**Supplementary Information**

***Fingerprint detection and process prediction by multivariate analysis of fed-batch monoclonal antibody cell culture data***

**I. Section S1**

**II. Supplementary Tables (S1 to S3)**

**III. Supplementary Figures (S1 to S5)**

**I. Section S1**

Six different ways of performing the regression analysis from section 3.3 are presented in Table S1. In the first model (denoted A1) the regression analysis was carried out using all rows of the variable-wise unfolded data set without missing values for the auto-scaled 12 X variables and the Y variable. Model B1 is an analogous regression model incorporating the high titer batches only. Finally, the effect of missing data imputation was investigated. The first imputation method linearly interpolated the missing values in each run separately for each of the thirteen variables (denoted models A2 and B2). The second method used the iterative algorithm for data imputation presented by Walczak et al.24 for the randomly distributed missing values in the X space. Every Y evolution was fit to a logistic function, which as shown in Figure 1B closely resembles the general titer evolution profile (denoted models A3 and B3), for the imputation of all the missing Y data. Using the latter strategy the imputed missing values were also extrapolated given the known trajectory information resulting in a larger number of overall observations. The results for the different models are summarized in Table S1, where the number of observations (N) for each model is also reported.

The quality of the regression improves when the high titer batches alone are considered. This effect can be visually observed in Figure S4, where the observed titer data versus the predicted ones are shown for all batches (Figure S4A) and for the high titer batches only (Figure S4B). In the latter case, points are deviating less from the diagonal, indicating that the more uniform pattern can be better captured by a PLSR model. Table S1 also shows that the prediction performance in this data set cannot be improved by neither of the methods of missing data imputation. In the case of interpolation, the prediction error remains almost unchanged, even if the number of observations is substantially larger. In the case of recovery by PCA and logistic functions, it even turns out to be detrimental to model predictability. One of the main reasons for this is likely to be the extension the imputation of the missing data in the evolutions to extrapolation. It is a good example for the trade-off between further information (data rows) which can be used in a model and the inaccuracy which is added by imputing the missing values in those rows.

II. Supplementary Tables

**Table S1** Quality of PLSR models applied to different (variable-wise unfolded) data sets. N denotes the number of observations, RMSECV describes the model accuracy as the root mean square error in cross-validation in the scale of the titer (g/L) and Q2 describes the explained variance in cross-validation. The three columns are related to the number of latent variables used in the model. For all six regression models, the incorporation of two latent variables can be considered as a good representation of the systems as only little improvement is achieved by additional latent variables.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Method \ Quality | N | RMSECV | | | Q2 | | |
| 1 | 2 | 3 | 1 | 2 | 3 |
| A1) All batches | 555 | 0.138 | 0.097 | 0.094 | 0.59 | 0.80 | 0.81 |
| A2) Imputation by interpolation | 991 | 0.128 | 0.094 | 0.092 | 0.58 | 0.77 | 0.78 |
| A3) Imputation by IA and logistic function | 1377 | 0.121 | 0.104 | 0.100 | 0.63 | 0.72 | 0.75 |
| B1) High titer batches | 441 | 0.092 | 0.083 | 0.078 | 0.82 | 0.85 | 0.87 |
| B2) Imputation by interpolation | 782 | 0.090 | 0.081 | 0.076 | 0.80 | 0.84 | 0.86 |
| B3) Imputation by IA and logistic function | 1063 | 0.108 | 0.096 | 0.089 | 0.74 | 0.79 | 0.82 |

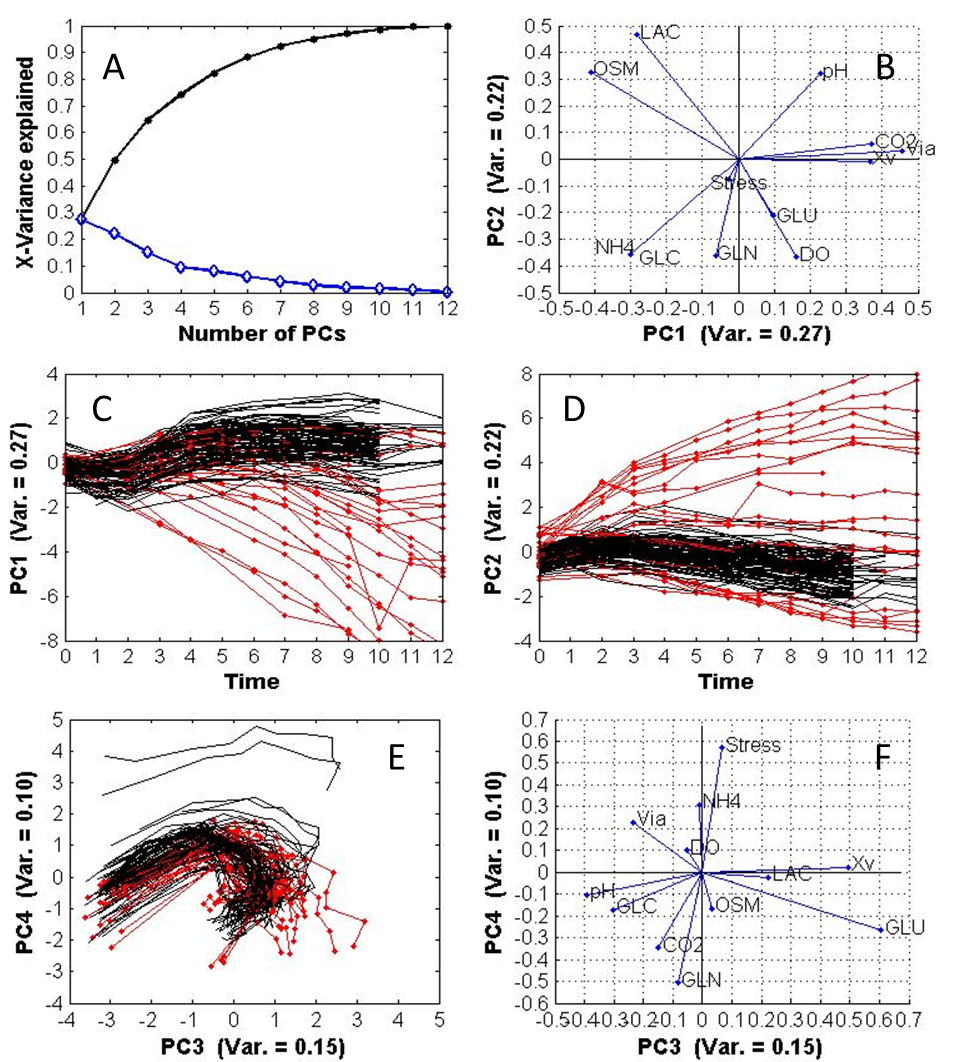
Table S2 Selection scheme of variables (columns) distinguished according to time (rows) by the genetic algorithm of Leardi et al.31, which was run for 10 times. The variables selected six times or more were considered as important (marked by cross). The controlled (time-invariant) variables DO and stress were never selected.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | pCO2 | pH | | GLC | | LAC | | GLN | | GLU | | NH4 | | OSM | | | Xv | | Via | |
| 0 |  |  | |  | |  | |  | |  | | **x** | |  | | |  | | x | |
| 1 | **x** |  | |  | |  | |  | |  | | **x** | |  | | |  | |  | |
| 2 |  |  | |  | |  | |  | |  | | **x** | |  | | |  | |  | |
| 3 |  | **x** | |  | |  | |  | |  | |  | |  | | |  | |  | |
| 4 |  |  | | **x** | |  | |  | |  | |  | |  | | | **x** | |  | |
| 5 |  |  | |  | |  | |  | |  | |  | |  | | | **x** | |  | |
| 6 |  |  | |  | |  | |  | |  | |  | |  | | | **x** | |  | |
| 7 |  |  | |  | | **x** | |  | |  | |  | |  | | | **x** | |  | |
| 8 |  |  | |  | |  | |  | | **x** | |  | |  | | | **x** | |  | |
| 9 | **x** |  | |  | | **x** | |  | | **x** | |  | | **x** | | | **x** | | **x** | |
| 10 | **x** | **x** | | **x** | |  | | **x** | | **x** | |  | | **x** | | | **x** | | **x** | |
| Table S3 RMSECV in g/L for the prediction of process titer at different points of time (columns) using historic data for the variables. The rows indicate the time period employed for all the X variables. The subscript represents the time point or period and the number of predictors incorporated in the model is shown in brackets. Those models incorporate further additional variables for pH and DO extending those variables to the profile observed in Figure 2: pHext(t) = (pH(t)-mean(pH(t)))2 and DOext = (DO-50)2. | | | | | | | | | | | | | | | | | | | | | | | |
|  | | | | Y0 | | Y1 | | Y2 | | Y3 | | Y4 | | Y5 | | Y6 | Y7 | | Y8 | | Y9 | | Y10 |
| X0 (12) | | | | 0.053 | | 0.052 | | 0.057 | | 0.050 | | 0.044 | | 0.046 | | 0.052 | 0.063 | | 0.082 | | 0.098 | | 0.109 |
| X0-1 (22) | | | |  | | 0.048 | | 0.053 | | 0.047 | | 0.039 | | 0.044 | | 0.046 | 0.060 | | 0.074 | | 0.092 | | 0.106 |
| X0-2 (32) | | | |  | |  | | 0.054 | | 0.046 | | 0.041 | | 0.042 | | 0.042 | 0.056 | | 0.069 | | 0.081 | | 0.086 |
| X0-3 (42) | | | |  | |  | |  | | 0.042 | | 0.038 | | 0.035 | | 0.032 | 0.046 | | 0.056 | | 0.067 | | 0.077 |
| X0-4 (52) | | | |  | |  | |  | |  | | 0.038 | | 0.034 | | 0.030 | 0.040 | | 0.048 | | 0.061 | | 0.073 |
| X0-5 (62) | | | |  | |  | |  | |  | |  | | 0.033 | | 0.026 | 0.036 | | 0.041 | | 0.055 | | 0.063 |
| X0-6 (72) | | | |  | |  | |  | |  | |  | |  | | 0.024 | 0.034 | | 0.038 | | 0.051 | | 0.063 |
| X0-7 (82) | | | |  | |  | |  | |  | |  | |  | |  | 0.034 | | 0.038 | | 0.051 | | 0.061 |
| X0-8 (92) | | | |  | |  | |  | |  | |  | |  | |  |  | | 0.037 | | 0.050 | | 0.058 |
| X0-9 (102) | | | |  | |  | |  | |  | |  | |  | |  |  | |  | | 0.049 | | 0.057 |
| X0-10 (112) | | | |  | |  | |  | |  | |  | |  | |  |  | |  | |  | | 0.054 |

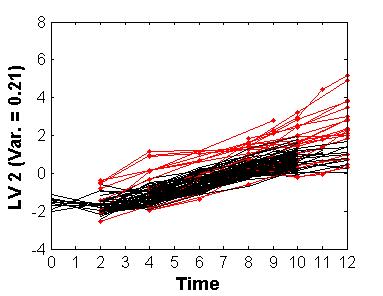
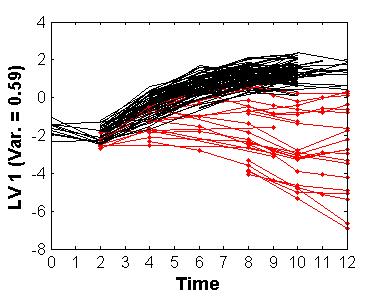
III. Supplementary Figures



**Figure S1** Time evolutions of viability, lactate concentration, osmolality, viability, glutamine, partial pressure of CO2 and ammonium concentration. The low titer batches are visualized by dot-markers.



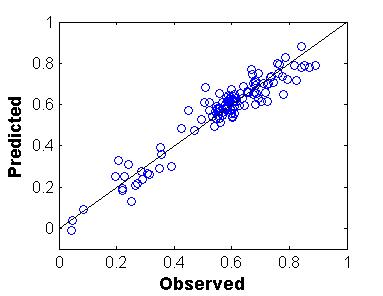
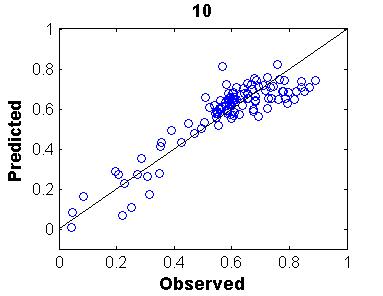
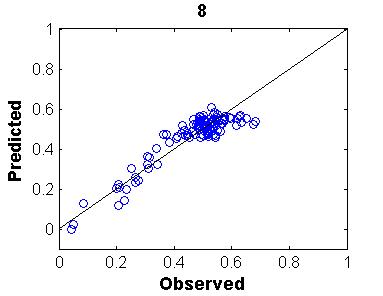
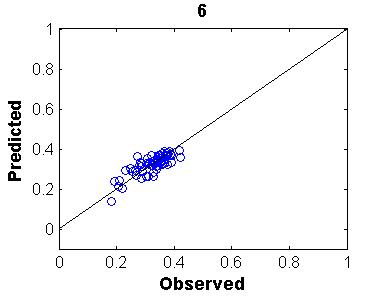
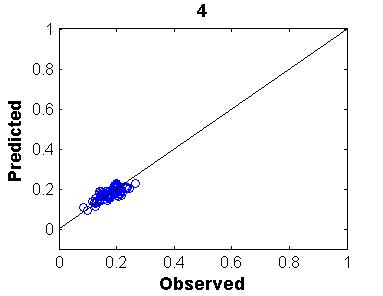
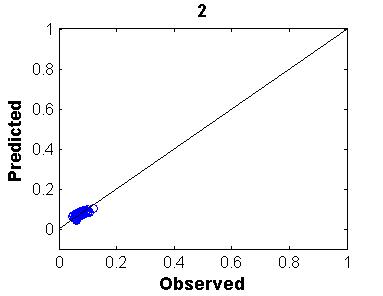
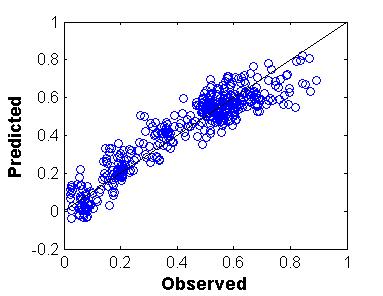
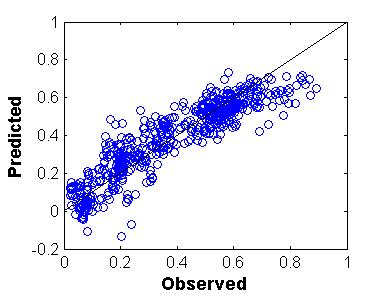
**Figure S2** Supplements to PCA Analysis in section 3.2: Cumulative (dots) and specific (diamonds) variance explained by PCs (A), Loading plot for the first two PCs (B), Time evolutions of scores of first (C) and second PC (D), Score plot showing trajectories (day 0 on the left) of low and high titer batches (E), Loading plot for the third and forth PCs (F). The low titer batches are visualized by dot-markers. The number in brackets refers to the variance explained by the respective PC.



A

B

**Figure S3** Time evolutions of scores of first (A) and second LV (B). The low titer batches are visualized by dot-markers. The number in brackets refers to the variance explained by the respective LV.



A

B

C

D

E

F

G

H

Figure S4 Comparison of observed and predicted titer (in g/L) using variable unfolded data with all (A, model A1 in Table S1) and only high titer batches (B, model B1 in Table S1), using time-specific models for the days 2, 4, 6, 8 and 10 (C to G, shown in Table 2) and using batch-wise unfolded data for prediction of the titer at day 10 with the full history (H, final model along the diagonal of Table 3).



Figure S5 VIP values for the variables pCO2, DO, pH, Stress, concentration of ammonium, Xv, concentrations of lactate and glutamine in the time-specific models for days 2, 4, 6, 8 and 10 (shown in Table 2). The solid line demonstrates the results in the overall observation evolution model (model A1 in Table S1).