Biotechnology Reports 31 (2021) e00640

Contents lists available at ScienceDirect

Biotechnology Reports

journal homepage: www.elsevier.com/locate/btre

Generic and specific recurrent neural network models: Applications for large and small scale biopharmaceutical upstream processes

Jens Smiatek [a,b,*], Christoph Clemens [c], Liliana Montano Herrera [d], Sabine Arnold [d], Bettina Knapp [e], Beate Presser [e], Alexander Jung [b], Thomas Wucherpfennig [d], Erich Bluhmki [e,f]

[a] Institute for Computational Physics, University of Stuttgart, D-70569 Stuttgart, Germany
[b] Boehringer Ingelheim Pharma GmbH & Co. KG, Digitalization Development Biologicals CMC, D-88397 Biberach (Riss), Germany
[c] Boehringer Ingelheim Pharma GmbH & Co. KG, Focused Factory Drug Substance, D-88397 Biberach (Riss), Germany
[d] Boehringer Ingelheim Pharma GmbH & Co. KG, Bioprocess Development Biologicals, D-88397 Biberach (Riss), Germany
[e] Boehringer Ingelheim Pharma GmbH & Co. KG, Analytical Development Biologicals, D-88397 Biberach (Riss), Germany
[f] University of Applied Sciences Biberach, D-88397 Biberach (Riss), Germany

ARTICLE INFO

Keywords:
Recurrent neural networks
Upstream processes
Generic and specific machine learning models
Temporal evolution
Principal component analysis
Simulation of bioreactor processes
Autocorrelation functions

MSC:
92B20
62M10
37N25
46N60

ABSTRACT

The calculation of temporally varying upstream process outcomes is a challenging task. Over the last years, several parametric, semi-parametric as well as non-parametric approaches were developed to provide reliable estimates for key process parameters. We present generic and product-specific recurrent neural network (RNN) models for the computation and study of growth and metabolite-related upstream process parameters as well as their temporal evolution. Our approach can be used for the control and study of single product-specific large-scale manufacturing runs as well as generic small-scale evaluations for combined processes and products at development stage. The computational results for the product titer as well as various major upstream outcomes in addition to relevant process parameters show a high degree of accuracy when compared to experimental data and, accordingly, a reasonable predictive capability of the RNN models. The calculated values for the root-mean squared errors of prediction are significantly smaller than the experimental standard deviation for the considered process run ensembles, which highlights the broad applicability of our approach. As a specific benefit for platform processes, the generic RNN model is also used to simulate process outcomes for different temperatures in good agreement with experimental results. The high level of accuracy and the straightforward usage of the approach without sophisticated parameterization and recalibration procedures highlight the benefits of the RNN models, which can be regarded as promising alternatives to existing parametric and semi-parametric methods.

1. Introduction

Over the last years, modelling and simulation has become an important field of research for biotherapeutical manufacturing and process development. Due to increasing computational power as well as the improved use of process analytical technologies, novel computational approaches for complex upstream and downstream processes are in the focus of recent interest [1–3]. While mechanistic kinetic-dispersive models are nowadays considered as standard methods for the study of capturing and polishing steps in downstream operations [4–10], there exist a plethora of distinct models for upstream processes with certain advantages and shortcomings. The large number of modelling approaches may be related to the importance of correlated molecular mechanisms at distinct length scales as well as the broad variability of biological parameters among living organisms.

At the largest length and time scales, active pharmaceutical ingredients (APIs) like monoclonal antibodies (mAbs) are produced by distinct cells in bioreactors whose optimal design is nowadays studied by computational fluid dynamics or Lattice-Boltzmann simulations [11–18]. At smaller or even molecular scale, one is usually interested in modelling the cell metabolism, which helps to identify optimal feeding strategies as well as improved process protocols for higher product titers and improved product quality [19]. Specifically often used standard Chinese hamster ovary (CHO) cells show a rather complex cell metabolism [20] in combination with diverse post-translational modification profiles [21], such that the understanding of the cell metabolism in

* Corresponding author at: Institute for Computational Physics, University of Stuttgart, D-70569 Stuttgart, Germany.
E-mail address: smiatek@icp.uni-stuttgart.de (J. Smiatek).

https://doi.org/10.1016/j.btre.2021.e00640
Received 4 February 2021; Received in revised form 24 April 2021; Accepted 27 May 2021
Available online 28 May 2021

terms of high API quality is of fundamental interest.

In addition to detailed metabolic flux pathway models [22–27], often also simpler mechanistic models are used to predict the time-dependent concentration profiles from standard cell cultures [28–35]. The mathematical framework is represented by coupled partial differential equations which may also include the temperature as well as pH values in order to provide a more detailed representation of experimental conditions. Although most models show an overall good agreement with the experimental results, certain systematic deviations are often evident, which can be attributed to an incomplete knowledge of the cell metabolism as well as the use of oversimplified pseudo first-order and Monod reaction kinetics [36]. As a specific example, complex and varying feed strategies in terms of bolus addition are often not reliably reproduced [36]. Thus, certain deviations from experimental outcomes as well as the neglected or simplified influence of intrinsic parameters like temperatures or pH values for mechanistic growth models become evident. Recently, so-called hybrid models were introduced in order to improve process simulations [37,38,36,39–42]. In combination with a mechanistic framework, experimental data are used to derive time-dependent rate constants in combination with relevant process parameters like the temperature and the pH value as well as feeding rates in terms of an artificial neural network approach or other advanced regression techniques [36–38,43–45,3,46,42,41]. Despite slight differences between the approaches, a hybrid model usually extracts the temporally varying rate $\nu_{p,x}(t)$ for a biomass-related parameter x or for the product titer p from an ANN approach, which is then introduced according to

$$\frac{\mathrm{d}}{\mathrm{d}t}x\left(t\right) = \left(\nu_{p,x}\left(t\right) - D\left(t\right)\right)\cdot x\left(t\right) \tag{1}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}p\left(t\right) = \nu_{p,x}\left(t\right)\cdot x\left(t\right)\cdot\theta\left(t\right) - D\left(t\right)\cdot p\left(t\right), \tag{2}$$

where $\theta(t)$ is the Heaviside function, which can be either 0 or 1, dependent on the presence or absence of induction in combination with the dilution factor $D(t)$, which contains information about bolus addition or sampling. The corresponding temporal values for $p(t)$ or $x(t)$ are then calculated by standard numerical integration schemes [43]. Hence, hybrid models are able to reproduce the growth and metabolic rates of fed-batch processes [36,43] in combination with complex feeding strategies. Such a detailed description is not achieved by mechanistic models, however, their benefit for simple predictions of growth parameters even for perfusion processes was recently demonstrated [33,47].

Notably, the determination of rate constants for mechanistic and hybrid models as well as the parameterization of the approaches is still a challenging task. Moreover, it has to be noted that the corresponding mechanistic framework provides a rather coarse-grained picture when compared to more sophisticated metabolic flux pathway models [26]. Hence, the corresponding insights in terms of Eqs. (1) and (2) into growth, death and production behavior are of limited value for more refined considerations due to simplified descriptions as well as unphysical temporal variations of the rate constants. Although hybrid models can be used as a beneficial tool to complement Design of Experiment studies with regard to an adequate exploration of the design space [44,41], it has to be noted that experimental work in terms of initial parameter scans is of essential need. Thus, given the limitations with regard to the complex parameterization procedure in combination with the rather limited insights, it can be assumed that straightforward non-parametric machine learning approaches provide comparable outcomes with less efforts. Moreover, such data-driven methods circumvent the consideration of temporal variations for the rate constants, which is thus in agreement with quasi-equilibrium thermodynamics.

Over the last years, a lot of effort was spent into the development of neural networks or further advanced regression algorithms [48–50] and their application for bioprocess control and prediction [51,52,1,53–55]. Often used approaches are artificial neural networks (ANNs) which can be regarded as highdimensional regression methods for connecting input parameters to target variables [56,48]. ANNs are nowadays widely used in the field of natural sciences, as can be seen by applications ranging from the calculation of molecular properties, prediction of chemical reactions and drug screening [57–64]. Although ANNs are well suited to connect static features, they are often limited for data showing temporal evolution. Promising candidates in this regard are multivariate recurrent neural network (RNN) models [65–67,55], whose benefits for the calculation and simulation of temporal process data in various contexts were recently described [51,55].

In this article, we present specific and generic RNN models for the simulation of multivariate large- and small scale upstream processes. Our approach can be used for the control and study of single product-specific large-scale manufacturing runs (specific RNN model) as well as small-scale evaluations in terms of combined processes for distinct products at development stage (generic RNN model). All RNN calculations rely on experimental data with broad variability. Certain variations at well-defined time points can be attributed to differences in process conditions as well as biological factors. Despite these challenges, our results only show small deviations between calculated and experimental values, which are significantly smaller than the ensemble experimental standard deviation. The main advantages of our method are the straightforward implementation without complex parameterization procedures in combination with a high predictive accuracy. In contrast to hybrid or mechanistic models, the proposed RNN approach can be used without further approximations, pre-defined boundary conditions or knowledge about the underlying metabolic connections. Moreover, the questionable introduction of temporally varying rate constants is avoided. Without further adaption, fully automatized and pre-trained RNN models can also be used by non-experts which promotes their usage for the calculation and simulation of modern biotherapeutical manufacturing and development processes in real time. The results for the platform-dependent generic RNN model approach underpin such assumptions.

The article is organized as follows. In the next section we provide a short introduction into the theoretical background of RNNs. Details about the numerical implementation and the data sets are presented in Section 3. All numerical results are shown in Section 4. We conclude and summarize in the last section.

2. Theoretical background: recurrent neural networks

Over the last years, recurrent neural networks (RNNs) attracted recent interest as promising approaches to process and to evaluate large amounts of temporal sequences [67]. Typical applications of RNNs include speech recognition [68,67] as well as weather, climate and finance forecasting [69–71]. In principle, RNNs can be regarded as a modified version of standard feed-forward ANNs [56,48,64]. The basic network structure is represented by one input layer, one or multiple hidden layers and one output layer with a varying number of nodes in each layer. In contrast to feed-forward ANNs, direct connections between two successive layers of nodes are implemented as recurrent loops. Hence, the RNN is able to process temporal sequences and to predict the evolution of outcomes.

The basic algorithm of an RNN [67,68] includes the consideration of an input interval $\mathbf{x} = (x_1, \ldots, x_T)$ of length T as fed into the nodes of the input layer, the hidden vector $\mathbf{h} = (h_1, \ldots, h_T)$ as calculated in the hidden layers and the final output vector $\mathbf{y} = (y_1, \ldots, y_T)$ where bold letters denote vectors. The following iterative algorithm connects the input sequence to the elements of the hidden vector

$$h_t = \mathcal{H}(\omega_{\mathrm{ih}}x_t + \omega_{\mathrm{hh}}h_{t-1} + b_h) \tag{3}$$

and hence also to the output vector

$$y_t = \omega_{ho} h_t + b_o, \quad (4)$$

respectively, with sequence or time points $t = 1, \ldots, T$, biases b_j and the weights ω_{jl}, where the indices $j, l \in \{i, h, o\}$ denote the corresponding input (i), hidden (h) and output (o) layers. The function $\mathcal{H}(\cdot)$ represents a standard hidden layer activation function like in ANNs which is typically a logistic, hyperbolic tangent or sigmoidal function with a smooth differentiable form [56]. A scheme of a standard RNN is shown in Fig. 1.

Notably, a significant improvement for the stability and accuracy of RNNs was the introduction of the long-short term memory (LSTM) approach [65] which allows the consideration of long times within sequences. The recent interest in modern RNN architectures can also be attributed to the development of advanced training algorithms [72]. A reliable and highly efficient training algorithm is of fundamental importance for all iterative multivariate regression approaches. For RNNs, a so-called backpropagation through time (BPTT) method [73] is often used which requires a temporal unfolding of the network in accordance with Fig. 1. We refer the reader to the supplementary material for more details on the LSTM approach, stacked hidden layers and advanced training algorithms.

3. Numerical details

In this section, we present the features and the characteristics of the corresponding process experimental data sets. Moreover, we discuss the numerical details of the specific and generic RNN models.

3.1. Data sets

3.1.1. Large scale process runs

The large scale data set for the specific RNN model considers individual manufacturing runs for a single API with values for the titer, total cell density (TCD), viable cell density (VCD), viability, glucose and lactate concentration at distinct time points. Non-considered process parameters like seeding cell densities, time points for bolus addition or feeds as well as set points for temperatures were identical for the runs. The data set included 118 process runs with 9 measurements each at different time points with roughly comparable time intervals of 24 h. All large scale data correspond to a validated process that is executed in a 12,000 L bioreactor. The measurements were performed by a standard set of analytical methods to determine the product, cell or metabolite concentration in samples taken from cell suspension. The raw process data is shown in the supplementary material.

3.1.2. Small scale process runs

The ensemble process data set for the generic RNN model combines the runs of four individual mAb production processes at development scale. All processes were subject to the same platform procedure including identical growth and feed media as well as CHO clone cells. The data set included 90 processes in total with 16 runs for mAb A, 25 runs for mAb B, 25 runs for mAb C and 24 runs for mAb D including values for the product titer, TCD, VCD, viability, glucose and lactate concentration, actual pH value, bioreactor volume and cultivation temperature. All parameters were systematically and consistently varied between the runs. The individual runs included 15 measurements from initial start time with comparable time intervals of roughly 24 h for each mAb production process. Values for the product titer were only measured for the last 6 time points due to nearly negligible values for the previous lag and exponential growth phase. The glucose concentration was manually changed for some processes at later process time points due to modified feeding strategies. Despite being platform processes, minor differences between the individual products and processes can be noted for the temperature, seeding cell density, upper pH values, power inputs, medium equilibration times, and gassing rates which vary slightly among the products and the processes. Moreover, the different mAbs were of similar product type, but had slight differences in their genetic sequence and hence expression behavior. In contrast to the large scale data set, further variability can be attributed to the analytical methods, the used equipment and the corresponding calibration procedure in non-good manufacturing practice (non-GMP) and hence non-validated environment. All key parameter raw data for mAbs A, B, C and D are shown in the supplementary material.

3.2. Details of the RNNs

3.2.1. Specific RNN model for large scale process runs

All RNN models were programmed in Python 3.7.1 by using the modules Pandas and Numpy. The architecture of the RNN was implemented through the Keras module (version 2.3.1) [74] relying on the TensorFlow backend (version 1.13.1) [75]. Each of the three hidden layers in the RNN was formed by 120 nodes and the first and the second layer (LSTM layers) were made recurrent while the third layer only considers a feed-forward connection to the dense output layer. A hyperbolic tangent (tanh) was chosen as corresponding activation

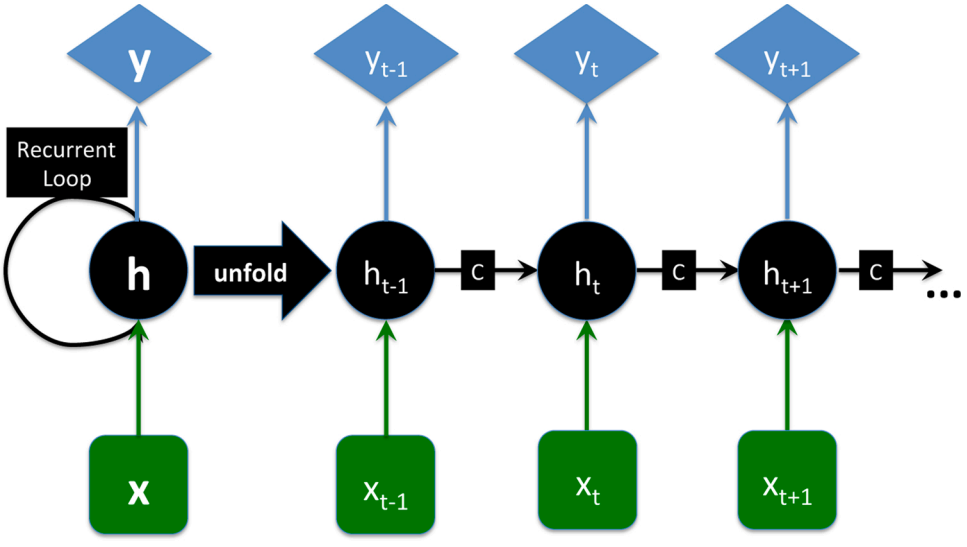

Fig. 1. Scheme of a recurrent neural network with one hidden layer. The green squares denote the input layer, the black circles the hidden layer and the blue diamonds the output layer. All arrows denote data flow and calculations in the corresponding direction. A compact structure of the RNN is shown on the left side with the recurrent loop. The temporal unfolding of the recurrent loop and the hidden layer shows the network structures on the right side. It can be seen that the individual representations focus on distinct time or sequence points $t - 1$, t, $t + 1$ as represented by the connections between the input and output sequence points. The recurrent loop is implemented as a connection between the nodes h_{t-1}, h_t, …, h_T such that h_{t-1} and h_t, respectively, are communicated (as denoted by the black square with C) to h_t and h_{t+1}. The dots on the right side mark the remaining and not shown unfolded connections with final sequence calculations for x_T, h_T and y_T. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

function. The learning rate was set to 0.001. For all input values of the generic and the specific RNN models, we used a robust scaler which removes the median and scales the data according to the interquartile range. The interquartile range is the range between the first quartile (25% quantile) and the third quartile (75% quantile). No further data preprocessing was performed. All calculations for the principal component analysis were performed with a standard scaler, which subtracts the mean value and normalizes by division with the standard deviation. For specific and generic RNN model training, we used the Adaptive Moment Estimation (Adam) optimizer [76] with the mean absolute error (MAE) as loss function. As a standard procedure to avoid overfitting, we added a dropout function [77] with a fraction of 0.1 to the LSTM layers.

We considered 118 experimental process runs at large scale and 20 randomly chosen runs were used for validation in terms of a standard training/test splitting procedure [48]. The test data are not included in the training data set. Input and target values were the titer, TCD, VCD, viability as well as glucose and lactate concentrations. The training phase initially included 500 epochs with an early stopping function [78]. As stopping criterion, we used a standard MAE loss function [48] which needs to show a convergent behavior within 20 epochs. The RNN batch size for data processing was chosen as 4 which was used for each input vector. Instead of predicting or learning the whole temporal sequence at once, we introduced an interval procedure which introduces the two input values from previous time points for the calculation of the output value at the next time point in terms of

$$(x_{t-1}, x_t) \rightarrow y_{t+1} \tag{5}$$

for all $t = 1, \ldots, T - 1$. Thus, only the first (initial measurement at process start) and the second value from the measurements in each run need to be known for the RNN calculations. The choice of this value was motivated by the presented results for the autocorrelation functions, which show a pronounced non-Markovian behavior. Notably, such an approach allows a simulation of process outcomes also from random starting configurations as outlined in the remainder of this article.

3.2.2. Generic RNN model for small scale runs

Due to the smaller data set for the development runs, the RNN included only two hidden layers with 120 nodes each and the first layer (LSTM layer) was chosen as recurrent while the second layer relies on a feed-forward connection to the dense output layer. The learning rate was set to 0.001. In order to avoid overfitting, we added a dropout function [77] with a fraction of 0.1 to the LSTM layer. All other hyperparameter settings and chosen algorithms were identical to the specific RNN models for the large scale runs as discussed in the previous subsection.

For purposes of training, we used 86 process experimental runs (15 from mAb A, 24 from mAb B, 24 from mAb C and 23 from mAb D) and one randomly chosen process experiment for each mAb in terms of validation procedures. As an extension of a simple leave-one-out procedure, further evaluation with regard to random shuffling of training and test data for 100 repetitions finally provided reliable estimates for important statistical quantities in terms of validation procedures. For all repetitions, we ensured that the test data was not included in the training data set. The training phase initially included 500 epochs with an early stopping function [78]. Input and output values included the titer, TCD, VCD, viability, glucose and lactate concentration, actual pH value, bioreactor volume and the cultivation temperature. In contrast to the specific RNN models, the bioreactor volume, the actual pH value and the considered cultivation temperatures (between 307.65 K and 308.65 K) were additionally taken into consideration. The RNN batch size was chosen as 8 which was used for each input vector. An identical interval learning procedure (Eq. (5)) like for the specific RNN model was used for all calculations.

3.3. Simulations: impact of different temperatures

The generic RNN model was also used for simulations including different temperatures. Although one can in principle study also other effects, *e.g.* pH variations, we concentrate on the impact of temperatures as these induce the most significant changes in the process outcomes. Each individual run was started at a fixed temperature of 307.65 K, 308.15 K and 308.65 K. For each temperature, we performed 2500 independent process simulations based on the pre-trained small scale RNN models. For the initial and the first time point, the corresponding values for the titer, TCD, VCD, viability, glucose and lactate concentration, pH value and bioreactor volume were drawn from a normal distribution with mean value $\mu_p(\tau)$ and variance $\sigma^2_{\text{Exp}}(\tau)$ where τ denotes the measurement time. The values for $\mu_p(\tau)$ and the variance $\sigma^2_{\text{Exp}}(\tau)$ were calculated for the individual process parameters from the original experimental process data sets at the corresponding first two measurement points in accordance with the simulated temperatures. The generic RNN model used these values as random input parameters drawn from normal distributions with the same moments and provides the corresponding outcomes for the later time points in terms of fixed interval calculations (Eq. (5)). The temperature was kept constant during the simulations while all other parameters were subject to intrinsic changes. The corresponding mean values and standard deviations for the combined simulation runs are calculated at distinct time points in order to study the influence of different temperatures on key process outcomes. For purposes of independent validation, experimental values of mAb E for the VCD, TCD, product titer and viability at comparable time points in terms of a platform process at fixed temperatures of 307.65 K, 308.15 K and 308.65 K were monitored. The corresponding processes and values related to mAb E were not used for training or validation of the RNN models and serve as an independent experimental confirmation of the simulations.

3.4. Validation methods

Each RNN model was validated by comparison between the computed and the experimental (target) values. As standard statistical quantities, we used the mean absolute error of prediction (MAE) and the root-mean-squared error of prediction (RMSE). When divided by the standard deviation of the ensemble experimental values for the process parameter $\sigma_{\text{Exp}}(\mathbf{x})$, the corresponding normalized MAEs (nMAEs) and normalized RMSEs (nRMSEs) provide an unbiased estimate for the precision of the predictions. All our results revealed minor values for nMAEs and nRMSEs (nMAE < 0.31 and nRMSE < 0.43), which highlights the fact that the corresponding RNN model achieved a significantly higher accuracy when compared to a simple standard $3\sigma_{\text{Exp}}(\mathbf{x})$ deviation criterion. In addition, we computed the corresponding values for the validation and the training data set in order to detect overfitting issues. With regard to the used dropout procedure in combination with the early stopping convention, all our results for the nMAEs and nRMSEs revealed that issues of overfitting can be largely ignored. Furthermore, the Pearson correlation coefficients showed high values (for most values $R^2 \geq 0.94$), which demonstrates the linear relationship between computed and experimental values. Detailed values will be discussed and presented in the remainder of the article. Noteworthy, the unknown functional relationship between the target and the input values does not allow us to compute confidence intervals in order to estimate the statistical accuracy of the calculations. In over to overcome this shortcoming, we splitted the experimental data sets for the small and the large scale runs into training and test data. The test data was not considered for the training of the RNN and the nRMSE and nMAE values were used for model validation. If the calculations for the test data reveal significantly lower nRMSE and nMAE values than unity, a higher precision when compared to randomly drawn parameter values from the underlying experimental ensemble distribution is assumed. Moreover,

such an approach also allows a straightforward detection of outliers.

4. Numerical results

In this section, we first discuss the application of the specific RNN model for large scale manufacturing processes. Hereafter, we present a generic RNN model for the prediction of distinct mAb production processes at small scale. The corresponding generic RNN approach will also be used to simulate process outcomes for different temperatures.

4.1. Specific RNN model for large scale processes

4.1.1. Principal component analysis and autocorrelation functions

In principle, one may ask why an RNN model should provide reasonable results for key process parameters? Such a question is closely related to the temporal evolution of variables as well as the corresponding Markovian properties. As a further important property, the correlation between the individual parameters can be studied through a principal component analysis (PCA) [48]. The covariance matrix $\underline{C}$ for a process parameter vector $\mathbf{x}$ is defined by

$$\underline{C} = \left\langle (\mathrm{x} - \left\langle \mathrm{x} \right\rangle) \cdot (\mathrm{x} - \langle \mathrm{x} \rangle)^T \right\rangle \tag{6}$$

where $\langle \cdot \rangle$ denote mean values. With regard to the use of orthogonal basis transformations to a new vector $\mathbf{z}$ in terms of

$$\mathrm{x} - \left\langle \mathrm{x} \right\rangle = \underline{T} z \tag{7}$$

and equivalently

$$z = \underline{T}^{\mathrm{T}} (\mathrm{x} - \left\langle \mathrm{x} \right\rangle), \tag{8}$$

one can obtain the following expression

$$\underline{C} = \underline{T}^{\mathrm{T}}\, \underline{\Omega}\, \underline{T} \tag{9}$$

with the diagonal matrix $\underline{\Omega}$, where the jth column of $\underline{\Omega}$ corresponds to the principal component PC_j with eigenvalue w_j. In addition to the introduction of independent and orthogonal eigenvectors (principal components), PCA also provides insights into essential fluctuations. Herewith, the explained variance can be calculated which sheds light onto concerted process parameter variations [48]. For such an analysis, we considered the K principal components as calculated from the experimental data set, such that the explained variance for the cumulative contribution of fluctuations including all principal components PC_j with $j = 1, \ldots, \alpha$ can be written as

$$\mathrm{EV}_\alpha = \frac{\sum_{j=1}^{\alpha} w_j}{\sum_{j=1}^{K} w_j} \tag{10}$$

with the condition $\alpha \leq K$. The corresponding results for the large scale runs are shown in Fig. 2. As can be seen, roughly 67% of all variations within the data set can be assigned to the principal component PC 1. In combination with PC 2, it follows that nearly 95% of all fluctuations and variations can be described by only two PCs. Such extremely high values for the first two PCs forming the essential subspace are remarkable and point to the fact that most of the process outcomes are highly correlated. In addition to the correlations, one can also observe a temporal evolution of the process outcomes as monitored by the first two principal components. Hence, the values in the lower left corner of Fig. 2 (right side) can be attributed to initial process conditions while the final values for key process parameters in terms of PC 1 and PC 2 are located in the upper right corner. The corresponding correlations as shown in the supplementary material reveal that PC 1 is mainly dominated by the titer, the viability and the glucose concentration, while PC 2 shows its highest correlations with the TCD and the lactate concentration.

Closely related, the results for the individual autocorrelation functions [79–82] in terms of actual process parameter values x as defined by

$$\mathrm{ACF}(\tau) = \frac{\sum_{\tau_0}^{\tau_N - \tau} (x_{\tau_0} - \left\langle x \right\rangle) \cdot (x_\tau - \left\langle x \right\rangle)}{\sum_{\tau_0}^{\tau_N} (x_{\tau_0} - \langle x \rangle)^2} \tag{11}$$

at certain time points τ and τ_0 with $\tau_0 \leq \tau \leq \tau_N$ provide an estimate for the temporal correlation and the full decorrelation time τ_D at ACF $(\tau_D) \approx 0$. The corresponding results for all process outcomes are

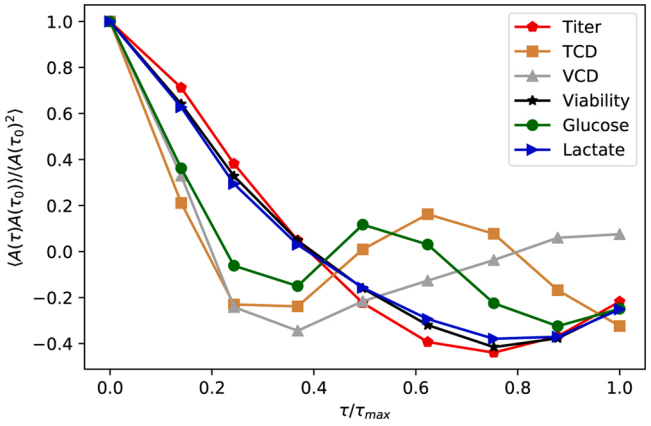

Fig. 3. Autocorrelation functions for the corresponding temporal process parameter changes with reference to distinct time points τ/τ_{max}.

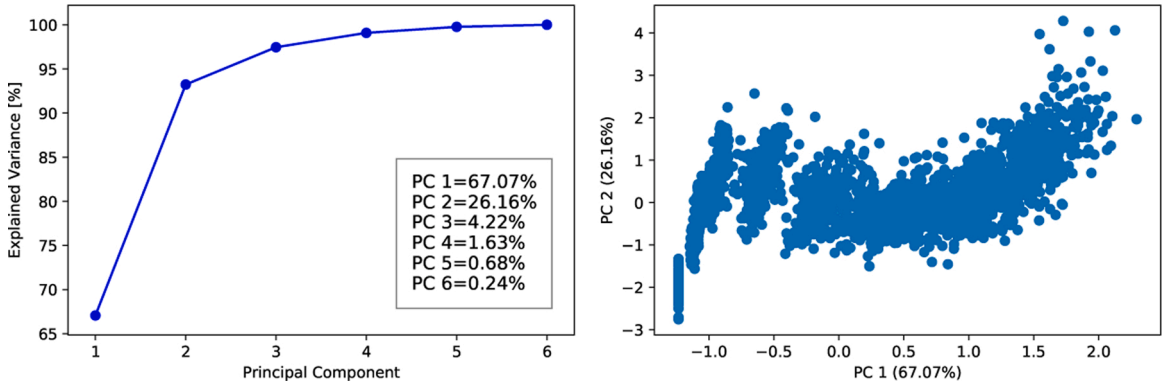

Fig. 2. Left side: Explained variance in terms of principal components (PC) for the large scale manufacturing runs. The values for the explained variance of the individual PCs are presented in the inset. Right side: Values of PC 1 and PC 2 for individual large scale manufacturing process runs.

presented in Fig. 3. As can be seen, the autocorrelation functions for the titer, lactate concentration and viability show a comparable decay and thus strong temporal correlations. In addition, all correlations vanish for these parameters at $\tau_D/\tau_{max} = 0.4$, where τ_{max} denotes the final time point. Notably, also the values for the TCD, VCD and glucose concentration show a concerted temporal decay with a shorter decorrelation time of $\tau_D/\tau_{max} = 0.2$. Such findings can be rationalized by the strong correlation between lactate and titer production as well as glucose consumption [83]. Due to different slopes, the individual phases of the process in terms of exponential growth phase and stationary non-growth phase can be clearly identified [84]. Despite the fact that one recognizes two relevant decorrelation times for the initial decay of the process variables, the broad comparability of the individual process parameter autocorrelation functions becomes evident. As already mentioned, such characteristics are highly beneficial for any RNN in terms of non-Markovian processes which facilitate meaningful predictions for reasonable changes in the process outcomes. Finally, the negative values for the ACF can be attributed to an anticorrelated behavior in which the temporal change of the process parameter values is reversed.

4.1.2. *Results of the specific RNN model*

The experimental data sets for the large scale runs in terms of mean values and standard deviations for certain time points are presented in the supplementary material. As expected for large scale manufacturing processes, individual variations due to slight process parameter changes are noticeable, but not highly pronounced. Nevertheless, such small variations are a challenging task for a specific RNN approach. Arbitrarily chosen experimental values for key parameters in combination with the outcomes of specific RNN model calculations are presented in Fig. 4. In general, one can recognize a good agreement between the calculated and the experimental values, which also includes accurate predictions for rapid changes in the glucose concentration at $\tau/\tau_{max} = 0.5 - 0.6$ as well as for the TCD (significant increase for $\tau/\tau_{max} > 0.6$). With regard to larger standard deviations at certain time points in the training data sets

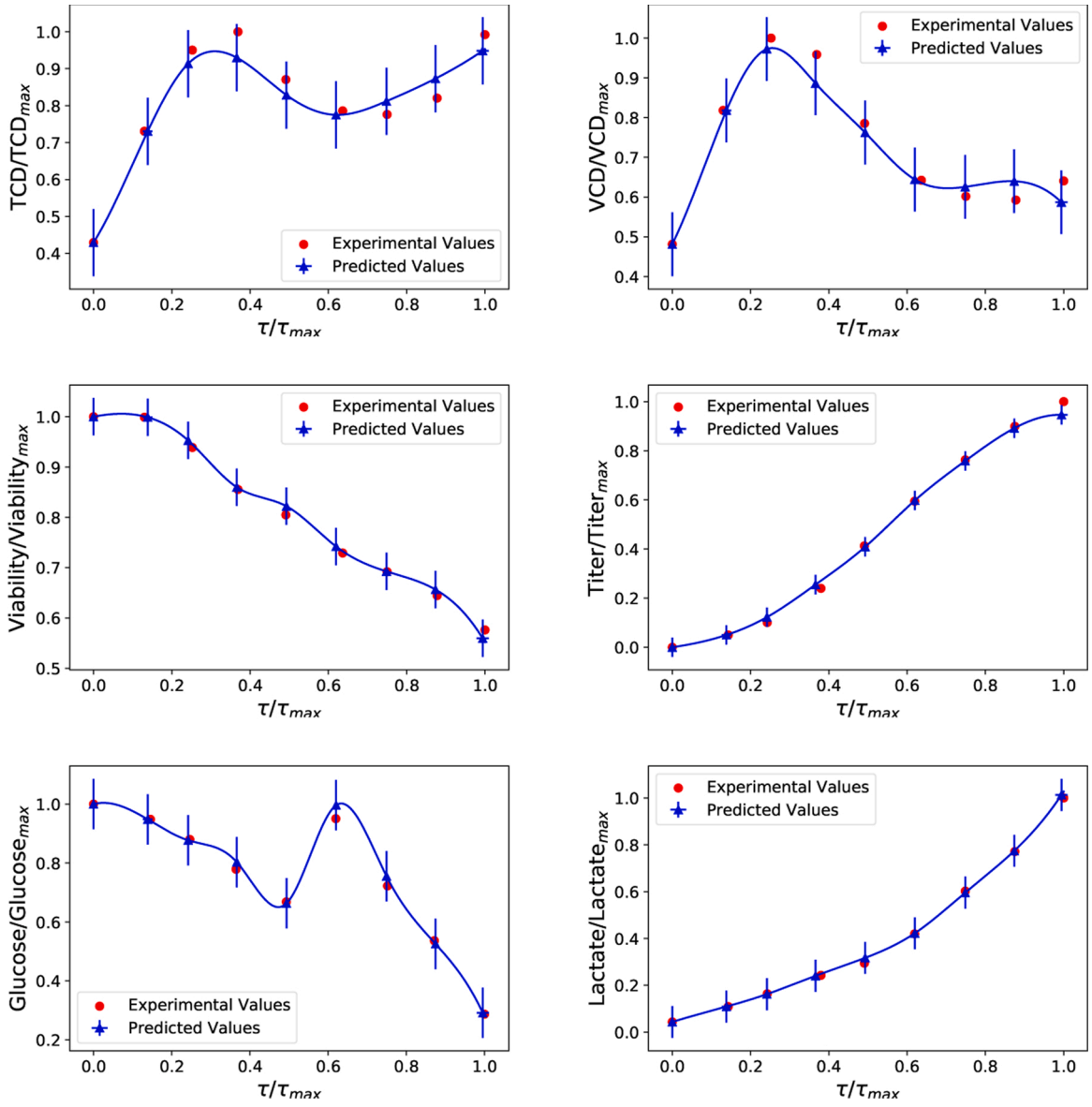

Fig. 4. Specific RNN model calculations (blue triangles) and experimental results (red circles) for the large scale process data sets in terms of randomly chosen process runs including TCD (top left), VCD (top right), viability (middle left), titer (middle right) as well as glucose (bottom left) and lactate concentration (bottom right). The blue lines correspond to cubic spline functions as guides for the eyes. The errorbars denote the global root-mean-squared errors of predictions for the RNN in terms of the test data set and the corresponding target variable (see text for more details). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(as shown in the supplementary material), one would specifically assume less precise RNN calculations for TCD and VCD values at $\tau/\tau_{max} > 0.2$. With reference to the RNN results, it can indeed be seen that the calculated TCD and VCD values show some slight variations at exactly these time points. In terms of the chosen interval approach (Eq. (5)), one would assume that such inaccuracies also progress to later time points, which rationalizes the slight discrepancies between experimental and computed results. Corresponding conclusions can also be drawn for some outliers in the glucose and lactate concentration at later process times. In terms of the experimental values as shown in the supplementary material, it can be seen that the standard deviation of the data points increases with process time. Hence, such an increasing variability can be regarded as a challenge for the interval learning approach (Eq. (5)) in terms of error progression which rationalizes the observed slight deviations. In addition, the glucose concentration is slightly changed by non-monitored external bolus additions whereas the lactate concentration strongly depends on the cell metabolism. Despite such slighter deviations, it has to be noted that all trends in the process parameters are well reproduced.

The results for the computed values based on the training and the test data sets are shown in Fig. 5 and the corresponding key statistical values are presented in Table 1. As can be seen, all computed results show a high linear correlation with the experimental data in terms of Pearson correlation coefficients $R^2 \geq 0.94$. Notably, the lowest values of R^2 can be observed for the values of VCD, TCD and the glucose concentration. The larger deviations from linearity for these process outcomes can be related to the more pronounced variations in the experimental data set as discussed before. Although some outliers can be identified which deviate significantly from the black line with unit slope, the vast amount of predictions reveals a high agreement with the corresponding experimental values due to rather low mean absolute errors for the validation data set as defined by

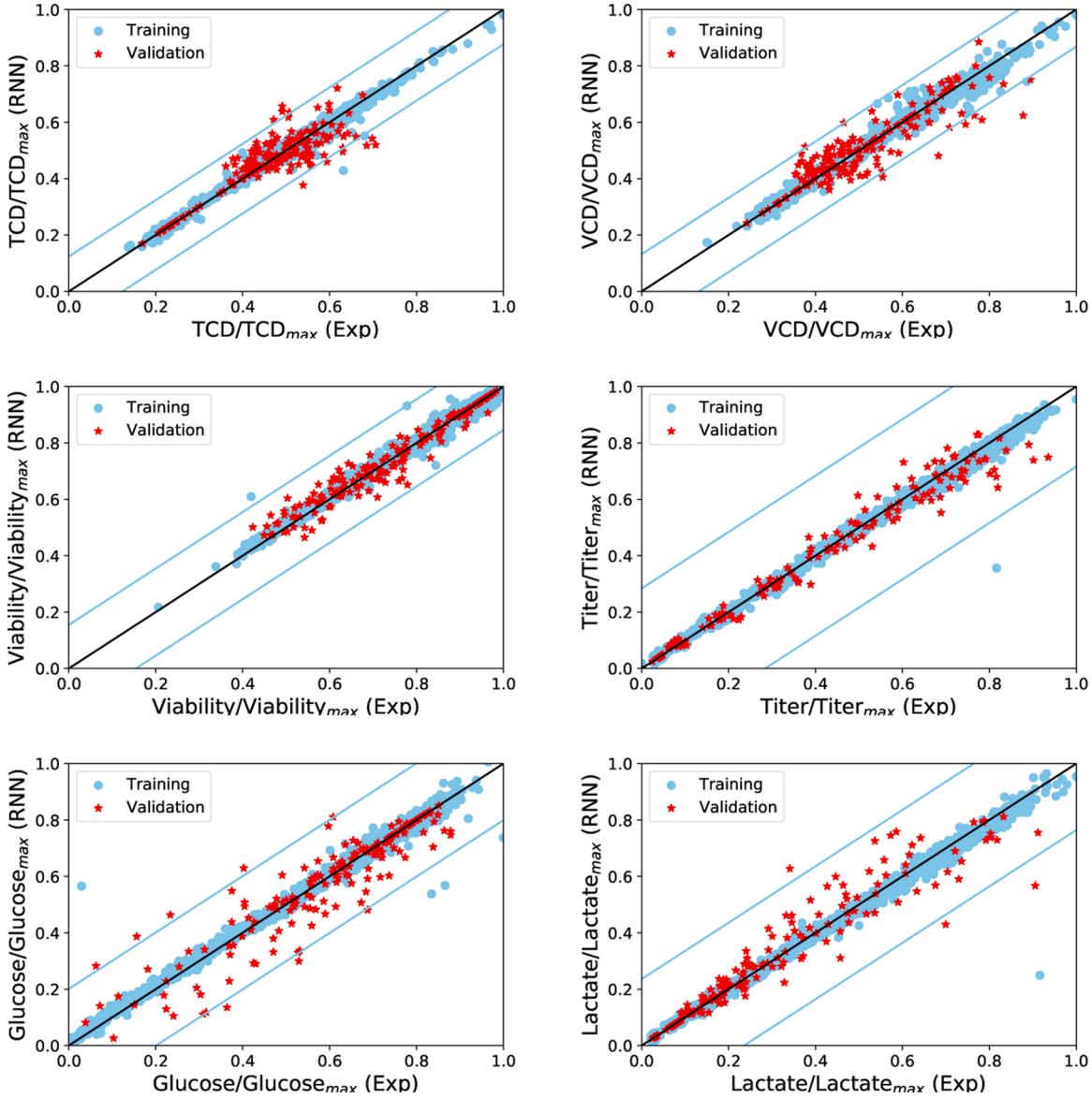

Fig. 5. Specific RNN model calculations for the test data set (red diamonds) and for the training data set (blue circles) with regard to the experimentally measured data (*x*-axis) and the predicted values (*y*-axis) including the TCD (top left), VCD (top right), viability (middle left),product titer (middle right), glucose (bottom left) and lactate concentration (bottom right). The black solid line has a slope of unity and represents full coincidence between measured and predicted values while the straight blue lines represent the ensemble standard deviation σ_{Exp} of the experimental data set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Mean Pearson correlation coefficients R^2, fraction of computed values x which are located within the ensemble experimental standard deviation $P(x < \sigma_{\mathrm{Exp}})$, normalized mean absolute error MAEs (nMAE) and normalized root-mean squared error (nRMSE) between computed and experimental values for the specific RNN model when averaged over the test data set (columns 3 and 4) and over the training data set (last two columns).

Value	R^2	$P(x < \sigma_{\mathrm{Exp}})$	nMAE	nRMSE	$\mathrm{nMAE_{tr}}$	$\mathrm{nRMSE_{tr}}$
Titer	0.99	1.0	0.09	0.14	0.04	0.06
TCD	0.94	0.95	0.30	0.42	0.07	0.11
VCD	0.94	0.95	0.30	0.40	0.08	0.12
Viability	0.99	1.0	0.16	0.22	0.07	0.09
Glucose	0.97	0.95	0.25	0.33	0.06	0.10
Lactate	0.99	0.98	0.16	0.23	0.04	0.08

$$\mathrm{MAE} = \frac{1}{N_s}\sum_{i}^{N_s} |\mathrm{y}_i - \widehat{\mathrm{y}}_i| \tag{12}$$

as well as the root-mean-squared error

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_s}\sum_{i}^{N_s} (\mathrm{y}_i - \widehat{\mathrm{y}}_i)^2} \tag{13}$$

where $\mathbf{y}_i$ and $\widehat{\mathrm{y}}_i$ denote the corresponding calculated and target values for N_s samples. Such assumptions are further underpinned by the values of the normalized MAEs

$$\mathrm{nMAE} = \frac{\mathrm{MAE}}{\sigma_{\mathrm{Exp}}} \tag{14}$$

and the normalized RMSEs

$$\mathrm{nRMSE} = \frac{\mathrm{RMSE}}{\sigma_{\mathrm{Exp}}} \tag{15}$$

where σ_{Exp} denotes the standard deviation of the experimentally measured ensemble data for the corresponding process parameter. Specifically the calculations for the titer, the viability as well as the lactate concentration reveal a high level of accuracy in terms of low nMAEs and nRMSEs (Table 1). Slighter deviations can be observed for the TCD, VCD and the glucose concentration. Nevertheless, the corresponding values for the nMAEs and nRMSEs are smaller than unity which highlights the applicability of the RNN model even for predictions of more complex process outcomes. As can be concluded, the results of the specific RNN model provide a significantly higher accuracy when compared to statistical estimates in terms of experimental standard deviations and the often used $3\sigma_{\mathrm{Exp}}$ criterion. In addition, a comparison of the nMAE and nRMSE values in Table 1 for the training and the test data reveals a comparable order. Hence, significant issues of overfitting can be largely ignored such that the aforementioned outliers can mainly be attributed to the broader experimental variability at the corresponding process stages. In consequence, the corresponding nMAEs and nRMSEs show a good predictive accuracy which rationalizes the use of this approach for large scale manufacturing runs. With regard to this point, it has to be noted that large scale processes reveal minor variations due to already well-defined process conditions when compared to exploratory small scale development processes. The application of RNNs for such processes will be discussed in more detail in the next subsection.

4.2. Generic RNN model for small scale processes

4.2.1. Principal component analysis and autocorrelation functions

With regard to the last section, it can be concluded that a specific RNN model for large scale runs provides meaningful results. However, the question remains if also a generic model can be developed which is able to compute the outcomes of mAb production processes at smaller scales. Such a model would be helpful to study optimal process conditions and to predict general trends for the performance of novel development candidates. Motivated by these points, we combined the small scale run data sets for four mAb products in order to study the properties as well as the validity of such a generic RNN model. As a first step, we performed a PCA on the corresponding data. The results for the explained variance of the combined process data as well as a projection of the process data on the first principle components are shown in Fig. 6. Due to the larger number of input variables, it has to be noted that we have to consider 9 PCs in contrast to the large scale runs. In consequence, the consideration of only two principle components PC 1 and PC 2 provides a reduced value for the explained variance in terms of roughly 58%. Hence, the corresponding values are a little bit smaller when compared to the large scale runs which can be rationalized by the larger number of input vectors as well as the distinct process characteristics (as shown in the supplementary material). A projection of the process data on the first two principle components is depicted on the right side of Fig. 6. As can be seen, the individual process data differ slightly in terms of mean positions and ranges, but significant overlap regions can also be identified. Thus, the individual processes show slight deviations but also some similarities which rationalizes their use for the development of a generic RNN model. Specifically the individual values for PC 2 highlight the clustering of the data into separated mAb processes. Noteworthy, the points in the lower left corner in Fig. 6 can be assigned to initial process parameter values while the symbols in the upper right corner correspond to final process outcomes. Such conclusions are further supported by the individual correlation coefficients of the principal components with the considered process parameters as shown in the supplementary material, which reveal high correlations of PC 1 with the product titer, the pH value and the viability as well as the product titer, the volume and the TCD (PC 2). In comparison to the specific RNN model, it can be assumed that the accuracy of the generic RNN approach will be less pronounced, which is due to the broader variation of the corresponding process parameters with regard to the individual platform projects.

Despite these slighter discrepancies, the results for the autocorrelation functions (Fig. 7) highlight a comparable temporal evolution of the corresponding process variables. Thus, all mAb process outcomes show a similar decay pattern for the titer, VCD and TCD with a decorrelation time of $\tau_D/\tau_{max} \approx 0.35$. In contrast to the large scale runs (Fig. 3), the temporal evolution of the titer is inherently coupled to the VCD and the decorrelation time is significantly larger. These findings can be rationalized by the complex biological metabolism of the CHO cells as described in the literature [83–85,20]. Notably, the comparable temporal decay of all process outcomes for distinct products can be considered as a consequence of the underlying platform process. With regard to this point, also the autocorrelation functions for the viability, as well as the glucose and lactate concentration reveal a comparable decay. In consequence, the outcomes of the PCA and the ACF highlight the potential applicability of a generic RNN model. Moreover, the pronounced non-Markovian behavior for the first three time points ($\tau/\tau_{\mathrm{max}} < 0.18$) rationalizes the use of the proposed interval learning scheme (Eq. (5)).

The raw data for the individual process runs are presented in the supporting material. In contrast to the large scale runs, the variance of the ensemble small-scale runs due to distinct mAb production processes is more pronounced. However, the question remains if a generic RNN model is able to distinguish between the four distinct mAbs in terms of individual process predictions. Examples for predicted values in terms of randomly chosen process runs for products mAb A, mAb B, mAb C and mAb D are presented in Fig. 8. As can be seen, the computed results show some larger variations which are expected in terms of the larger variance in the experimental data as shown in the supplementary material. However, it is worth to notice that the RNN is even able to reproduce the complex glucose concentration profiles as observed in the experiments. Notably, the deviations for all values become larger at $\tau/\tau_{max} \geq 0.75$

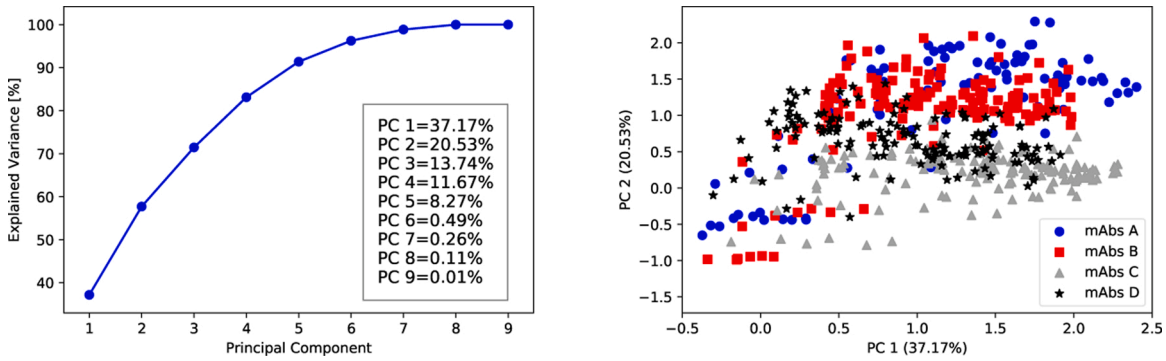

Fig. 6. Left side: Explained variance in terms of principal components (PC) for the small scale process runs. The corresponding value of the explained variance for the individual PCs is presented in the inset. Right side: Values for principal component 1 and principal component 2 in terms of individual process runs for mAb A (blue circles), mAb B (red squares), mAb C (gray triangles) and mAb D (black diamonds). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

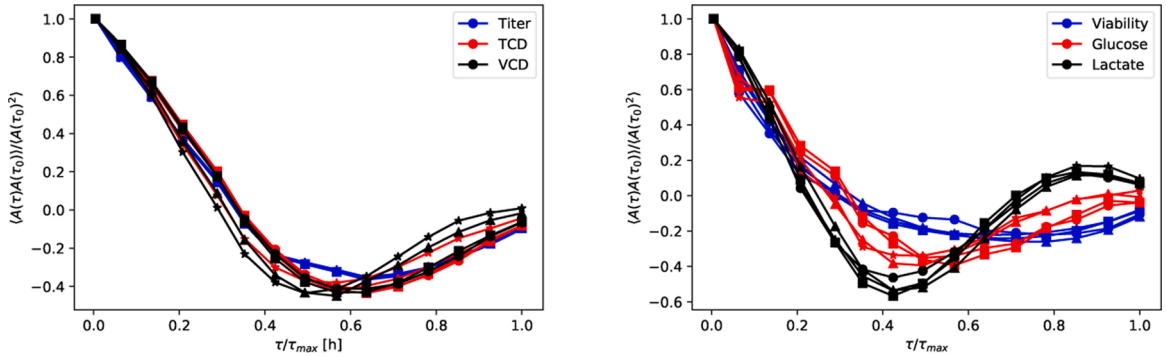

Fig. 7. Autocorrelation functions for the four mAb production processes as denoted by circles (mAb A), squares (mAb B), triangles (mAb C) as well as diamondoids (mAb C) for the product titer (blue color), TCD (red color) and VCD (gray color). The corresponding results for the viability, the glucose and the lactate concentration are shown on the right side. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which can be related to the propagation behavior of uncertainties as discussed in the previous subsection. Despite some discrepancies, it becomes evident that the corresponding results reveal a good agreement with the experimental data such that general trends are well reproduced.

The comparison between all predicted and experimental values for different mAbs is shown in Fig. 9. As can be seen, the accuracy is not that high when compared to the specific RNN predictions for the large scale runs, but still establish a reasonable agreement in comparison with the experimental outcomes. With regard to the corresponding statistical values in Table 2, it can be seen that the normalized MAEs and normalized RMSEs reveal a satisfying accuracy. Slighter deviations can mainly be observed for the glucose and the lactate concentrations which are subject to modified feeding strategies within the process as well as metabolic properties. In summary, the corresponding results for the titer, the TCD, the VCD and the viability reveal a high predictive accuracy. Despite the fact that certain predictions for individual mAb process outcomes differ from the experimental values, *e.g.* larger differences for mAb A between predicted and experimental values in Fig. 9, one can conclude that the generic RNN model is validated for processes with comparable parameter variation ranges. Such conclusions are also underpinned by the low nMAE and nRMSE values which rationalize the validity of our approach.

4.3. Simulated processes: temperature effects

In this subsection, we use the generic RNN model to study the influence of distinct conditions on small scale process outcomes. Here, we explicitly focus on the influence of different temperatures and how these affect the corresponding key process outcomes. With regard to this point, we simulated artificial process runs for fixed temperatures $T = 307.65$ K, 308.15 K and 308.65 K. The corresponding results with standard deviations (vertical bars) are presented in Fig. 10. The corresponding simulations are compared to experimental outcomes for key parameters of mAb E (filled symbols) in terms of a comparable platform process. Noteworthy, the values for mAb E were not used for training of the generic or specific RNN models. As can be seen, the corresponding values for the product titer, TCD, VCD and viability are mainly located within the standard deviation of the process simulations. Slighter deviations can only be observed for the viability at the lowest considered temperature. Despite these differences, it becomes evident that the computed results are in good agreement with independent experimental values as monitored for the product mAB E subject to the same platform process.

Besides predictions, one can obtain insights into the impact of distinct temperatures on growth rates and metabolite concentrations. For instance, it becomes evident that increasing temperatures establish higher titer as well as TCD and VCD values. In contrast, the values for the viability decrease with increasing temperatures. These findings can be rationalized by the faster metabolism at higher temperatures as known for mammalian cells [85]. In consequence, it can be concluded that the generic RNN model can be used to achieve deeper insights into modified process conditions and how they affect the process outcomes.

5. Summary and conclusions

We presented a novel approach for the calculation and prediction of upstream process outcomes in terms of specific and generic RNN models which do not rely on specific calibration procedures when compared to

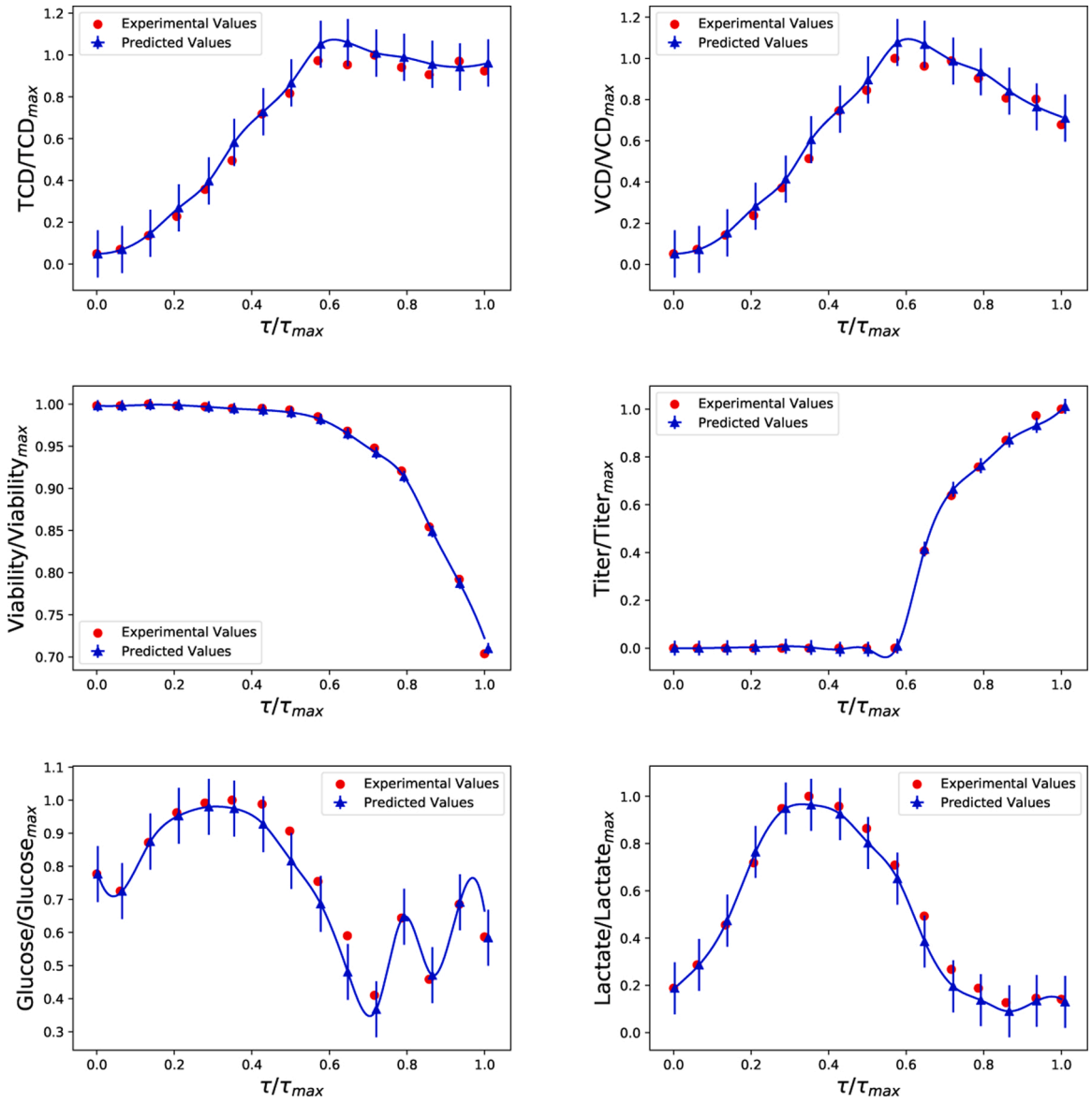

Fig. 8. Specific RNN model results (blue diamonds) for randomly chosen processes from four mAb development candidates in combination with the corresponding experimental results (red squares) including the TCD (top left), VCD (top right), viability (middle left), titer (middle right) as well as glucose (bottom left) and lactate concentration (bottom right). Measured data for the titer at $\tau/\tau_{max} \leq 0.6$ are not available. The predicted profiles (blue lines) are cubic spline functions which connect the outcomes of the individual RNN calculations. The errorbars denote the global mean absolute errors of calculations for the RNN in terms of the corresponding target variables (see text for more details). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

semi- or full parametric approaches. We demonstrated the validity of the models for large scale runs as well as for distinct individual small scale processes in terms of a platform-dependent generic RNN model. The corresponding results reveal a reasonable and good agreement with the experimental data which highlights the validity of our approach. All calculated values show minor variations when compared to ensemble experimental standard deviations such that the normalized MAE and normalized RMSE values are smaller than unity. Thus, our models provide a high accuracy which can also be used to simulate key process outcomes for small scale upstream processes in order to support the identification of suitable process conditions. In principle, one can use such simulations for the study of varying temperatures, pH values or other process parameter variations like modified feeding strategies with regard to the growth rates as well as metabolite concentrations. Even for large scale runs with minor parameter variations, the corresponding approach can be considered as an useful alternative to hybrid or mechanistic models. In particular, the proposed method reveals its benefits in terms of tighter process control and the identification of potential outliers.

In contrast to parametric models like mechanistic approaches, the proposed RNN modelling strategy is also able to consider intense parameters like temperatures, pH values or dissolved oxygen content. Comparable conclusions can be drawn with regard to modified bolus additions or feeding strategies, which often require a singular and ad-hoc change of the parameters in mechanistic models. Noteworthy, such variations contradict the differentiable form of reaction dynamics in thermodynamic equilibrium and also violate the minimum entropy production principle [86], thereby pointing to the fact that mechanistic models which only rely on mass balance conditions reveal certain shortcomings. Similar conclusions are also valid for hybrid models, which crucially rely on temporally varying rate constants. In agreement with mechanistic models, certain aspects of these models are inconsistent with equilibrium thermodynamics as well as linear non-equilibrium thermodynamics in terms of rapid and non-continuous changes of the

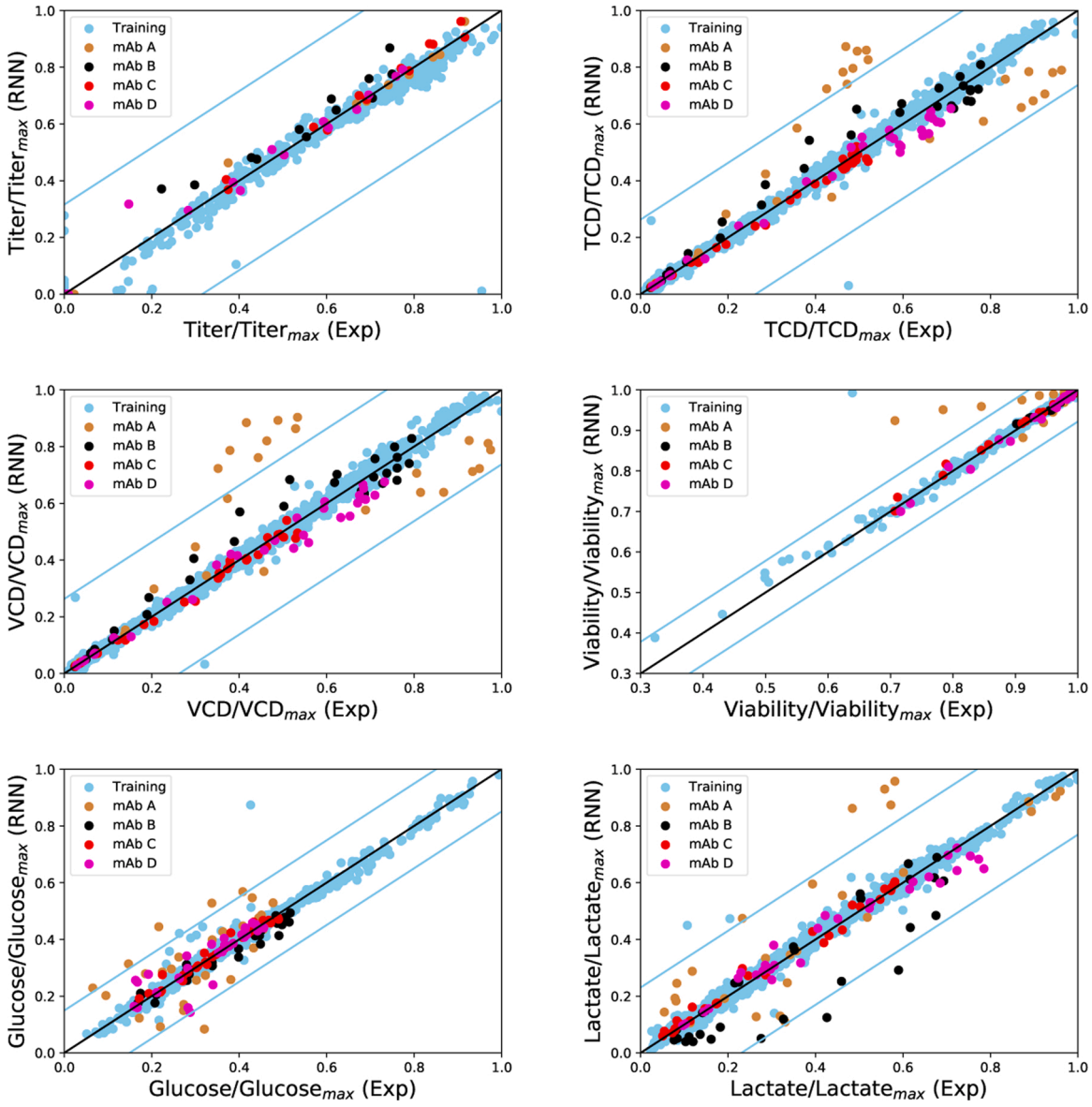

Fig. 9. RNN calculations for the training data set (blue diamonds) and for test process data from mAbs A (brown circles), mAbs B (black circles), mAbs C (red circles) and mAbs D (magenta circles) with regard to the experimentally measured data (x-axis) and the predicted values (y-axis) for the titer (top left), TCD (top right), VCD (middle left), viability (middle right), glucose (bottom left) and lactate concentration (bottom right). The black solid line has a slope of one and represents full coincidence between measured and predicted values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Mean Pearson correlation coefficients R^2, fraction of computed values x which are located within the ensemble experimental standard deviation $P(x < \sigma_{\mathrm{Exp}})$, normalized mean absolute error MAEs (nMAE) and normalized root-mean squared error (nRMSE) between computed and experimental values for the generic small-scale RNN model when averaged over the test data set (columns 3 and 4) and over the training data set (last two columns).

Value	R^2	$P(x < \sigma_{\mathrm{Exp}})$	nMAE	nRMSE	$\mathrm{nMAE_{tr}}$	$\mathrm{nRMSE_{tr}}$
Titer	0.98	0.99	0.08	0.18	0.05	0.12
TCD	0.99	0.93	0.20	0.22	0.06	0.09
VCD	0.99	0.93	0.22	0.23	0.06	0.09
Viability	0.99	0.98	0.13	0.14	0.04	0.12
Glucose	0.83	0.95	0.29	0.38	0.07	0.14
Lactate	0.95	0.94	0.30	0.39	0.07	0.10

entropy production. Thus, the RNN models circumvents the missing detailed knowledge about the underlying reactions, such that a prediction of process outcomes only relies on non-Markovian properties. Hence, although hybrid models may provide a comparable functionality and predictive capability when compared to the RNN approach, it has to be stated that these are often in conflict with the underlying thermodynamic principles.

In consequence, we highlight the straightforward and fast development of RNN models for cultivation processes. The underlying conflicts with thermodynamic boundary conditions can be circumvented by the proposed non-parametric functional form. To the best of our knowledge, such a broad applicability for generic and specific process description was yet not established for any other modeling approach. Although it has to be noted that hybrid as well as mechanistic models reveal their benefits depending on the level of considered detail [87], a comparable complex parameter calibration procedure as known for parametric models is not needed for our approach. Furthermore, intrinsic parameter

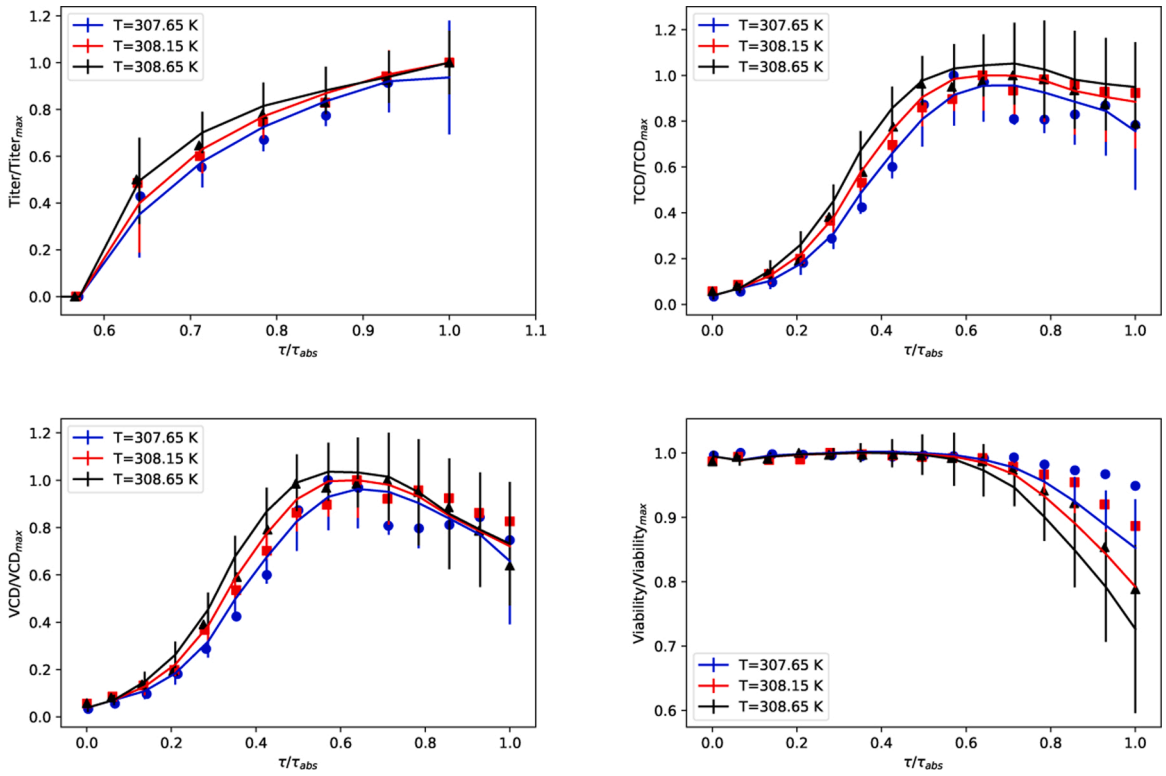

Fig. 10. Generic RNN model simulations for key process parameters outcomes product titer, TCD, VCD, and viability in terms of fixed temperatures $T = 307.65$ K, 308.15 K and 308.65 K. The individual results are obtained from 2500 simulations each with random starting conditions. The corresponding bars represent standard deviations for the individual temperatures as obtained by averaging over the 2500 simulations. The circles denote individual experimental values for one process run of mAb E at $T = 307.65$ K, while the squares and the triangles represent the outcomes for 308.15 K and 308.65 K, respectively.

values like the temperature as well as the pH value which are not part of mass balance conditions can be straightforwardly included in the model. Moreover, the use of non-parametric methods also provides a fast and straightforward retraining of the model if more experimental data become available. The straightforward and fast calculation procedures in terms of full automatization and thus in-line process control can be seen as the largest benefits when compared to other parametric or semi-parametric models. With regard to the recent discussions about the importance of integrated process models, digital twins as well as holistic process models [1], it also has to be noted that RNN approaches can be implemented straightforwardly in any software platform. In summary, the presented RNN models are highly flexible, straightforward to train and they can be used for distinct platform projects in upstream as well as downstream development and manufacturing.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

The authors thank Jan C. Schöning, Samet Yildirim, Verena Nold, Christoph Hold, Eugen Probst, Joachim Bär, Thomas Wärner, Roy Sailer, Mona Bäumli, Carina Gülch, Valerie Schmieder, Raphael Drerup, Fabian Stiefel, Heiko Babel, Alireza Ehsani, Simon Fischer, Tanja Krumpe, Christina Yassouridis, Hermann Schuchnigg, Edmund Salzmann, Stefan Minning, Jochen Schaub and Ogsen Gabrielyan for valuable discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.btre.2021.e00640.

References

[1] J. Smiatek, A. Jung, E. Bluhmki, Towards a digital bioprocess replica: computational approaches in biopharmaceutical development and manufacturing, Trends Biotechnol. 38 (2020) 1141–1153.
[2] D. Roush, D. Asthagiri, D.K. Babi, S. Benner, C. Bilodeau, G. Carta, P. Ernst, M. Fedesco, S. Fitzgibbon, M. Flamm, et al., Toward in silico cmc: an industrial collaborative approach to model-based process development, Biotechnol. Bioeng. (2020), https://doi.org/10.1002/bit.27520.
[3] H. Narayanan, M.F. Luna, M. von Stosch, M.N. Cruz Bournazou, G. Polotti, M. Morbidelli, A. Butté, M. Sokolov, Bioprocessing in the digital age: the role of process models, Biotechnol. J. 15 (1) (2020) 1900172.
[4] G. Guiochon, A. Felinger, D.G. Shirazi, Fundamentals of Preparative and Nonlinear Chromatography, Elsevier, 2006.
[5] F. Rischawy, D. Saleh, T. Hahn, S. Oelmeier, J. Spitz, S. Kluters, Good modeling practice for industrial chromatography: mechanistic modeling of ion exchange chromatography of a bispecific antibody, Comput. Chem. Eng. 130 (2019) 106532.
[6] G. Wang, T. Hahn, J. Hubbuch, Water on hydrophobic surfaces: mechanistic modeling of hydrophobic interaction chromatography, J. Chromatogr. A 1465 (2016) 71–78.
[7] S. Großhans, G. Wang, C. Fischer, J. Hubbuch, An integrated precipitation and ion-exchange chromatography process for antibody manufacturing: process development strategy and continuous chromatography exploration, J. Chromatogr. A 1533 (2018) 66–76.
[8] T. Briskot, F. Stückler, F. Wittkopp, C. Williams, J. Yang, S. Konrad, K. Doninger, J. Griesbach, M. Bennecke, S. Hepbildikler, et al., Prediction uncertainty assessment of chromatography models using bayesian inference, J. Chromatogr. A 1587 (2019) 101–110.
[9] D. Saleh, G. Wang, B. Müller, F. Rischawy, S. Kluters, J. Studts, J. Hubbuch, Straightforward method for calibration of mechanistic cation exchange chromatography models for industrial applications, Biotechnol. Prog. (2020), https://doi.org/10.1002/btpr.2984.
[10] D. Saleh, G. Wang, B. Mueller, F. Rischawy, S. Kluters, J. Studts, J. Hubbuch, Cross-scale quality assessment of a mechanistic cation exchange chromatography model, Biotechnol. Prog. (2020), https://doi.org/10.1002/btpr.3081.
[11] D.W. Hutmacher, H. Singh, Computational fluid dynamics for improved bioreactor design and 3d culture, Trends Biotechnol. 26 (4) (2008) 166–172.
[12] J. Wutz, A. Lapin, F. Siebler, J.E. Schäfer, T. Wucherpfennig, M. Berger, R. Takors, Predictability of kla in stirred tank reactors under multiple operating conditions using an euler-lagrange approach, Eng. Life Sci. 16 (7) (2016) 633–642.

[13] J. Wutz, R. Steiner, K. Assfalg, T. Wucherpfennig, Establishment of a cfd-based kla model in microtiter plates to support cho cell culture scale-up during clone selection, Biotechnol. Prog. 34 (5) (2018) 1120–1128.
[14] J. Wutz, B. Waterkotte, K. Heitmann, T. Wucherpfennig, Computational fluid dynamics (cfd) as a tool for industrial uf/df tank optimization, Biochem. Eng. J. (2020) 107617.
[15] S. Succi, The Lattice Boltzmann Equation: For Fluid Dynamics and Beyond, Oxford University Press, 2001.
[16] M. Bernaschi, S. Melchionna, S. Succi, Mesoscopic simulations at the physics-chemistry-biology interface, Rev. Mod. Phys. 91 (2) (2019) 025004.
[17] J. Smiatek, M. Sega, C. Holm, U.D. Schiller, F. Schmid, Mesoscopic simulations of the counterion-induced electro-osmotic flow: a comparative study, J. Chem. Phys. 130 (24) (2009) 244702.
[18] O.A. Hickey, C. Holm, J. Smiatek, Lattice-boltzmann simulations of the electrophoretic stretching of polyelectrolytes: the importance of hydrodynamic interactions, J. Chem. Phys. 140 (16) (2014) 164904.
[19] M. Brunner, K. Kolb, A. Keitel, F. Stiefel, T. Wucherpfennig, J. Bechmann, A. Unsoeld, J. Schaub, Application of metabolic modeling for targeted optimization of high seeding density processes, Biotechnol. Bioeng. (2021), https://doi.org/10.1002/bit.27693.
[20] S. Pereira, H.F. Kildegaard, M.R. Andersen, Impact of cho metabolism on cell growth and protein production: an overview of toxic and inhibiting metabolites and nutrients, Biotechnol. J. 13 (3) (2018) 1700499.
[21] K.T. Schjoldager, Y. Narimatsu, H.J. Joshi, H. Clausen, Global view of human protein glycosylation pathways and functions, Nat. Rev. Mol. Cell Biol. (2020) 1–21.
[22] L.-E. Quek, S. Dietmair, J.O. Krömer, L.K. Nielsen, Metabolic flux analysis in mammalian cell culture, Metabol. Eng. 12 (2) (2010) 161–171.
[23] W.S. Ahn, M.R. Antoniewicz, Towards dynamic metabolic flux analysis in cho cell cultures, Biotechnol. J. 7 (1) (2012) 61–74.
[24] M.R. Antoniewicz, Dynamic metabolic flux analysis-tools for probing transient states of metabolic networks, Curr. Opin. Biotechnol. 24 (6) (2013) 973–978.
[25] M.R. Antoniewicz, Methods and advances in metabolic flux analysis: a mini-review, J. Ind. Microbiol. Biotechnol. 42 (3) (2015) 317–325.
[26] H. Fouladiha, S.-A. Marashi, F. Torkashvand, F. Mahboudi, N.E. Lewis, B. Vaziri, A metabolic network-based approach for developing feeding strategies for cho cells to increase monoclonal antibody production, Bioproc. Biosyst. Eng. (2020) 1–9.
[27] F.V. Ritacco, Y. Wu, A. Khetan, Cell culture media for recombinant protein expression in Chinese hamster ovary (cho) cells: history, key components, and optimization strategies, Biotechnol. Prog. 34 (6) (2018) 1407–1426.
[28] S. Sha, Z. Huang, Z. Wang, S. Yoon, Mechanistic modeling and applications for cho cell culture development and production, Curr. Opin. Chem. Eng. 22 (2018) 54–61.
[29] B. Frahm, P. Lane, H. Atzert, A. Munack, M. Hoffmann, V.C. Hass, R. Pörtner, Adaptive, model-based control by the open-loop-feedback-optimal (olfo) controller for the effective fed-batch cultivation of hybridoma cells, Biotechnol. Prog. 18 (5) (2002) 1095–1103.
[30] J. Schaub, C. Clemens, H. Kaufmann, T.W. Schulz, Advancing biopharmaceutical process development by system-level data analysis and integration of omics data. Genomics and Systems Biology of Mammalian Cell Culture, Springer, 2011, pp. 133–163.
[31] J. Möller, K.B. Kuchemüller, T. Steinmetz, K.S. Koopmann, R. Pörtner, Model-assisted design of experiments as a concept for knowledge-based bioprocess development, Bioprocess Biosyst. Eng. 42 (5) (2019) 867–882.
[32] A. Ehsani, S. Niedenfuehr, T. Eissing, S. Behnken, A. Schuppert, How to use mechanistic metabolic modeling to ensure high quality glycoprotein production, Comput. Aided Chem. Eng. 40 (2017) 2839–2844.
[33] S. Ulonska, P. Kroll, J. Fricke, C. Clemens, R. Voges, M.M. Müller, C. Herwig, Workflow for target-oriented parametrization of an enhanced mechanistic cell culture model, Biotechnol. J. 13 (4) (2018) 1700395.
[34] A. Ehsani, C.D. Kappatou, A. Mhamdi, A. Mitsos, A. Schuppert, S. Niedenfuehr, Towards model-based optimization for quality by design in biotherapeutics production, Comput. Aided Chem. Eng. 46 (2019) 25–30.
[35] M. Kornecki, J. Strube, Accelerating biologics manufacturing by upstream process modelling, Processes 7 (3) (2019) 166.
[36] H. Narayanan, M. Sokolov, M. Morbidelli, A. Butté, A new generation of predictive models-the added value of hybrid models for manufacturing processes of therapeutic proteins, Biotechnol. Bioeng. 116 (2019) 2540–2549.
[37] M. von Stosch, S. Davy, K. Francois, V. Galvanauskas, J.-M. Hamelink, A. Luebbert, M. Mayer, R. Oliveira, R. O'Kennedy, P. Rice, et al., Hybrid modeling for quality by design and pat-benefits and challenges of applications in biopharmaceutical industry, Biotechnol. J. 9 (6) (2014) 719–726.
[38] M. von Stosch, J.-M. Hamelink, R. Oliveira, Hybrid modeling as a qbd/pat tool in process development: an industrial e. coli case study, Bioprocess Biosyst. Eng. 39 (5) (2016) 773–784.
[39] H. Narayanan, M.F. Luna, M. von Stosch, M.N. Cruz Bournazou, G. Polotti, M. Morbidelli, A. Butté, M. Sokolov, Bioprocessing in the digital age: the role of process models, Biotechnol. J. 15 (2020) 1900172.
[40] S. Nargund, K. Guenther, K. Mauch, The move toward biopharma 4.0: insilico biotechnology develops "smart" processes that benefit biomanufacturing through digital twins, Genet. Eng. Biotechnol. 39 (6) (2019) 53–55.
[41] D. Zhang, E.A. Del Rio-Chanona, P. Petsagkourakis, J. Wagner, Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization, Biotechnol. Bioeng. 116 (11) (2019) 2919–2930.
[42] R. Simutis, A. Lübbert, Hybrid approach to state estimation for bioprocess control, Bioengineering 4 (1) (2017) 21.
[43] B. Bayer, M. von Stosch, G. Striedner, M. Duerkop, Comparison of modeling methods for doe-based holistic upstream process characterization, Biotechnol. J. (2020) 1900551.
[44] B. Bayer, G. Striedner, M. Duerkop, Hybrid modeling and intensified doe: an approach to accelerate upstream process characterization, Biotechnol. J. (2020) 2000121.
[45] B. Bayer, B. Sissolak, M. Duerkop, M. Von Stosch, G. Striedner, The shortcomings of accurate rate estimations in cultivation processes and a solution for precise and robust process modeling, Bioprocess Biosyst. Eng. 43 (2) (2020) 169–178.
[46] P. Zürcher, M. Sokolov, D. Brühlmann, R. Ducommun, M. Stettler, J. Souquet, M. Jordan, H. Broly, M. Morbidelli, A. Butté, Cell culture process metabolomics together with multivariate data analysis tools opens new routes for bioprocess development and glycosylation prediction, Biotechnol. Prog. (2020) e3012.
[47] L. Stepper, F.A. Filser, S. Fischer, J. Schaub, I. Gorr, R. Voges, Pre-stage perfusion and ultra-high seeding cell density in cho fed-batch culture: a case study for process intensification guided by systems biotechnology, Bioprocess Biosyst. Eng. (2020), https://doi.org/10.1007/s00449-020-02337-1.
[48] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning, 12th Symposium on Operating Systems Design and Implementation 16) (2016) 265–283.
[50] M. Seeger, Gaussian processes for machine learning, Int. J. Neural Syst. 14 (02) (2004) 69–106.
[51] P. Petsagkourakis, I.O. Sandoval, E. Bradford, D. Zhang, E.A. del Rio-Chanona, Reinforcement learning for batch bioprocess optimization, Comput. Chem. Eng. 133 (2020) 106649.
[52] Y. Jin, S.J. Qin, Q. Huang, V. Saucedo, Z. Li, A. Meier, S. Kundu, B. Lehr, S. Charaniya, Classification and diagnosis of bioprocess cell growth productions using early-stage data, Ind. Eng. Chem. Res. 58 (30) (2019) 13469–13480.
[53] J. Pinto, C.R. de Azevedo, R. Oliveira, M. von Stosch, A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development, Bioprocess Biosyst. Eng. 42 (11) (2019) 1853–1865.
[54] M. Karim, S. Rivera, Comparison of feed-forward and recurrent neural networks for bioprocess state estimation, Comput. Chem. Eng. 16 (1992) S369–S377.
[55] W.C. Wong, E. Chee, J. Li, X. Wang, Recurrent neural network-based model predictive control for continuous pharmaceutical manufacturing, Mathematics 6 (11) (2018) 242.
[56] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[57] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discov. Today 20 (3) (2015) 318–331.
[58] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, The rise of deep learning in drug discovery, Drug Discov. Today 23 (6) (2018) 1241–1250.
[59] D.C. Elton, Z. Boukouvalas, M.D. Fuge, P.W. Chung, Deep learning for molecular design-a review of the state of the art, Mol. Syst. Des. Eng. 4 (4) (2019) 828–849.
[60] C.W. Coley, W.H. Green, K.F. Jensen, Machine learning in computer-aided synthesis planning, Acc. Chem. Res. 51 (5) (2018) 1281–1289.
[61] M.H. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature 555 (7698) (2018) 604–610.
[62] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. Von Lilienfeld, K.-R. Muüller, A. Tkatchenko, Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. Lett. 6 (12) (2015) 2326–2331.
[63] K.T. Schütt, F. Arbabzadah, S. Chmiela, K.R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, Nat. Commun. 8 (1) (2017) 1–8.
[64] J. Yang, M.J. Knape, O. Burkert, V. Mazzini, A. Jung, V.S. Craig, R.A. Miranda-Quintana, E. Bluhmki, J. Smiatek, Artificial neural networks for the prediction of solvation energies based on experimental and computational data, Phys. Chem. Chem. Phys. 22 (42) (2020) 24359–24364.
[65] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[66] A.C. Tsoi, A. Back, Discrete time recurrent neural network architectures: a unifying review, Neurocomputing 15 (3–4) (1997) 183–223.
[67] L.X. Yu, L. Raw, L. Wu, C. Capacci-David, Y. Zhang, S. Rosencrance, FDA's new pharmaceutical quality initiative: knowledge aided assessment & structured applications, Int. J. Pharm. 1 (2019) 1–4.
[68] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, Proc. Mach. Learn. Res. 32 (2014) 1764–1772.
[69] C.L. Giles, S. Lawrence, A.C. Tsoi, Noisy time series prediction using recurrent neural networks and grammatical inference, Mach. Learn. 44 (1–2) (2001) 161–183.
[70] I. Maqsood, M.R. Khan, A. Abraham, An ensemble of neural networks for weather forecasting, Neural Comput. Appl. 13 (2) (2004) 112–122.
[71] Y.B. Dibike, P. Coulibaly, Temporal neural networks for downscaling climate variability and extremes, Neural Netw. 19 (2) (2006) 135–144.
[72] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Cognit. Model. 5 (3) (1988) 1.
[73] M.C. Mozer, A focused back-propagation algorithm for temporal pattern recognition, Complex Syst. 3 (4) (1989) 349–381.
[74] F. Chollet, et al., Keras, 2015. https://keras.io.
[75] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner,

I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, Software Available From tensorflow.org, 2015. http://tensorflow.org/.

[76] D.P. Kingma, J. Ba, Adam, A Method for Stochastic Optimization, 2014 (arXiv Preprint), arXiv:1412.6980.

[77] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[78] L. Prechelt, Automatic early stopping using cross validation: quantifying the criteria, Neural Netw. 11 (4) (1998) 761–767.

[79] G.E. Box, G.M. Jenkins, G.C. Reinsel, G.M. Ljung, Time Series Analysis: Forecasting and Control, John Wiley & Sons, 2015.

[80] B.J. Berne, J.P. Boon, S.A. Rice, On the calculation of autocorrelation functions of dynamical variables, J. Chem. Phys. 45 (4) (1966) 1086–1096.

[81] J. Smiatek, Osmolyte effects: impact on the aqueous solution around charged and neutral spheres, J. Phys. Chem. B 118 (3) (2014) 771–782.

[82] A. Narayanan Krishnamoorthy, C. Holm, J. Smiatek, Local water dynamics around antifreeze protein residues in the presence of osmolytes: the importance of hydroxyl and disaccharide groups, J. Phys. Chem. B 118 (40) (2014) 11613–11621.

[83] M. Brunner, P. Doppler, T. Klein, C. Herwig, J. Fricke, Elevated pco2 affects the lactate metabolic shift in cho cell culture processes, Eng. Life Sci. 18 (3) (2018) 204–214.

[84] X. Pan, C. Dalm, R.H. Wijffels, D.E. Martens, Metabolic characterization of a cho cell size increase phase in fed-batch cultures, Appl. Microbiol. Biotechnol. 101 (22) (2017) 8101–8113.

[85] R.J. Masterton, C.M. Smales, The impact of process temperature on mammalian cell lines and the implications for the production of recombinant proteins in cho cells, Pharm. Bioprocess. 2 (1) (2014) 49–61.

[86] S.R. De Groot, P. Mazur, Non-Equilibrium Thermodynamics, Dover Publications, 1984.

[87] A. Moser, C. Appl, S. Bruning, V.C. Hass, Mechanistic mathematical models as a basis for digital twins, Adv. Biochem. Eng. Biotechnol. (2020), https://doi.org/10.1007/10_2020_152.