Abschlussbericht Kennziffer 2627

Monitoring und Kontrolle von Bioreaktionsprozessen zur Herstellung heterologer Proteine mit *Escherichia coli*

Förderperiode 2005/2007

Kurzfassung: Es wurde ein komplexes Bioreaktorsystem zur Beobachtung und Führung bioverfahrenstechnischer Prozesse aufgebaut. Für ein robustes Monitoring wurden basierend auf der 2D-Fluoreszenzspektroskopie Künstliche Neuronale Netze für die Online-Überwachung wichtiger biologischer Schlüsselgrößen trainiert. Dabei wurden die Hauptkomponentenanalyse sowie die unabhängige Komponentenanlyse für die Datenreduktion eingesetzt und die erzielten Ergebnisse miteinander verglichen.

1 Einleitung

Der Einsatz von rekombinanten Mikroorganismen in der produzierenden Industrie nimmt immer mehr zu. Gerade im Bereich biotechnologischer Pharmastoffe (API[1]'s) unterliegen die Herstellungsprozesse jedoch strikten Regularien. Viele der in der Forschung und Entwicklung eingesetzten Messsysteme zur Identifikation und Überwachung von Bioprozessen bieten nicht die notwendige Robustheit und/oder genügen nicht den strengen Ansprüchen von GMP[2]-gemäßen Herstellungsprozessen, was die Überwachung und Qualitätssicherung im Produktionsvorgang erheblich erschwert.

Im Rahmen dieses Projektes konnten unter Verwendung eines hochinstrumentierten Forschungsbioreaktors die Prozessgrößen Zelldichte (c_{XL}), Glukosekonzentration (c_{S1M}) und die Zielproteinkonzentration (c_{P1L}) online dargestellt werden. Basierend auf der 2D-Fluoreszenzspektroskopie konnte unter Einsatz von Künstlichen Neuronalen Netzen (KNN oder ANN[3]) ein robustes Monitoring realisiert werden. Dabei wurden unterschiedliche chemometrische Methoden zur Reduktion des Eingangs-datenraumes verwendet und die damit erzielten Ergebnisse verglichen und beurteilt.

2 Material

Alle Kultivierungen wurden in einem 15 l BIOSTAT® C (Sartorius Stedim Biotech GmbH, Melsungen) Edelstahlreaktor durchgeführt. Das instrumentelle Setup ist in Abb. 1 schematisch dargestellt. Neben der Standardinstrumentierung für Temperatur, Druck, Gelöstsauerstoffkonzentration und pH-Wert, wurden auch die Trübung, die Leitfähigkeit und Impedanz in der Kulturbrühe (inline) sowie die Ammoniumkonzentration, die Glucosekonzentration und die Acetatkonzentration in direkter Anbindung an den Prozess (atline) überwacht. Für die Anwendung von Massenbilanzen und Online-Berechnungen wurden die Massen des Fermenters, der Vorlagen, der Korrekturmittel und der zellfreien Probenahme für die Atline-Analytik sowie die Probenahme mittels Oberschalenwaagen vom SCADA[4]-System MFCS/win aufgezeichnet.

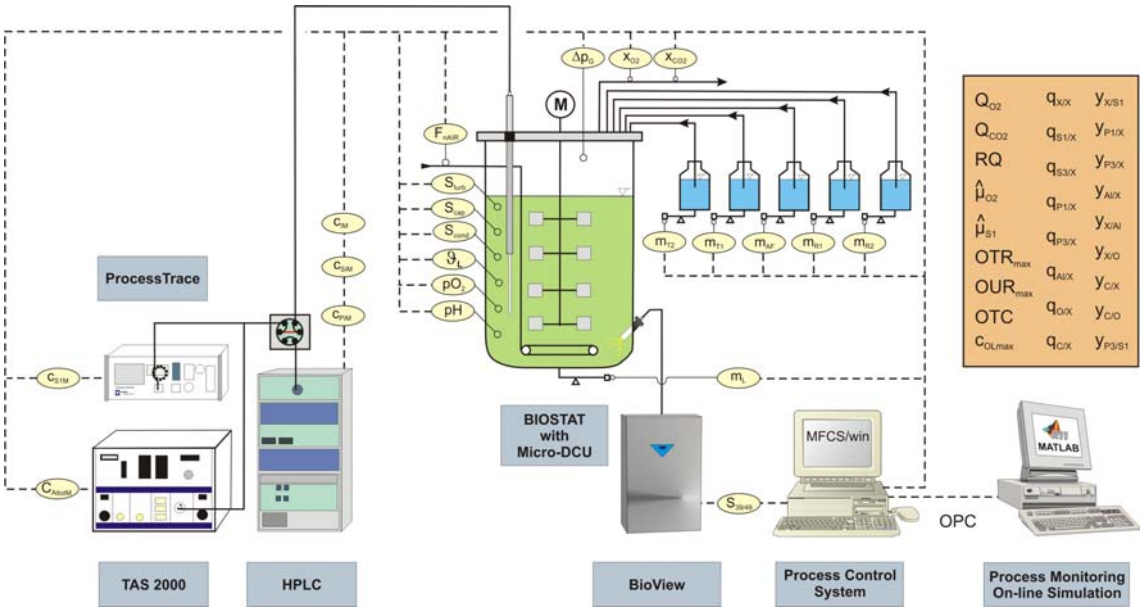

Abb. 1: Setup der Forschungsbioreaktors BIOSTAT® C

[1] API := Active Pharmaceutical Ingredients
[2] GMP := Good Manufacturing Practice
[3] ANN := Artificial Neural Networks
[4] SCADA := Supervisory Control and Data Acquisition

Alle Kultivierungen wurden auf M9 Minimalmedium mit Glukose als Hauptsubstrat durchgeführt. Es wurde ein *Escherichia coli* DH5α-Stamm (6897, DSMZ Braunschweig) mit einem pTrcHisB Plasmid (Invitrogen) verwendet, in welches die Sequenz für das rekombinante Protein T-Sapphire-GFP integriert wurde und der mit IPTG[5] induzierbar ist.

Die in diesem Projekt eingesetzte 2D-Fluoreszenzspektroskopie (BioView, DELTA, Hørsholm, Dänemark) genügt den industriellen Ansprüchen an ein robustes Messgerät. Diese Messung findet inline statt. Sie ist nicht invasiv und nicht produktberührend. Somit ist der unproblematische Einsatz in einer GMP-Umgebung gewährleistet.

Der eingesetzte BioView-Sensor verfügt über zwei Filterräder, von denen eines für die Excitation und das andere für die Emissionsmessung verwendet wird. Die Anregungswellenlängen liegen zwischen 270 und 550 nm, wobei der Abstand sowie die Bandbreite jeweils 20 nm betragen. Die Emissionsmessung erfolgt zwischen 310 und 590 nm mit dem gleichen Abstand und der gleichen Bandbreite wie für die Anregung. Ausserdem stehen noch Neurtraldichtefilter (ND) mit einer Transmission oberhalb von 450 nm für die Excitation (ExND) sowie für die Messung des wellenlängenabhängigen Streulichts (NDEm) zur Verfügung. Insgesamt ergeben sich damit 150 Wellenlängenkombinationen. Das sich ergebende Fluoreszenzspektrum mit den entsprechenden Fluoreszenzbereichen ist in Abb. 2 dargestellt.

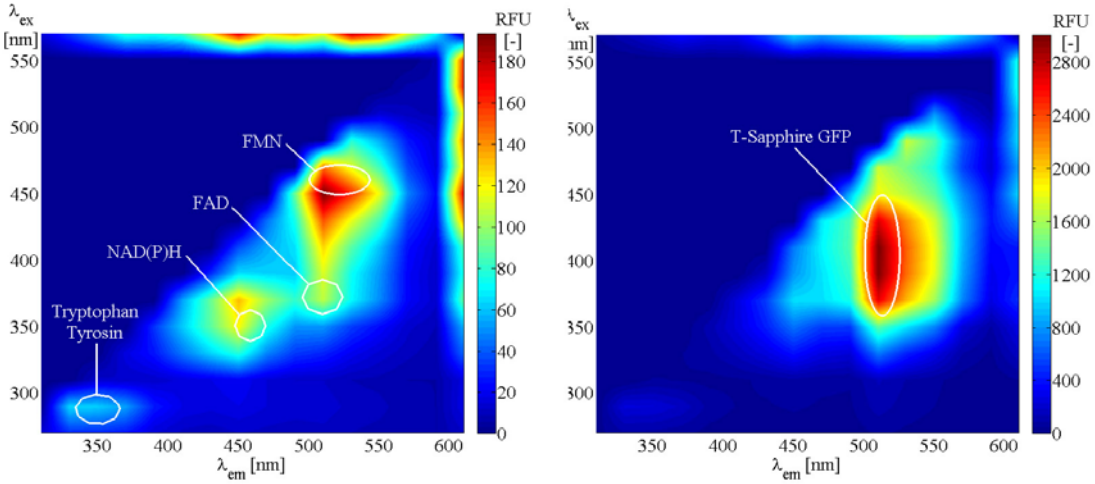

Abb. 2: Darstellung der ausgewählten Fluoreszenzbereiche im 2D-Fluoreszenzspektrum
links: Batch-Phase; rechts: Produktion während der Fed-Batch Phase

3 Theoretische Grundlagen

Durch die gewonnenen BioView-Daten sind fluoreszierende Komponenten in der Kulturbrühe im Spektrum direkt sichtbar. Auf Grund von Sättigungseffekten (Innerfilter-Effekt), Überlagerung der Fluoreszenzantworten unterschiedlicher Komponenten aber auch die Excitation von Fluorophoren durch die Emission anderer Komponenten (Kaskadeneffekt) und die starke Abhängigkeit der Fluoreszenz von den Umgebungsbedingungen (pH, Temperatur, pO_2), ist die direkte Interpretation der Signale nicht immer möglich.

Die Komponenten, die keine Eigenfluoreszenz besitzen, sind teilweise über die an Stoffwechselwegen beteiligten Fluorophore bestimmbar. Da diese Komponenten an mehreren Stoffwechselwegen gleichzeitig mitwirken, erfordert die Erstellung strukturierter Modelle einen erheblichen analytischen Aufwand

[5] IPTG := Isopropylthiogalactosid

Der Einsatz von Künstlichen Neuronalen Netzen bietet sich überall dort an, wo explizites Wissen nicht in ausreichendem Maße vorhanden ist und implizites Wissen in Form einer ausreichenden Anzahl an Datensätzen vorhanden ist.

Für nichtlinear separierbare und komplexe Probleme werden mehrschichtige Netze eingesetzt, so genannte Multi Layer Perceptrons (MLP). Diese Netze besitzen eine nichtlineare Schicht mit nichtlinearen Transferfunktionen (z.B. Sigmoide), die eine Projektion des nichtlinear separierbaren Problems zu einem linear separierbaren Problem ermöglichen, welches von der Ausgangsschicht, die wiederum linear ist, separiert werden kann. Diese MLP´s werden auch als Universalapproximator bezeichnet, da es möglich ist, nahezu jede Funktion damit zu approximieren.

Da bei steigender Dimension des Eingangsdatenraumes der Bedarf an Referenzvektoren exponentiell steigt, müssen geeignete Methoden zur Datenreduktion eingesetzt werden. Die hier eingesetzten Methoden zur Datenreduktion sind zum Einen die Hauptkomponentenanalyse (PCA[6]) und zum anderen die Independent Component Analysis (ICA). Bei Letzterer handelt es sich um einen etablierten Algorithmus aus der „blind source separation“. Die ICA wird bereits in vielen Bereichen der Technik erfolgreich zur Quellentrennung eingesetzt

In der Hauptkomponentenanalyse wird die ursprüngliche Datenmatrix X in zwei kleinere, orthogonale Matrizen, die Hauptkomponenten T und die Ladungen L,

$$X = T \cdot L^T , \tag{1}$$

zerlegt.

Die Daten werden dadurch so rekonstruiert, dass sich neue, unkorrelierte Variablen ergeben. Die Hauptkomponenten werden nach dem Kriterium maximaler Varianz bestimmt. Sie geben dabei den Verlauf der Ursprungsdaten in einem neuen Koordinatensystem wieder. Die Ladungen enthalten die Information über den Ursprungsort (hier die Wellenlängenkombinationen) der Fluoreszenzdaten.

Die Ladungsmatrix L wurde unter MATLAB unter Verwendung der Single Value Decomposition (SVD) mit zentrierten Daten durchgeführt.

Geht man davon aus, dass sich die 2D-Fluoreszenzspektren x_j als Mischung aus n unabhängigen Komponenten s_i mit den Gewichtungsfaktoren a_{ij} zusammensetzen, so kann der Datenraum der BioView-Spektren über

$$x = A \cdot s \tag{2}$$

beschrieben werden.

Mit Kenntnis von A wäre es ein Leichtes, die Signale s mit $W = A^{-1}$ aus den Messwerten x über

$$s = W \cdot x \tag{3}$$

zu rekonstruieren.

---

[6] PCA := Principal Component Analysis

Da nun A aber unbekannt ist, muss W anders berechnet werden. Das Problem dabei ist, dass die Komponenten von x meist stark korreliert sind. Daher muss ein Verfahren gefunden werden, um diese zu dekorrelieren. Dieses erfolgt entweder mittels PCA, wobei hier Statistiken 2. Ordnung (Kovarianzmatrix) verwendet werden, oder besser über die ICA, welche Statistiken aller Ordnungen zulässt. Damit erreicht man nicht nur das Entkoppeln aller Korrelationen sondern auch eine weitestgehend stochastische Unabhängigkeit der Komponenten.

Die ICA wurde mit dem im Internet frei verfügbaren ICASSO-Algorithmus unter MATLAB iterativ ausgeführt.

4 Ergebnisse

Für das Training der Netze standen acht Kultivierungen zur Verfügung. Die optimale Netzstruktur wurde iterativ über die Minimierung des RMSP[7] bestimmt. Bei der PCA wurden zunächst die Hauptkomponenten aller Kultivierungen gemittelt. Da bereits die ersten 3 PCs[8] 99 % der in den Messdaten enthaltenen Varianz beschreiben, wurden für das Training der Netze maximal die ersten vier Hauptkomponenten verwendet. Der Suchraum zur Auffindung des optimalen Netzes bestand aus 1 – 4 PCs und 1 - 50 nichtlineare Zwischenschichtneuronen.

Bei Einsatz der ICA ist die Mittelwertbildung aus allen Komponenten nicht möglich. Da sowohl s als auch A (Gl. (2)) unbekannt sind, kann jeder skalare Faktor in s in A wieder ausgeglichen werden. Die unabhängigen Komponenten müssen deshalb für eine Kultivierung bestimmt werden und dann auf alle restlichen Daten angewendet werden. Der Suchraum für die optimale Netzarchitektur lag bei 2 – 6 ICs und 1 - 50 nichtlineare Zwischenschichtneuronen.

Die Validierung der gewählten neuronalen Netze wurde während zwei Kultivierungen durchgeführt. Die Rekonstruktion der Zielproteindaten für einen Prozess ist in Abb. 3 dargestellt. Beide eingesetzten Verfahren führen zu guten Ergebnissen, wobei sowohl beim Training als auch bei der Validierung die ICA einen geringeren RMSP aufweist. Der Fehlervergleich beider Methoden und die gewählte Netzarchitektur sind in Tab. 1 und Tab. 2 angegeben.

Tab. 1: Fehler der verwendeten neuronalen Netze

Prozessgröße	Methode	PCA	ICA
c_{P1L}	Training	31.15	15.49
	Validation	43.74	23.50
c_{XL} 0 – 30 gl^{-1}	Training	0.3488	0.3959
	Validation	0.8487	0.5643
c_{XL} 30 – 70 gl^{-1}	Training	1.664	0.6082
	Validation	3.705	3.897
c_{S1M}	Training	0.7924	0.6410
	Validation	0.8924	0.9324

[7] RMSP := Root Mean Square Error of Prediction
[8] PC := Principal Component

Tab. 2: Gewählte Netzarchitektur

Prozessgröße	Methode	PCA	ICA
c_{P1L}	Anzahl Zwischenschichtneuronen	3	5
	Anzahl Eingangsvektoren	3	6
c_{XL}	Anzahl Zwischenschichtneuronen	10	6
0 – 30 gl^{-1}	Anzahl Eingangsvektoren	4	6
c_{XL}	Anzahl Zwischenschichtneuronen	2	2
30 – 70 gl^{-1}	Anzahl Eingangsvektoren	4	3
c_{S1M}	Anzahl Zwischenschichtneuronen	2	2
	Anzahl Eingangsvektoren	2	4

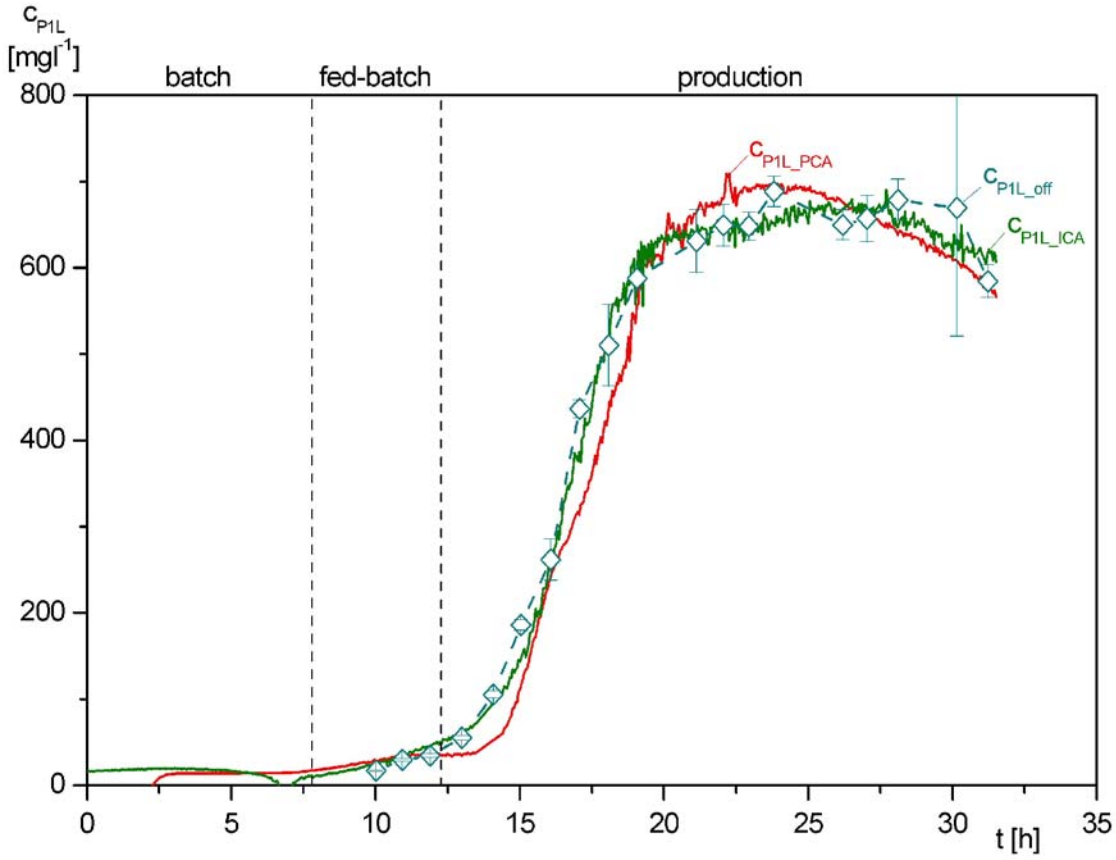

Abb. 3: Zielproteinverlauf (c_{P1L}) während eines typischen Produktionsprozesse
c_{P1L_off} := offline bestimmte Zeilproteinkonzentration
c_{P1L_PCA} := Zielproteinkonzentration über ANN bestimmt, Datenreduktion mittels PCA
c_{P1L_ICA} := Zielproteinkonzentration über ANN bestimmt, Datenreduktion mittels ICA

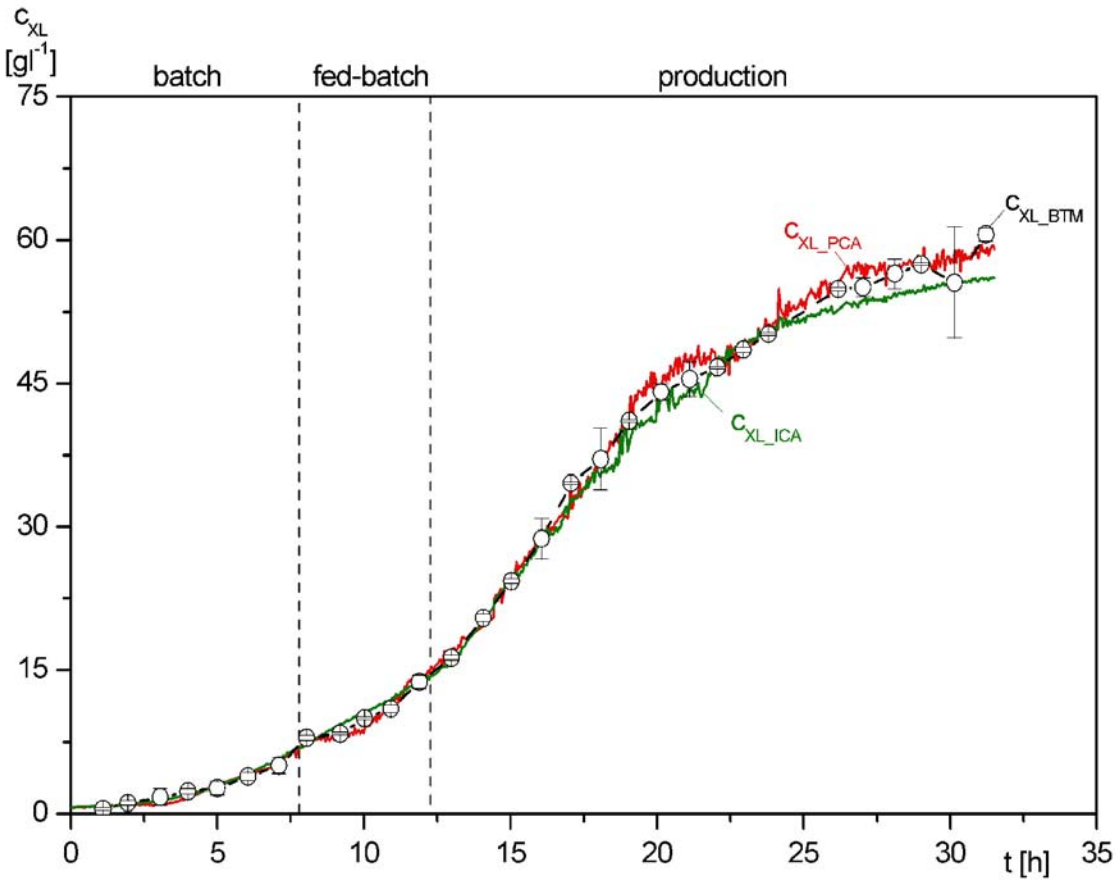

Abb. 4: Zelldichteverlauf (c_{XL}) während eines typischen Produktionsprozesse
c_{XL_off} := offline bestimmte Zelldichte
c_{XL_PCA} := Zelldichte über ANN bestimmt, Datenreduktion mittels PCA
c_{XL_ICA} := Zelldichte über ANN bestimmt, Datenreduktion mittels ICA

Für die Darstellung der Zelldichte über den gesamten Kultivierungszeitraum konnte kein Netz zufriedenstellende Ergebnisse liefern. Dies ist hauptsächlich auf die starke Überlagerung vieler Spektren durch die T-Sapphire-GFP Fluoreszenz sowie auf einsetzende Sättigungseffekte bei höheren Zelldichten zurückzuführen. Aus diesem Grund wurde je ein Netz für die Zelldichte zwischen $0 – 30\ gl^{-1}$ und ein Netz für $30 – 70\ gl^{-1}$ Biotrockenmasse erstellt. Die erzielten Ergebnisse sind in Abb. 4 dargestellt und geben den unbekannten Zelldichteverlauf online sehr gut wieder. Die Datenreduktion mittels ICA und PCA führten in diesem Fall zu etwas sehr ähnlichen Ergebnissen (Tab. 1; Abb. 4).

Zur Online-Überwachung des Hauptsubstrates Glukose während der Batch-Phase wurde ein weiteres Netz trainiert. Die Anpassung ist in Abb. 5 aufgezeigt. Die PCA basierten Ergebnisse liefern hier eine etwas exaktere Wiedergabe der Glukosekonzentration und können den Zeitpunkt des vollständigen Verbrauches der C-Quelle besser wiedergeben.

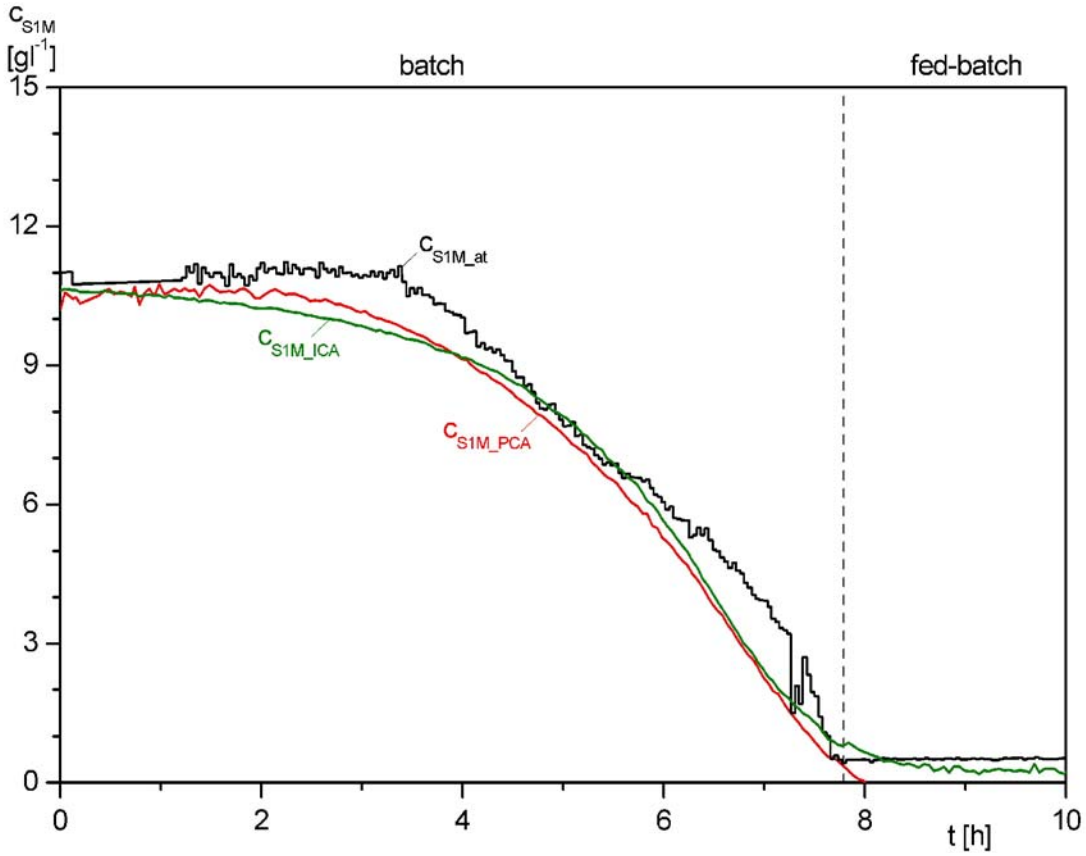

Abb. 5: Glucoseverlauf (c_{XL}) während der Batchphase
c_{S1M_at} := mittels GlucoTrace (TraceAnalytics, Braunschweig) bestimmte Glucosekonzentration
c_{S1M_PCA} := Glucosekonzentration über ANN bestimmt, Datenreduktion mittels PCA
c_{S1M_ICA} := Glucosekonzentration über ANN bestimmt, Datenreduktion mittels ICA

5 Fazit

Es konnten geeignete Netze trainiert werden, die ein robustes Online-Monitoring der Schlüsselgrößen Zelldichte, des fluoreszierenden Zielproteins über die gesamte Kultivierungsdauer, sowie der Glukosekonzentration während der Batch-Phase zulassen. Dabei konnten mit der ICA als Datenreduktionsmethode im Vergleich zur klassischen PCA etwas präzisere Ergebnisse erzielt werden.