumed for the on-line measurement of volume \( V_L \), denaturant \( c_{DL} \) and total protein concentration \( c_{PL} \) based on feed balancing (Igwe et al.).

\[
\begin{align*}
\Sigma_{I,k} &= 0.323 \cdot e^{-16.00 \cdot c_{IL}(t_k)} + 0.0154 \\
\Sigma_{N,k} &= 0.323 \cdot e^{-16.00 \cdot c_{NL}(t_k)} + 0.0154 \\
\Sigma_{V,k} &= 0.0186 \cdot V_L(t_k) \\
\Sigma_{D,k} &= 0.0186 \cdot c_{DL}(t_k) \\
\Sigma_{P,k} &= 0.0186 \cdot c_{PL}(t_k)
\end{align*}
\]

Performance of the state-observer is quantified by the NRMSE (eq. 15 of the states \( c_{IL}, c_{NL} \) and \( c_{AL} \) based on the at-line and off-line measurements, calculated for both the prior and posterior distribution.

2.7 Real-time architecture

A schematic overview of the real-time architecture for the monitoring of the protein refolding process can be seen in figure 1. The state-observer is running on a network device using Python (Version 3.10). Connection between the state-observer and the refolding process and data evaluation is established via RestAPI. On-line measurements of the feed balance are available every 12 seconds. Estimation of the feed-rate from balance signals is performed on every iteration interval of the state-observer \( (t_{s} = 30 \text{ seconds}) \). At-line measurements are available every 60 minutes and results are delayed by 30 minutes due to sample preparation, the SE-HPLC method run-time and subsequent evaluation. Without new measurements, the filter performs only the prediction and estimation steps. Upon arrival of the new measurements, the particle distribution is updated and recalculation occurs from the time of measurement according to the simple rerun PF described by Oreshkin et al. because storage and computational cost were not limiting and minimal error was desired.
FIGURE 1  Schematic overview of the implemented real-time architecture. Experimental setup with constant feed of solubilized protein in the refolding vessel, at-line SE-HPLC measurements every 60 minutes with 30 minute delay for the measurement result. Logging of the reservoir balance signal every 12 seconds and on-line calculation of feed-rate, volume, denaturant and total protein concentration every 30 seconds. The particle filter (PF) as bayesian state-observer predicts the current most probable state $x_{k+1}$ using the current state vector $x_k$ and input $u_k$ with a sampling time of 30 seconds. $\hat{x}_k$ is the predicted state before the update step with $x_{k+1} = \hat{x}_k$ if no new measurement is available. If a new measurements vector $y_k$ is available, the PF updates the particle distribution and resamples if necessary. Recalculation is performed to compensate for the time delay between sample time and the availability of the measurement result. Communication and data-storage in a process and data management system is carried out via RestAPI.

3 | RESULTS AND DISCUSSION

3.1 Model calibration and analysis

For model calibration a fed-batch refolding experiment was carried out (Experiment P1, see table 1 ).

First, the model was fit to the dataset estimating all four model parameters together. Subsequently, local sensitivity, identifiability and parameter correlation was analyzed. The linear correlation of model parameter subsets was assessed using the parameter collinearity index $\gamma_h$ (equation 22) for all possible sub-sets with at least two parameters and the full parameter set. The resulting collinearity indices are presented in table 2. Most parameter sub-sets have an index value of below 20, a threshold defined by Brun et al. for a tolerable collinearity. However, the full set has a collinearity index above 36, suggesting that not all model parameters should be estimated at the same time. Subsequently, the model was analyzed for the importance of each parameter on the model output using the parameter importance ranking (equation 17). As it can be seen in figure 2 b the parameter $a_h$ has only a small impact on the overall model output. Thus, the subset consisting of the model parameters $a_n$, $b_n$ and $b_a$ with a $\gamma_h$ value of 6.9758 was used for model calibration, while $a_h$ was fixed to the previously estimated value from the full data set. Subsequently the three parameters were fixed with the identified values and $a_h$ was refitted on the calibration dataset. As described by Brun et al., using a fixed value from a previous estimation of a collinear model parameter set is inherently collinear. However, setting a true valid value for such a fixed parameter is difficult, especially if the system is insufficiently characterized. Even after repeating the above procedure for three full cycles with alternating start values, changes to the identified model parameters were insignificant (data not shown).

The resulting parameters can be found in table 3 and the model fit as feed-forward simulation in figure 4, blue solid line, with state errors (NRMSE, see equation 15) of 6.37 %, 4.31 % and 12.33 % for the protein states $c_{IL}$, $c_{NL}$ and $c_{AL}$, respectively. The standard deviation was calculated from the 68 % confidence intervals as indicated in section 2.4.3 and are given in table.
Basic protein refolding model with states for solubilized protein ($c_{SL}$), folding intermediates ($c_{IL}$), native protein ($c_{NL}$), aggregated protein ($c_{AL}$) and the respective reaction rates. Simplification of the model by omitting $c_{SL}$ and reaction rate $k_i$.

**FIGURE 2** Schematic refolding model with reaction rates, model parameters the model parameter importance ranking.

**TABLE 2** Collinearity index of all possible model parameter subsets $h$ with at least two parameters including the full set.

<table>
<thead>
<tr>
<th>Subset</th>
<th>$\gamma_h$</th>
<th>Subset</th>
<th>$\gamma_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_n, b_n$</td>
<td>5.3121</td>
<td>$a_n, b_n, a_a$</td>
<td>5.8865</td>
</tr>
<tr>
<td>$a_n, a_a$</td>
<td>5.5867</td>
<td>$a_n, b_n, b_a$</td>
<td>6.9758</td>
</tr>
<tr>
<td>$a_n, b_a$</td>
<td>3.9280</td>
<td>$a_n, a_a, b_a$</td>
<td>9.0080</td>
</tr>
<tr>
<td>$b_n, a_a$</td>
<td>4.5777</td>
<td>$b_n, a_a, b_a$</td>
<td>7.9346</td>
</tr>
<tr>
<td>$b_n, b_a$</td>
<td>5.9273</td>
<td>$a_n, b_n, a_a, b_a$</td>
<td>36.1599</td>
</tr>
<tr>
<td>$a_a, b_a$</td>
<td>7.0151</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: $\gamma_h$, collinearity index

It can be noted that the relative error for parameter $a_n$ is around 119% and for $a_a$ around 94% while the errors of $b_n$ & $b_a$ are lower at 8% & 15%, respectively. One of the highly uncertain parameter $a_a$ has a low sensitivity on the model states and therefore not a big influence on them. The other $a_n$ has both, a high sensitivity and uncertainty and is therefore a potential weak-point of model applicability.

Overall the model captures the refolding dynamics with acceptable accuracy, especially in the state of the native protein. This is consistent with earlier work, where Hevehan and De Bernardez Clark and Dong et al. achieved comparably good fits of the native protein. Direct comparison of the model parameter values is skewed, because both authors used high denaturant & protein concentrations and assumed an aggregation reaction of third order. Thus, their $a_n$ and especially $a_a$ have far higher values.

To use the model for real-time monitoring, its observability has to be assessed. Based on the available online and at-line measurement information consisting of native protein, folding intermediates, volume, denaturant concentration and total protein concentration, the combined observability and identifiability was determined as described in section 2.4.4. Both the observability as well as the identifiability matrix have full rank suggesting that with the available measurement information all states can be reconstructed and model parameters re-identified.

### 3.2 Observer tuning

The calibration data-set (P1) was used for the tuning of the PF. Simulations were performed with constant iteration time $t_i$ of 30 seconds for 24 hours. To prevent further collinearities during on-line state-estimation, noise was only added to the identifiable parameter sub-set $a_n, b_n$ and $b_a$ with the highest sensitivities on the model states. The PF as described in section 2.5 was initialized with $N$ particles drawn from a normal distribution with initial condition

$$\mu_{0} = (0, 0, 0, 0.8, 1.33, -8.68, 12.05, -16.79)^T$$
TABLE 3 Mean and standard deviation for model parameter values from the model calibration of experiment P1.

<table>
<thead>
<tr>
<th>$a_n$</th>
<th>$b_n$</th>
<th>$a_\alpha$</th>
<th>$b_\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>h$^{-1}$</td>
<td>-</td>
<td>Lg$^{-1}$h$^{-1}$</td>
<td>-</td>
</tr>
<tr>
<td>Value</td>
<td>1.3343</td>
<td>-8.6824</td>
<td>12.0465</td>
</tr>
<tr>
<td>SD ±</td>
<td>1.5837</td>
<td>0.7182</td>
<td>11.3641</td>
</tr>
</tbody>
</table>

Abbreviations: SD, standard deviation

and diagonal covariance matrix for initialisation

$$\Sigma_{w0} = \text{diag} \left( 10^{-6}, 10^{-6}, 10^{-6}, 10^{-6}, 10^{-3}, 2.5080, 0.5159, 0.0, 6.6129 \right)$$

The variances of the protein states and the denaturant were assumed to be small, because the starting buffer in the refolding vessel was prepared protein and denaturant free while the variance of the reactor volume was assumed to be higher due to greater inaccuracy during the reactor setup. The variances for the model parameters were taken from the model calibration step except for $a_n$ which was excluded from propagation through the augmented model by setting the noise to 0.

The process noise added during the update steps $k$ was the same variance used for the initial distribution of the model parameters without any additional noise on the system states ($c_{IL}, c_{NL}, c_{AL}, c_{DL}$, & $V_L$). Thus, the diagonal process noise covariance matrix was:

$$\Sigma_{w,k} = \text{diag} \left( 0.0, 0.0, 0.0, 0.0, 0.0, 2.5080, 0.5159, 0.0, 6.6129 \right)$$

To tune the PF, the effect of the number of particles $N$ on the filter performance (NRMSE) as well as the mean runtime was analyzed. The filter was run as five-fold simulations for particle numbers of $N \in \{50, 250, 500, 750, 1000, 1500, 2000\}$. Equal random seeds ($\mu_0$ & $\Sigma_{w0}$) were used for the simulations of each particle number and iteration to allow comparability. Figure 3 shows that with increasing number of particles, the NRMSEs decrease until they stagnate at around 750 - 1000 particles.

The highest error originates from the not directly measured aggregated protein, thus, its NRMSE is not stable and shows high standard deviations obtained from the five-fold simulation (figure 3, blue circles). This is, because the aggregated protein is not part of the at-line measurement vector, and thus fluctuations in the measurements have no impact on the filter performance. When looking at the summed NRMSE of the folding intermediate and native protein - the protein states with which the filter is updated (figure 3, orange squares) - a clear decaying behaviour with decreasing standard deviation is visible. This is expected, because at higher particle numbers, it is more likely to have one or more particles in close proximity to the incoming measured value, and thus resulting in lower error values. Consequently, the filter gets more robust with increasing number of particles until the filter can not describe the uncertain biological system any better and a residual error remains. The overall runtime (3, green triangles) of the filter increases linearly with the number of particles. As no more improvement is gained beyond 1000 particles at the cost of higher computational burden, $N = 1000$ was chosen as the optimum number of particles and used within the rest of this study. The overall runtime of $\approx 19$ minutes translates to a tolerable calculation time of $\approx 400$ ms per filter iteration, which in the real process takes place every 30 seconds.

### 3.3 Particle filter vs. feed-forward simulation

To show the superiority of the PF with optimized settings and simultaneous state and parameter estimation its performance was compared to the feed-forward model simulation of the calibration as well as the two validation experiments.

#### 3.3.1 Tuning experiment

The simulations of P1 are presented in figure 4. As calibrated on the same dataset the model simulation can overall follow the measurements with reasonably low errors (NRMSE). At certain points it can be observed that the simplified model can not accurately represent the underlying dynamics, such as the slight S-shape of $c_{IL}$ (figure 4, A), the increase of $c_{AL}$ towards the end (figure 4, B) as well as fluctuations in $c_{AL}$ between 15 and 20 hours (figure 4, C).
FIGURE 3  Mean NRMSE of the posterior distributions of the aggregated protein \(c_{AL}\) (blue circles) and the sum of the protein states available via at-line measurements (folding intermediates, \(c_{IL}\); native protein, \(c_{NL}\); orange squares) as well as mean total runtime (green triangles) of the PF for different numbers of particles \(N\) including standard deviation from five-fold simulation.

The PF using additional at-line information of \(c_{IL}\) & \(c_{NL}\) and non-static parameters \(a_n\), \(b_n\) and \(b_a\) (figure 4 D, F & G) was able to better follow these trends for the folding intermediate and the native protein. Compared to the simple feed-forward simulation the PF differs the most from the aggregated protein (figure 4 C). The reason for this is a combination of: first the aggregated protein is not part of the at-line measurement vector, and thus fluctuations in the measurements have no impact on the filter performance. Secondly, the measurement of the aggregated protein shows high variance, possibly from unreliable sample preparation or chromatographic evaluation. Third, the volume calculation omits missing volume taken during sampling for the measurements. Although compared to the model simulation the reconstructed aggregate concentration from the PF is underestimating the concentration it better displays the observed fluctuations between 15 and 20 hours.

The trend of the model parameters \(a_n\), \(b_n\) and \(b_a\) (figure 4 D, F & G) shows how the parameters evolve over time to best describe the at-line measured states. The proportional factor of the refolding reaction \(a_n\) increases during the first 5 hours while the penalty factor \(b_n\) stays constant, thus boosting the refolding rate \(k_n\) (figure 4 H) to better represent the strongly rising concentration of the native protein associated with a stronger decrease of \(c_{IL}\). In the following period the formation of native protein is decreased leading to a reduction of \(a_n\) as well as \(b_n\), which also describes the slight rise in \(c_{IL}\). Furthermore, between 5 to 10 hours \(b_a\) rises a small amount, to balance the stagnating formation of the native protein with increased aggregate formation. However, the impact of the changes on the aggregation rate \(k_a\) (figure 4 I) overall are very limited, because the high negative value of \(b_a\) penalizes the aggregation rate heavily with rising denaturant concentration. During the next five hours (10-15 h) the concentration of the folding intermediates rises stronger despite constant feed-rates for the whole experiment. Thus, both penalizing parameters \(b_n\) and \(b_a\) decrease their values to reflect the accumulation of the folding intermediates. The sharp outlier in \(b_a\) after about 17 hours can be explained by quite a substantial resampling step, possibly caused by the disturbance in the feed balance signal. Furthermore, the denaturant concentration approaches 0.4 mol L\(^{-1}\) at this point, thus the aggregation rate approaches 0 and changes in \(b_a\) have very little effect on the rate. In the last part, the native protein concentration starts to increase again. This rise is reflected by sharply rising values for \(a_n\) & \(b_n\) leading to a visible increase in the refolding rate \(k_n\).

3.3.2 Validation experiments

Two fed-batch refolding experiments (P2, P3) were performed to validate the established PF and show that it can be applied to new refolding processes (see table 1). The PF was initialized analog to the settings of the tuning experiment and the optimal number of \(N = 1000\) particles were used. The estimation results of the PF compared to the feed-forward simulation of the calibrated model can be found in figure 5. Table 4 summarizes the NRMSE values for both experiments and the feed-forward simulations.

In both validation processes it can be seen that the feed-forward model simulation is poor and therefore the model without any real-time correction is not transferable. Similarly, other studies have shown recently that PFs can be used to improve monitoring of a biological process by adapting model parameters in real-time, for example adaptation of metabolic and growth capabilities
FIGURE 4 Weighted means of the protein states (A-C), model parameters (D-G) and reaction rates (H-I) for simulations with a particle filter (N=1000). Measurements for all three protein states are shown, but the off-line measurements of the aggregated protein are not considered in the filter calculation. Comparison of a particle filter with random noise added onto the model parameters (Particle Filter, solid line) and without noise addition (feed-forward simulation, dashed line). Additionally, the 68% confidence interval from the weighted particle variance on model parameters and reaction kinetics (shaded area) and the recalculation steps are shown (dotted lines).

of Penicillium chrysogenum, Corynebacterium glutamicum or E. coli (Kager et al.; Sinner et al.; Müller et al.). The PF with adaptable model parameters shows a better agreement on both experiments. It can be observed that the PF effectively corrects the S-shape of the folding intermediates and overestimated native protein concentration. Furthermore, the fit of the aggregated protein is improved. The change of the variable model parameters $a_n$, $b_n$ & $b_a$ is represented in figure 6 A, B & C, while $a_a$ was kept constant. During the first four hours, $a_n$ increases sharply in both experiments while $b_n$ stays stable to boost the refolding reaction rate $k_n$, allowing the filter to track the rising concentration of the native protein. Subsequently, the native protein concentration stagnates. The overestimation of the model is compensated by a decrease of $a_n$ as well as $b_n$ to reduce the refolding rate $k_n$. The slight delay of the stagnation phase comparing P2 and P3 is also represented in the parameter trajectory. Furthermore, the accumulation of the folding intermediates facilitates a decrease of $b_a$ and thus $k_a$ after 6 hours. However, due to the already small value of $k_a$, this influence is very small. The one outlier in $b_a$ is propagated from a potential outlier.
FIGURE 5  Weighted means of the folding intermediates $c_{IL}$, the native protein $c_{NL}$ and the aggregated protein $c_{AL}$ for both validation experiments (P2; A,C,E) and (P3; B,D,F) with a particle filter ($N=1000$, solid line) in comparison with the corresponding feed-forward simulation (dashed line). Measurements for all three protein states are shown, but the off-line measurements of the aggregated protein are not considered in the filter calculation. Additionally, the recalculation steps are shown (dotted lines).

measurements of $c_{IL}$ at 9 hours in P2 leading to substantial resampling. Due to the robustness of the filter both the state and model parameter trajectories are corrected with the next at-line measurement.

3.3.3 | Quantification of the filter performances

For better quantification, the NRMSE values of the PF and the feed-forward simulation were calculated and presented in table 4. The prior is defined as the weighted mean of the filter before the distribution is updated at the time of measurement (orange lines in figures 4 and 5). The posterior refers to the weighted mean after the distribution is updated at the time of measurement (start of the black dotted lines in figures 4 and 5).

The calculations (table 4) show that the PF of P1 has lower errors in the states $c_{IL}$ and $c_{NL}$ compared to the feed-forward simulation. Due to resampling the posterior distributions have an even lower estimation error leading to reduction of the NRMSE of 3.7-fold ($c_{IL}$) and 4.1-fold ($c_{NL}$) to total errors of 1.73 % and 1.06 %, respectively. The error of $c_{AL}$ however is almost twice as large compared to the feed-forward simulation and even increases slightly in the posterior distribution. Overall both the pre-calibrated model as well as the PF are able to accurately follow the process P1, especially the most important states of native protein - as it is the product - and the folding intermediates - as they can fold into the native protein. After incorporation of the process noise to the model parameters, the PF can outperform the feed-forward simulation in describing $c_{IL}$ and $c_{NL}$ at the cost of more uncertain predictions of $c_{AL}$. The true superiority of the PF towards the feed-forward simulation is shown when the PF is applied to the validation processes, as the model is transferred without re-calibration of the model parameters. The PF shows lower errors in all states compared to their respective feed-forward simulation. The smallest reduction is seen in $c_{AL}$ of P2 with 42 % and the biggest improvement in $c_{NL}$ of P2 with an error reduction of 93 % (posterior distributions). The corresponding improvement of the prior distributions is in the range of 18 % ($c_{IL}$, P2) and 81 % ($c_{NL}$, P2). The NRMSE in the native protein is the lowest with 4.88 % & 3.44 % for the posterior and 12.55 % & 8.91 % for the prior distributions of P2 &
FIGURE 6 Model parameters (A-C) and reaction rates (D,E) for both validation experiments (P2, P3) with a particle filter (N=1000). Additionally, the 68% confidence interval from the weighted particle variance (shaded area) and the recalculation steps are shown (dotted lines).

P3, respectively. The PF also yields a better representation of the not-directly measured aggregate concentration, but this state remains the most uncertain at 20.45% (P2, posterior) and 16.80% (P3, posterior). Compared to the observer tuning experiment (P1) both validation experiments perform worse. This is expected, because the model and PF were calibrated to P1. Though, the errors and their reduction in the validation experiments remain reasonable and are in the same range as recent studies in bacterial upstream processes Sinner et al. (a); Müller et al.

The calculated errors of the posterior distributions are better compared to the prior distributions. However, to utilize the information provided by the PF in real-time for example for process control, it is necessary to focus on the prior distribution. In the current monitoring system only the prior distribution is visible in real-time as the measurement results are delayed. The posterior distribution is therefore not immediately available but observability of the process model is only given with the combination of on-line and at-line measurements. Analog to Kager et al., it was shown that the PF is able to handle large time delays of the at-line measurements by recalculating from the time where the sample was drawn. The posterior distribution could be used if the measurement delay of the at-line measurements is small, thus leading to a reduced error. However, due to the long time delay, this is not feasibly in this case. The higher error of the prediction is the price for a fully observable protein refolding process model. Thus, focus of further work is necessary to reduce the delay time of the at-line measurements to increase the measurement frequency and in turn reduce the error difference between prior and posterior distribution. Furthermore, the decision to use the model parameter error as process variance to be added on the model parameter subset is the logical choice, because the values originated from the analysis of the model itself including the applied measurement uncertainties (Sinner et al., (a). As described above, the model fit is good, thus this choice worked well in this application, but investigating the effect of the applied error for example originating from monte carlo simulations - on the model parameters might result in an increased model fit as well.

4 | CONCLUSION

In this contribution a concept for full-state estimation of a fed-batch protein refolding process was developed by combining a mechanistic process model with available on-line and time delayed at-line measurements. First, the process model was calibrated to experimental data and the model analysed. Augmentation of the model parameters into the state vector allows the
TABLE 4 Comparison of normalized root mean squared error values of the protein states (folding intermediates, native protein, aggregated protein) for the calibration and validation experiments. The estimation column is divided into the noise free feed-forward simulation (FFS) as well as the prior and posterior distribution of the particle filter.

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Process</th>
<th>Estimation</th>
<th>NRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>folding intermediates</td>
<td>native protein</td>
</tr>
<tr>
<td>Calibration</td>
<td>P1</td>
<td>FFS</td>
<td>6.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prior</td>
<td>3.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Posterior</td>
<td>1.73</td>
</tr>
<tr>
<td>Validation</td>
<td>P2</td>
<td>FFS</td>
<td>14.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prior</td>
<td>11.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Posterior</td>
<td>5.40</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>FFS</td>
<td>12.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prior</td>
<td>8.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Posterior</td>
<td>6.62</td>
</tr>
</tbody>
</table>

Abbreviations: NRMSE, normalized root mean squared error; FFS, feed-forward simulation

re-estimation of the model parameters in real-time by adding random normally distributed noise onto the selected subset. For this purpose, a strategy is proposed to use a combination of the parameter importance ranking and collinearity index to find a suitable parameter subset with tolerable collinearity. Calculated uncertainties from the model analysis were used as variances for the initial distribution and noise addition with the update steps of the PF. Furthermore, to assure observability, the at-line measurements are added to the on-line measurements. Consequently, a recalculation step is necessary to cope with the large time delay of the at-line measurements. The calibration data-set was used to find the optimal number of particles $N$ by calculating NRMSE values of the state-observer, while keeping the mean calculation runtime of the filter in a reasonable range. Finally, the PF was validated with two fed-batch refolding processes. The different protein states of both processes can be followed by the filter with satisfactory accuracy. In all cases the state-observer outperformed the feed-forward model simulation.

The addition of noise on the model parameters allows the filter to adapt to poor model calibration and inherent changes in model parameters, resulting in enhanced performance of the filter and a deeper insight into ongoing reaction kinetics. Most significant improvements were especially visible in the two most important states of the system - the native protein and the folding intermediates - leading to NRMSE values in the range of 3.86 % to 2.54 % for the prior and 1.73 % to 1.06 % for the posterior distribution of these two states in the tuning experiment (P1). Furthermore, it was shown that the filter can be transferred to new refolding processes. Although expectedly worse compared to the tuning experiment, the NRMSE values of the native protein and folding intermediates are in the reasonable range of 8.78 % to 12.55 % for the prior and 3.44 % to 6.62 % for posterior distributions. Furthermore, the PF reduced the NRMSE compared to the feed-forward simulations by up to 5- and 13-fold ($c_{NL}$, P2) for the prior and posterior distribution. The NRMSE values of the posterior distribution were shown to be superior to the prior distribution due to the incorporation of the new at-line measurements, but the high time delay only allows for the prior distribution to be used as input for possible further real-time optimization or control strategies. Nonetheless, the developed particle filter can be seen as the first step towards the closed loop model-control of a protein refolding process, because the necessary requirement of full-state information in real-time is achieved.

To improve on this concept, strategies should be made to reduce the error of the prior distribution. First, one can tackle the delay between the time where the sample was drawn and when the measurement result is available. Thus, the measurement frequency could be increased leading to a lower difference between prior and posterior distribution. Alternatively, to get measurement updates at every filter iteration, other on-line PAT could be explored that is capable of handling changing background signals due to the fed-batch process mode and generally low protein concentrations. Furthermore, adaptations to the
The process model could be made to better reflect the true underlying process and thus intrinsically reduce the NRMSE of the prior distribution.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ACKNOWLEDGMENTS

The authors acknowledge financial support through the COMET Centre CHASE, funded within the COMET Competence Centers for Excellent Technologies programme by the BMK, the BMDW and the Federal Provinces of Upper Austria and Vienna. The COMET programme is managed by the Austrian Research Promotion Agency (FFG). Further financial support was provided by the Novo Nordisk Foundation (Start Package grant NNF220C0081250). The Authors thank DTU Library as member of the Royal Danish Library consortium for their open access funding program. The authors also acknowledge support and guidance from Prof. Dr. Oliver Spadiut of TU Vienna and insights and support provided by the project partner Bilfinger Life Science.

AUTHOR CONTRIBUTIONS

Jan Niklas Pauk: Conceptualization, Experimental execution, Writing – Original draft preparation, review & editing. Chika Linda Igwe: Experimental execution, Writing - Review & editing. Christoph Herwig: Supervision, writing - Review & editing, Funding acquisition. Julian Kager: Conceptualization, Writing - Review & editing, Funding acquisition.

ORCID

Jan Niklas Pauk https://orcid.org/0000-0002-3309-837X
Chika Linda Igwe https://orcid.org/0000-0003-3583-8850
Christoph Herwig https://orcid.org/0000-0003-2314-1458
Julian Kager https://orcid.org/0000-0001-7274-6678

References


Kager, J., Herwig, C., Stelzer, I.V., . State estimation for a penicillin fed-batch process combining particle filtering methods with online and time delayed offline measurements 177, 234–244. URL: https://linkinghub.elsevier.com/retrieve/pii/S0009250917307388 doi:.


Kieffer, T., Rudolph, R., Kohler, H.H., Buchner, J., . Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation 9, 825–829. doi:.


**How to cite this article:** J. N. Pauk, C. L. Igwe, C. Herwig, and J. Kager (2023), Full-state monitoring of protein refolding reactions using particle filters and delayed measurements, *Biotech. Bioeng.*, 2023;00:1–6.