Title	Performance Monitoring of a Mammalian Cell Based Bioprocess using Raman spectroscopy.
Author(s)	Ryder, Alan G.; Li, Boyan; Ray, Bryan H.
Publication Date	2013
Publication Information	B. Li, B.H. Ray, K.J. Leister, and A.G. Ryder (2013) 'Performance Monitoring of a Mammalian Cell Based Bioprocess using Raman spectroscopy'. Analytica Chimica Acta, 796 :84-91.
Link to publisher's version	http://dx.doi.org/10.1016/j.aca.2013.07.058
Item record	http://www.sciencedirect.com/science/article/pii/S0003267013010349; http://hdl.handle.net/10379/3938
DOI	http://dx.doi.org/10.1016/j.aca.2013.07.058

Performance Monitoring of a Mammalian Cell Based Bioprocess using Raman spectroscopy. B. Li, B.H. Ray, K.J. Leister, and A.G. Ryder. *Analytica Chimica Acta*, 796, 84-91, (2013). DOI: http://dx.doi.org/10.1016/j.aca.2013.07.058

SUPPLEMENTAL INFORMATION:

TABLE OF CONTENTS

S1. SPECTRUM PRE-PROCESSING

Raman spectra were collected [1] and then subjected to a series of sequential pre-processing techniques prior to chemometric analysis. For each measurement composed of nine spectra, spectra containing cosmic interference were discarded prior to averaging and multiplicative scatter correction [2]. Then an asymmetric weighted least squares algorithm [3] was implemented to remove baseline offsets before the individual spectra were normalized to an internal reference band (water bending vibration at 1636 cm^{-1}). Due to the fact that all the broth samples were dilute aqueous solutions with low analyte concentrations, the Raman bands of water tend to mask and distort the analyte bands. The water band was subtracted using an orthogonal projection procedure [4]. The resulting data exhibit less extraneous sources of variance and more linear response of the variables. The first derivative transform [5] was finally performed to further improve the data for PLS analysis by accentuating analyte signals while reducing baseline artifacts.

S2. ADDITIONAL SPECTRA

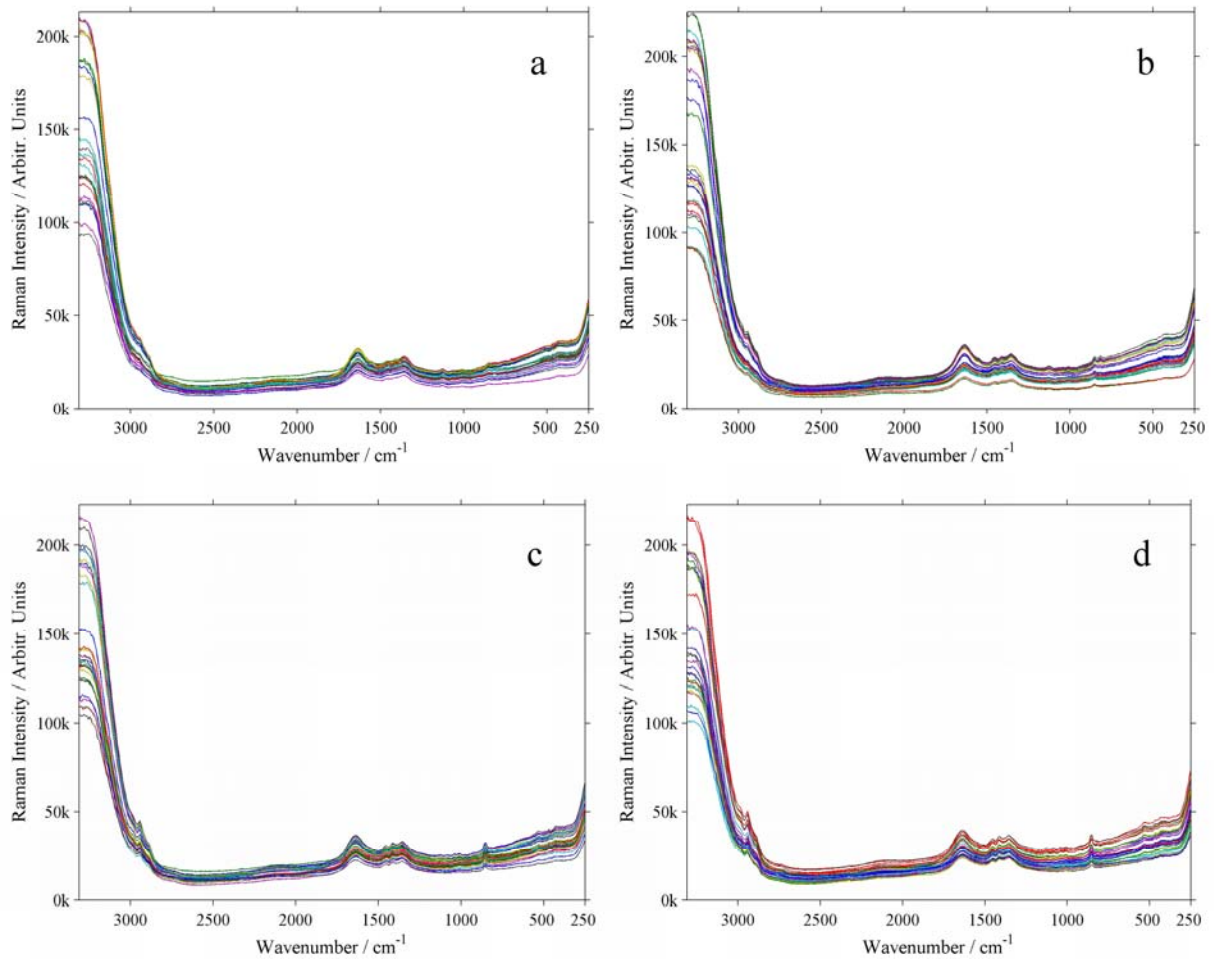

Figure S-1: Raw Raman spectra collected from the final bioreactor stage in the range of 3311–250 cm^{-1}: (**a**) After inoculation (DS9), (**b**) 5-day post inoculation (DS10), (**c**) 10-day post inoculation (DS11), and (**d**) Prior to transfer (DS12).

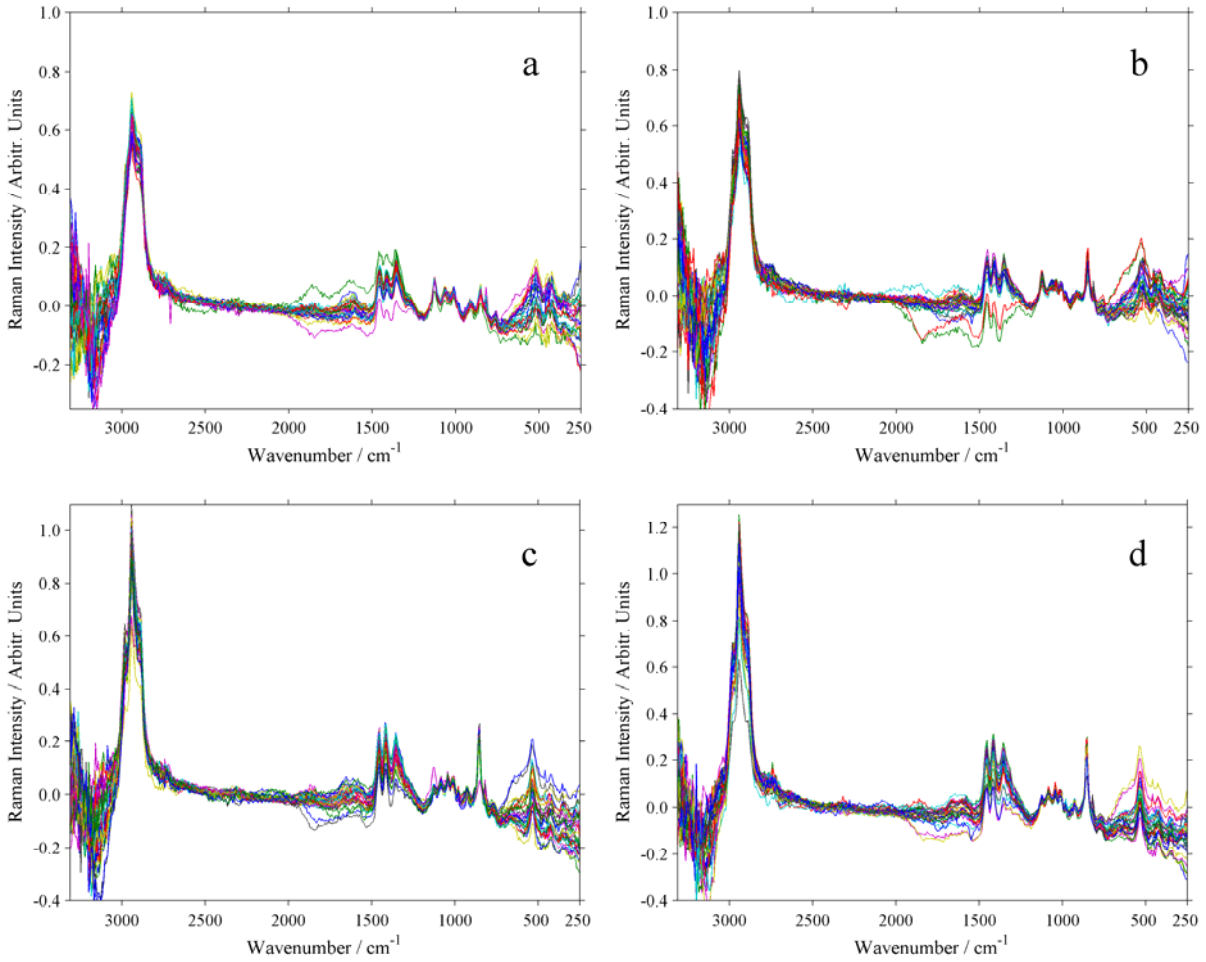

Figure S-2: Baseline-corrected and water subtracted Raman spectra collected in the range of 3311–250 cm^{-1} from the final bioreactor stage: (**a**) After inoculation (DS9), (**b**) 5-day post inoculation (DS10), (**c**) 10-day post inoculation (DS11), and (**d**) Prior to transfer (DS12).

S3. QUANTITATIVE ANALYSIS: DETERMINATION OF PLS COMPONENTS

The accurate selection of PLS components (latent variables or factors or model dimensionality/rank) is very crucial to constructing optimal models. Retaining too few components implies that the calibration data are under-fitted and there is still information left that can be modeled. Choosing too many components will lead to over-fitting. This means that although the calibration data are well described, the model will have poor predictive ability.[6] Many methods have been developed to tackle the over-fitting problem of which the best is to obviously use a sufficiently large independent validation/test sample set. However, sufficient validation samples are not always available and in practice, PLS component selection is

frequently carried out using some sort of validation, such as leave-one-out (LOO) and r-fold cross validation (CV).[2]

The CV-based method economically handles the available data, but like any data-based statistical test gives an interval results and hence sometimes gives either an under-fit or an over-fit, that is they reach the minimum RMSEV for a lower or higher model rank than would be achieved using an independent validation set. Generally, the CV method does not compensate for the fact that the same samples are used for both calibration and validation, therefore, the method has a tendency to select too many components, and as a result, over-fitting the calibration data is an issue. Recently Monte Carlo (MC) cross validation has been introduced to reduce the risk of over-fitting.[6] However, a major concern with this method is whether making a calibration model with a small number of samples will be truly representative. Another issue to consider is the fact that with CV based methods; the RMSEP estimates do not often exhibit a clear global minimum, because they carry a considerable uncertainty with respect to bias and variance. Therefore, for complex low-sample number systems it is not always clear as to what is the optimal number of PLS components.

In our study, we had a limited number of samples (<37) made available to us, and therefore we first used both LOO and MC cross validation methods to determine the number of PLS components for each model. The resultant RMSECs showed very low values, around half of the RMSEPs/RMSEVs. Taking the DS4 sample set as example, Figure S-3 and Table S-1 shows the simplest LOOCV result for the selection of the PLS components using both the CoAdReS and ACO refined variables respectively.

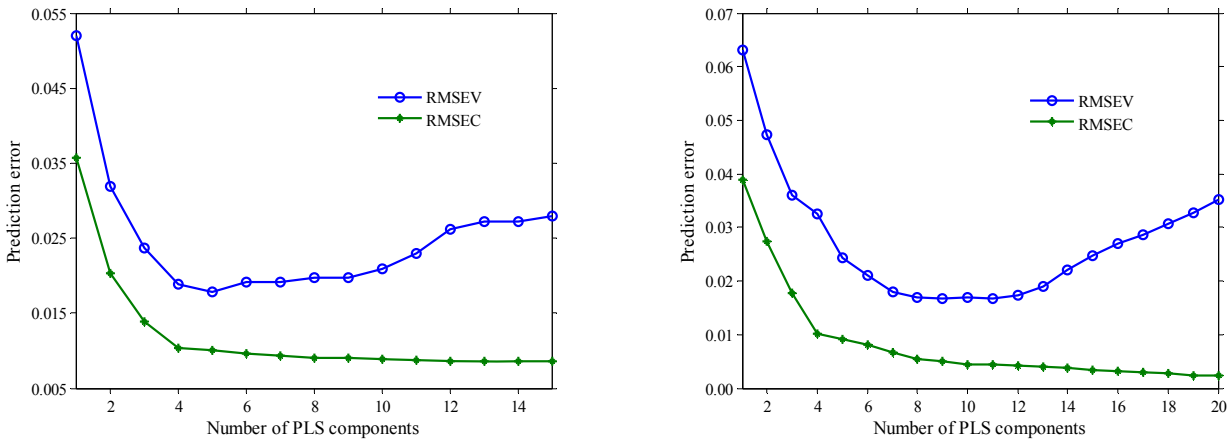

Figure S-3: Selection of the PLS components using LOOCV with the CoAdReS (left) and ACO (right) refined variables for DS4 sample set.

Table S-1: Variance (in %) explained by each PLS component, obtained from the LOOCV implementation on the CoAdReS and ACO selected variables of the DS4 sample set.

# LVs	Percentage variance (%) explained by PLS components

	CoAdReS-selected variables	Titre	ACO-selected variables	Titre
1	15.44	65.19	12.83	58.62
2	19.25	23.55	23.22	20.69
3	15.25	6.03	12.41	12.04
4	9.16	2.28	5.67	5.88
5	**13.79**	**0.20**	16.95	0.53
6	10.01	0.18	**9.11**	**0.44**
7	5.73	0.17	3.24	0.56
8	2.76	0.13	2.70	0.44

One can observe that, the RMSEP/RMSEV reached a clear global minimum at the 5th PLS component with the CoAdReS-refined variables, while for the ACO-refined variables 8 PLS components were suggested for constructing PLS model because a small plateau began from the 8th PLS component. The implementation of MCCV gave similar results (Figure S-4). Although ad hoc minimum RMSEPs were obtained for most of the models, based on the PLS component selection with the CV method, the RMSEPs values are however, overly optimistic due to over-fitting.

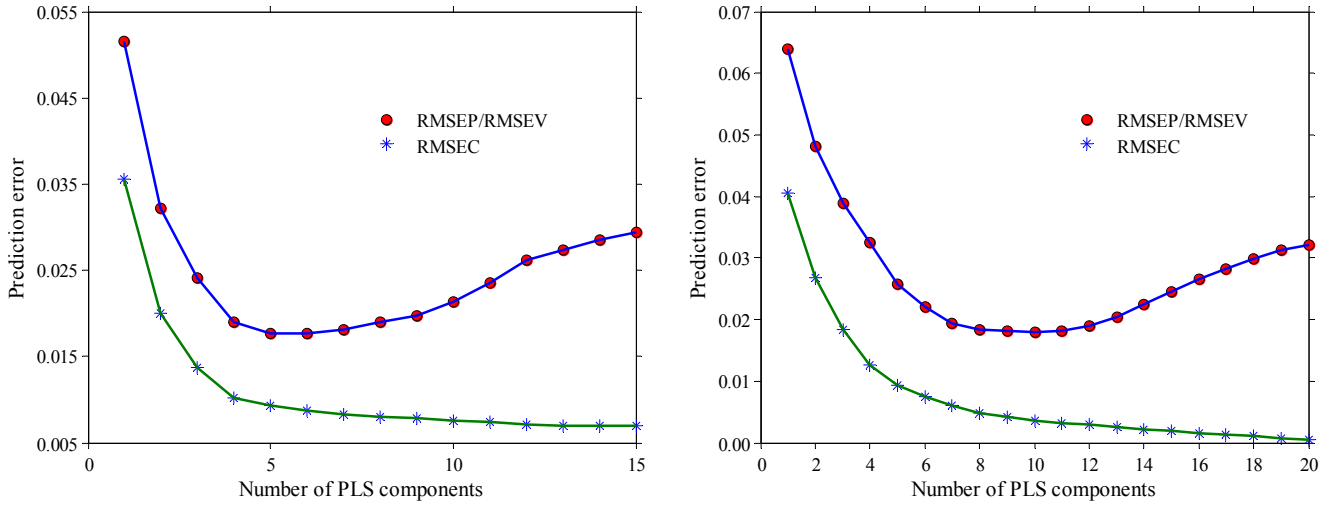

Figure S-4: Selection of the PLS components using MCCV with the CoAdReS (left) and ACO (right) refined variables for DS4 sample set. 5 validation samples were randomly used for each of 500 MC repeating runs.

Next we implemented a newly proposed distribution-based statistical test method, *i.e.*, randomization test [7], to determine the PLS components for each sample set, and thus generate more correct PLS models. In contrast to CV, this pragmatic data-driven approach assesses the statistical significance of each individual component that enters the PLS model, with no

requirement to exclude any data, and thus avoid over-fitting problem related to data exclusion. The method principle was briefly described as follows.

The randomization method scrambles the elements of Y (property of interest, *e.g.*, glycoprotein product titer in this paper) while keeping the corresponding values in X (*e.g.*, spectra), and hence destroys any relationship that might exist between the X- and Y-variables. A series of test PLS regression models were then generated that should reflect the absence of a real association between the X- and Y-variables. The covariance between the PLS components and the Y-value is calculated as a test statistic for each of these PLS models, which should be indistinguishable from a chance fluctuation (*i.e.*, a null-distribution of noise value), except for the small remaining correlation to the original Y. More important is that a critical value is derived from the null-distribution as the value exceeded by a certain percentage of noise values (say 5 or 10%). This critical value follows as a percentage point of a data-driven histogram (which is generated by repeating the randomization and calculation a number of times, *e.g.*, 1000 randomizations) of noise values. The statistic obtained for the original data—the value under test—is then compared with the critical value. From this comparison one can much more clearly decide if the PLS factor is significant or not.

The DS4 sample set is shown as an example of how the randomization test method performed PLS component selection for both the CoAdReS and ACO selected variables. 1000 randomizations were run to generate a histogram, and then the risk of over-fitting (in %) for individual PLS components was estimated. Figure S-5 shows the comparison of the histogram of noise values and the value under test for 8 PLS components obtained from the CoAdReS selected variables. It can be readily seen that the current randomization test yielded small significance levels for the first 4 PLS components, whereas the last 4 (from the 5^{th} to the 8^{th}) components are clearly insignificant by this test. Table S-2 details the risk of over-fitting (in %) for individual PLS components and we can thus conclude that 4 components should be employed for appropriate PLS modelling. The use of 5 PLS components as suggested by the CV method would lead to over-fitting of the data.

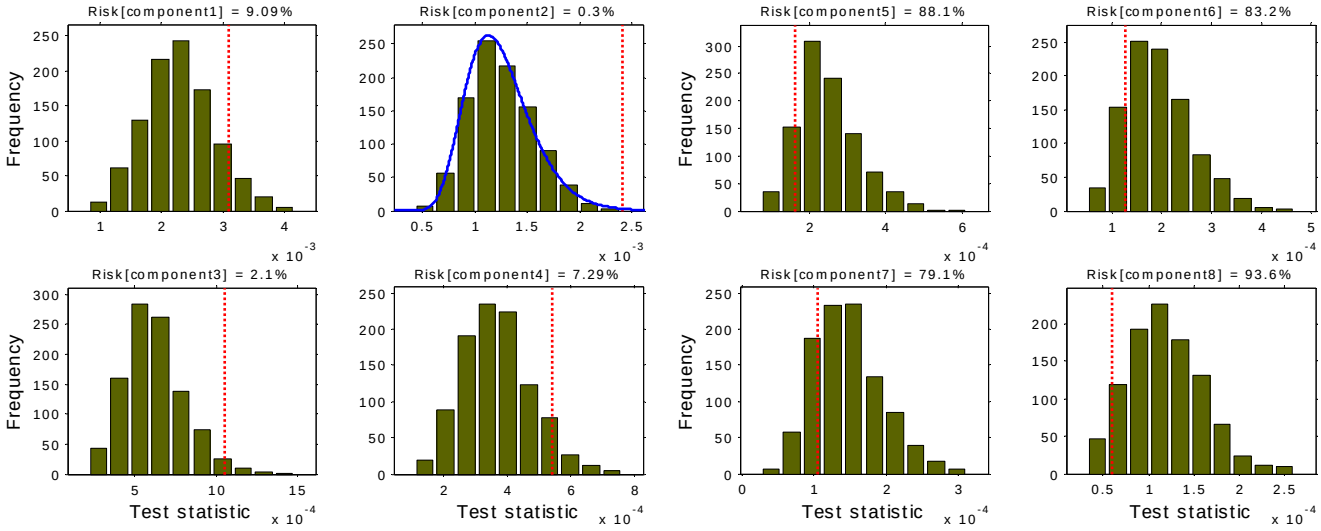

Figure S-5: DS4 sample set with CoAdReS-selected variables: comparison of the histogram of noise values and the value under test (···) for 8 PLS components.

Table S-2: Risk of over-fitting (in %) for individual components, estimated from 1000 randomizations.

DS4 sample set	PLS component							
	1	2	3	4	5	6	7	8
CoAdReS-selected variables	9.09	0.3	2.1	7.29	88.1	83.2	79.1	93.6
ACO-selected variables	34.7	1.1	2.2	11.8	22.7	23.5	62.6	37

This method was also implemented on the ACO-selected variables of the DS4 sample set for the PLS component selection (Figure S-6 and Table S-2). As a consequence, 6 components were suggested for PLS modelling (down from 8 suggested by CV method).

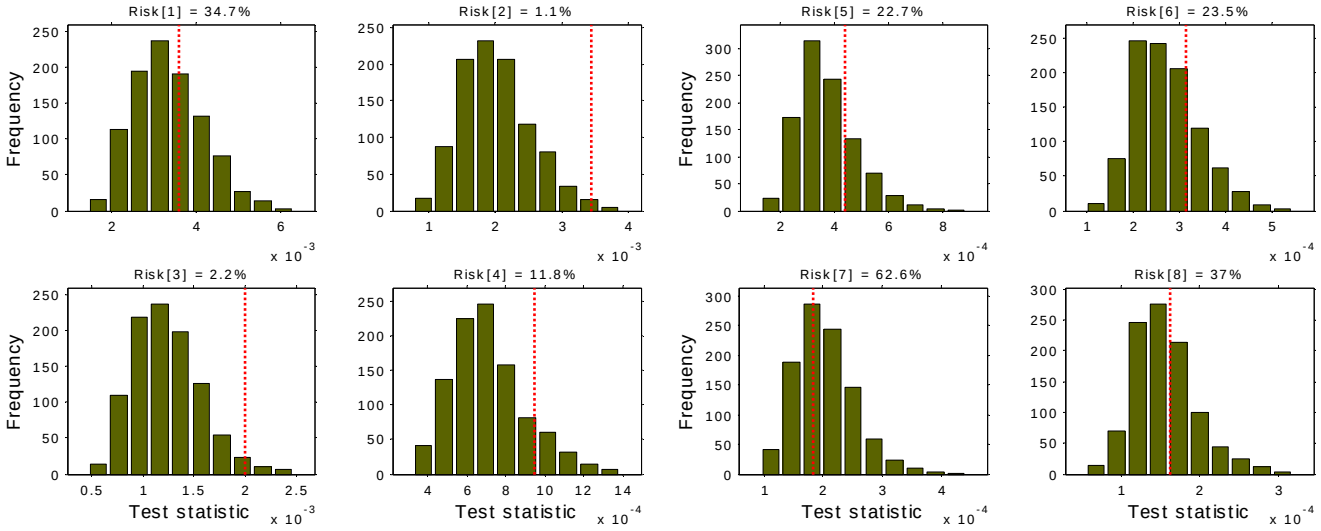

Figure S-6: DS4 sample set with ACO-selected variables: comparison of the histogram of noise values and the value under test (···) for 8 PLS components.

Both the CV and randomization test methods were carried out on all sample sets and Table S-3 shows the results. It is noted that the randomization test method consistently selected fewer PLS components, and thus the probability of data over-fitting was significantly reduced. Therefore, the PLS components as determined by the randomization test method were used for the quantitative PLS models, Table 2 in the manuscript.

Table S-3: Overview of PLS component selection for all the sample sets used in this study.

Dataset		# PLS components	
		CV method	Randomization test
DS4	CoAdReS	**5**	**4**
	ACO	**8**	**6**
DS5	CoAdReS	**8**	**5**
	ACO	**8**	**6**
DS6	CoAdReS	**8**	**6**
	ACO	**8**	**5**
DS7	CoAdReS	**7**	**5**
	ACO	**7**	**5**
DS8	CoAdReS	**8**	**7**
	ACO	**9**	**7**
DS9	CoAdReS	**9**	**8**
	ACO	**9**	**5**

DS10	CoAdReS	**9**	**7**
	ACO	**9**	**7**
DS11	CoAdReS	**8**	**6**
	ACO	**9**	**5**
DS12	CoAdReS	**8**	**7**
	ACO	**8**	**6**

The complexity/simplicity of the PLS model depends on a combination of the data set complexity, the appropriate data pretreatment techniques, and the proper estimate of the optimal number of PLS components. Bearing in mind the fact that:

- The samples analyzed in this study were from an industrial bioprocess, therefore very complex and composed of a large number of constituents,
- The Raman spectra are also very complex, with many overlapping bands, and the Raman signal of water is the strongest component. Therefore we do not explicitly observe signals that we can unambiguously assign to specific components.
- Each calibration set had a very limited number of samples available; therefore it is not easy to generate a multivariate calibration model with a minimal number of components.

To demonstrate the limitations due to sample set size (and corresponding influence of LV number) we selected the DS12 sample set as an example. A simple leave-one-out cross validation (LOOCV) was employed upon both the full spectra and ACO-selected variable set. PLS models were generated with varying sample numbers from 14 to 29 (selected by production lot number). For each model we extracted the RMSEC/RMSEP values for different numbers of latent variables from 1 to 14. All this information was then visualized in a 3D plot (Figure S-7) to show the interdependence of the parameters.

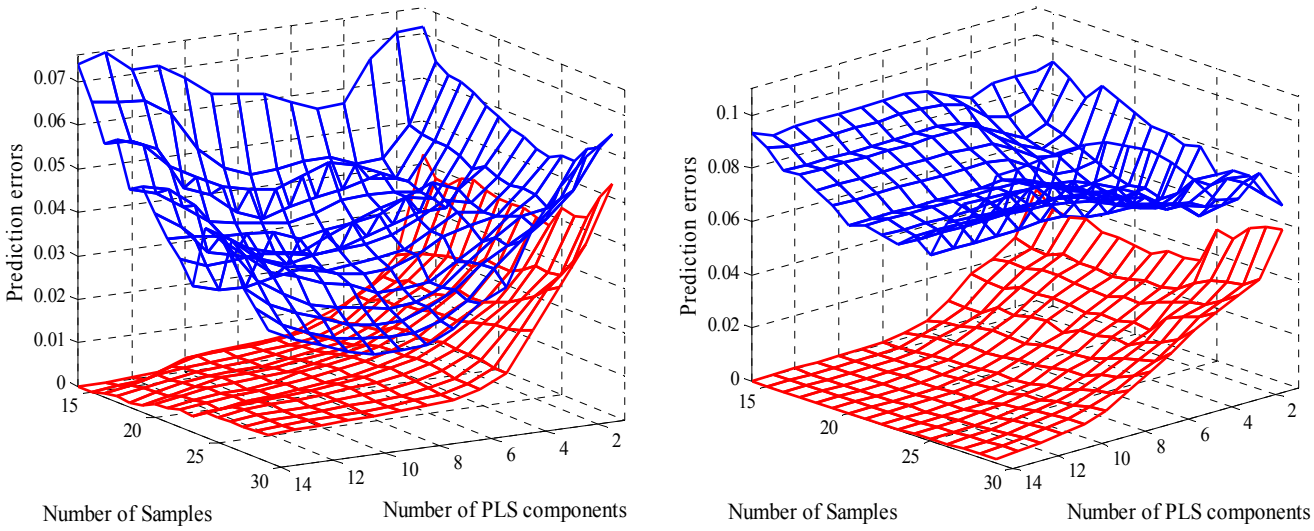

Figure S-7: Prediction errors showing the relation with the varying numbers of samples and PLS components: (left) ACO-selected variable set, and (right) full spectra in the 1850~400 cm^{-1} range. Blue represents the RMSEV values and red denotes the RMSEC values.

The plot clearly shows the dramatic improvement in prediction errors with the use of the reduced variable set (ACO in this instance) compared to the wide spectral range data set. It also shows that for the ACO variable set the RMSEV value tends to a minimum (~0.02 g L^{-1}) value with ~8–10 LVs with ~29 samples. The overall downward trend with increasing sample number is to be expected and does indicate that the correlation with yield is indeed real. RMSEC values tend to converge at LVs=6 once the sample number was ≥20, and the value stays nearly constant with increasing LV number and/or sample number. This would tend to suggest that 6 LVs and an RMSEC of ~0.01 g L^{-1} will be the best theoretical result obtainable (using this type of Raman data). If we were in a position to double or treble the sample number (which unfortunately we are not) then we would fully expect this trend to continue and the RMSEC/RMSEV values to converge.

S4. Quantitative Study of the DS12 Sample Set

Both the CoAdReS and ACO methods were applied to the sample set, DS12, to sort the informative variables. 23 variables were selected by CoAdReS with a threshold histogram value of 0.29, while 24 variables selected by ACO with a histogram threshold of 0.42 (Figure S-8). Most of the selected variables were located in the fingerprint region (1800–900 cm^{-1}).

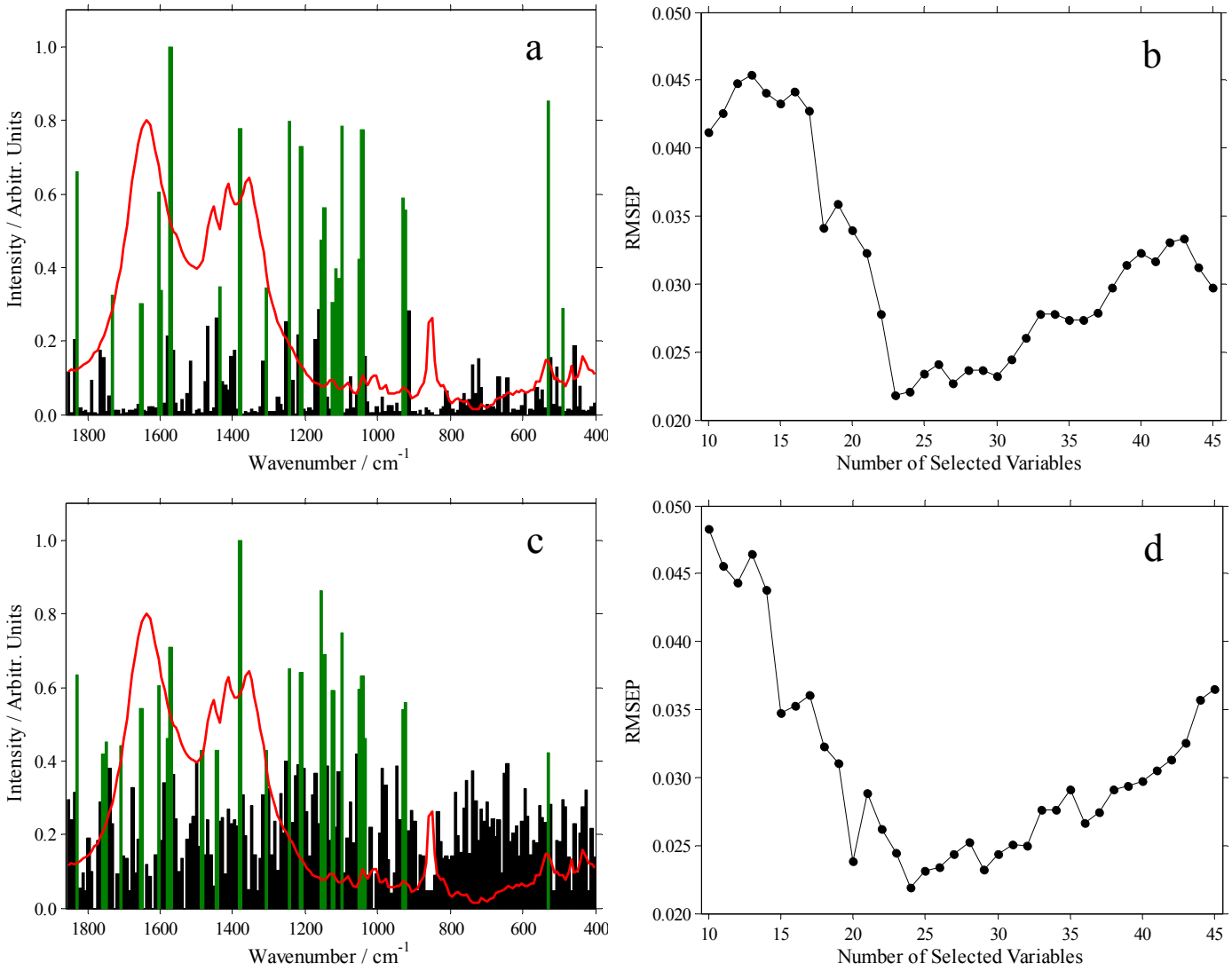

Figure S-8: (**a**) CoAdReS variable selection result for the DS12 (1853–400 cm^{-1} range). The green bars show the histogram values ≥0.29, and the black bars those with values smaller than 0.29. Superimposed is the mean baseline-corrected Raman spectrum in arbitrary vertical scales. (**b**) Determination of number of the selected variables. (**c**) ACO variable selection result for the DS12 in the range of 1853–400 cm^{-1}. The green bars show the histogram values ≥0.42, and the black bars those with values smaller than 0.42. Superimposed is the mean baseline-corrected Raman spectrum in arbitrary vertical scales. (**d**) Determination of number of the selected variables.

These informative variables were then used in PLS regression models (Figure S-9) to predict product yield. These models were resulted from averaging 500 PLS computations using 24 random samples for calibration and 5 samples for Monte Carlo cross-validation in each PLS modelling. The average RMSEC, RMSEP, REP%, and R^2 values were calculated and outlined on the figure. It is pronounced that the model quality in terms of reliability and accuracy was

thus greatly improved, compared to the case where the full 1853–400 cm^{-1} spectral range was used.

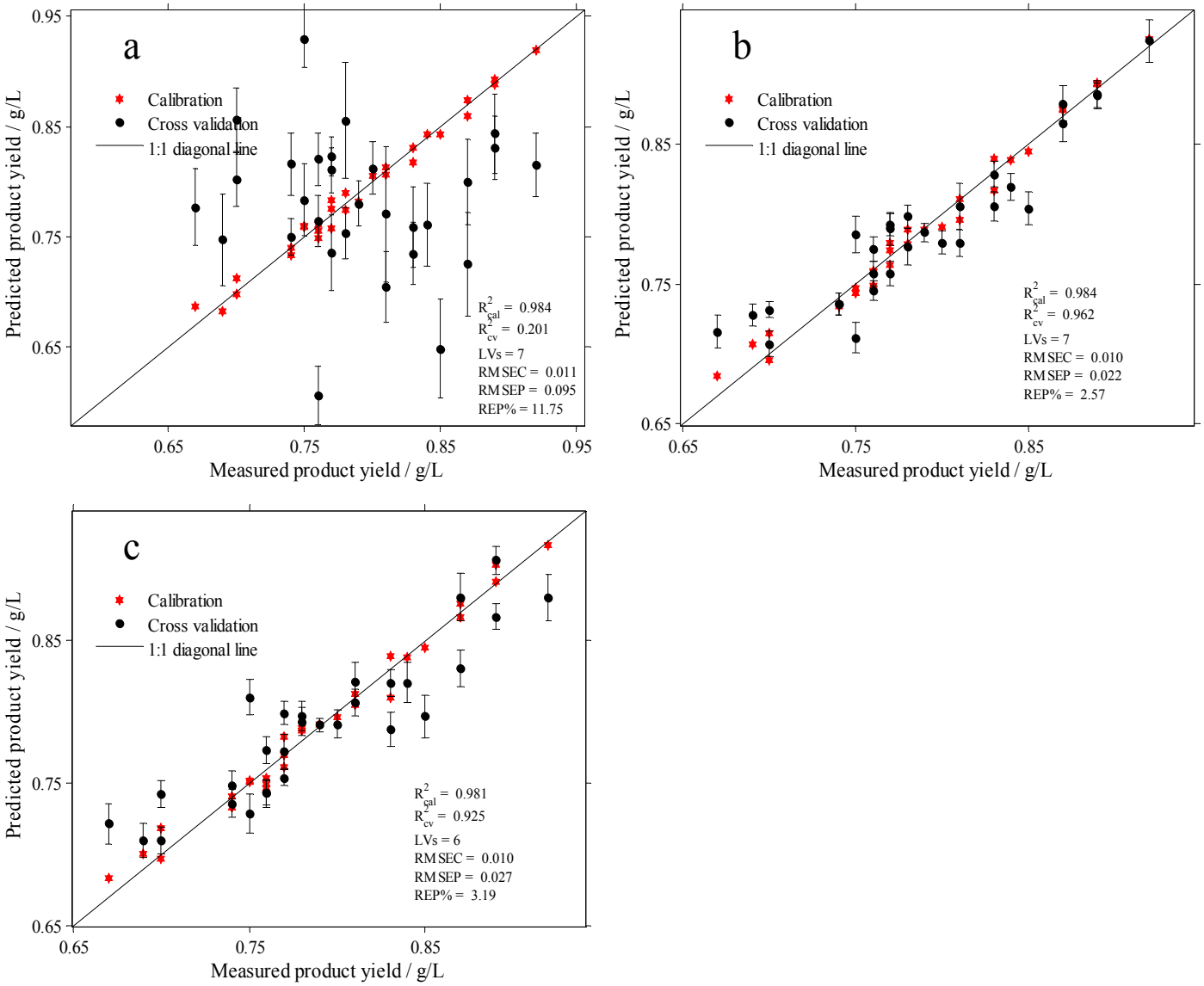

Figure S-9: PLS models for the correlation between the Raman spectral variables of DS12 and product yield (titre in g L^{-1}), which were obtained from averaging 500 PLS computations using 24 random samples for calibration and 5 samples for Monte Carlo cross-validation in each PLS modelling by means of: (**a**) full 1853–400 cm^{-1} spectral range, (**b**) CoAdReS selected variables, and (**c**) ACO selected variables.

S5. PLS MODELLING OF THE DS4 SAMPLE SET

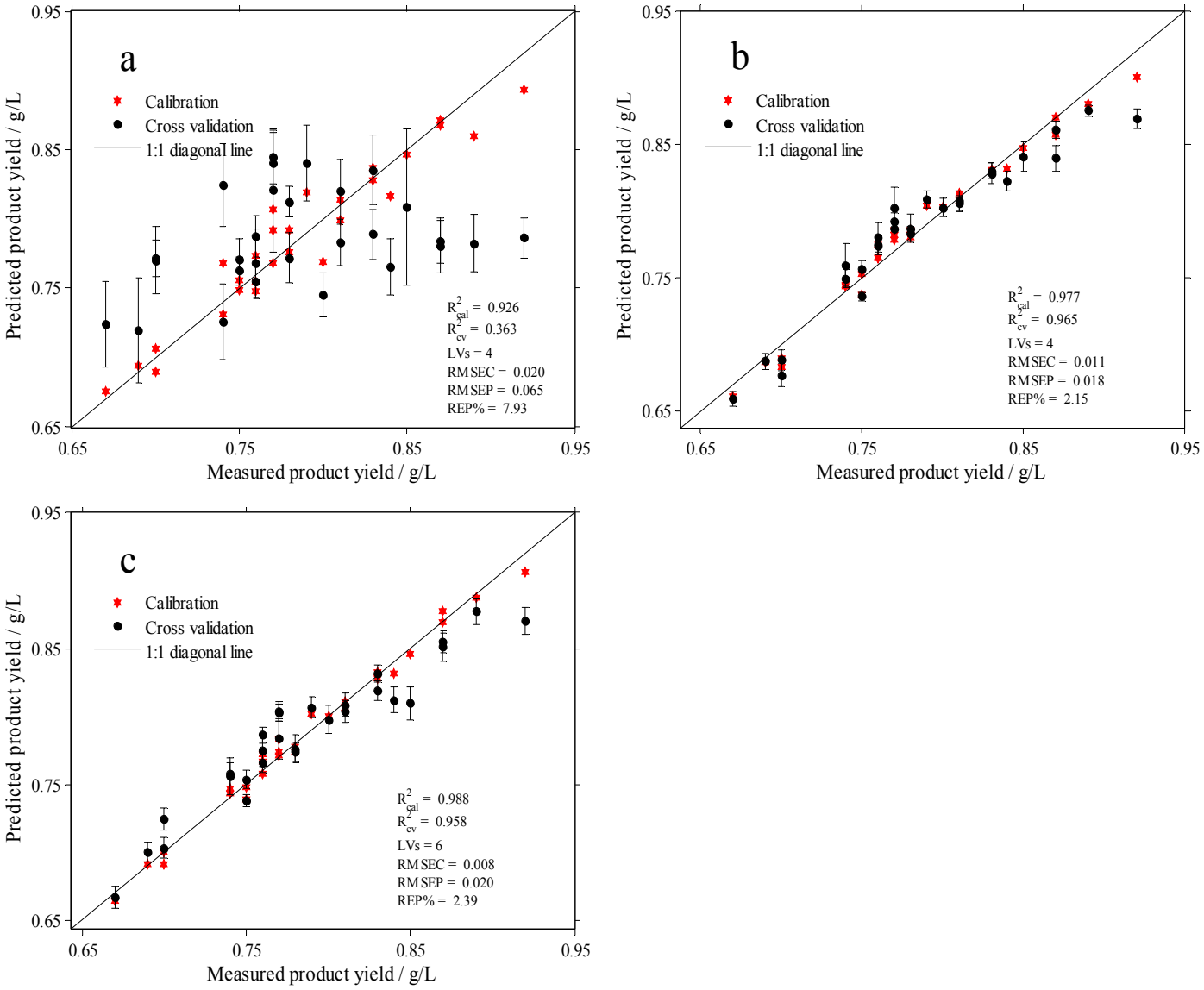

Figure S-10: PLS models for the correlation between the Raman spectral variables of DS4 and product yield (titre in g L^{-1}), which were obtained from averaging 500 PLS computations using 23 random samples for calibration and 5 samples for Monte Carlo cross-validation in each PLS modelling by means of: (**a**) full 1853–400 cm^{-1} spectral range, (**b**) CoAdReS selected variables, and (**c**) ACO selected variables.

Performance Monitoring of a Mammalian Cell Based Bioprocess using Raman spectroscopy. B. Li, B.H. Ray, K.J. Leister, and A.G. Ryder. *Analytica Chimica Acta*, 796, 84-91, (2013). DOI: http://dx.doi.org/10.1016/j.aca.2013.07.058

S6. Principal Component Analysis of Basal Media and Bioprocess Broths from a Single Production Lot

Principal component analysis (PCA) was carried out on the baseline corrected Raman spectra (1853–400 cm^{-1}) of basal media and bioprocess broth samples, which were pulled from a single production lot at twelve set time points in four consecutive bioreactors. A 3-PC model explained 98.75% variance of the data. The result revealed significant changes among these broths, which were linked to the growth phase. The evidently visible changes with process time are the increase in intensity of the bands at 530, 854, 1413, and 1453 cm^{-1} which were associated to the critical components. In particular, the two broths (#11 and #12) sampled towards the end of the fermentation processing especially show differences along PC2, because in the phase of production, more protein product is being produced.

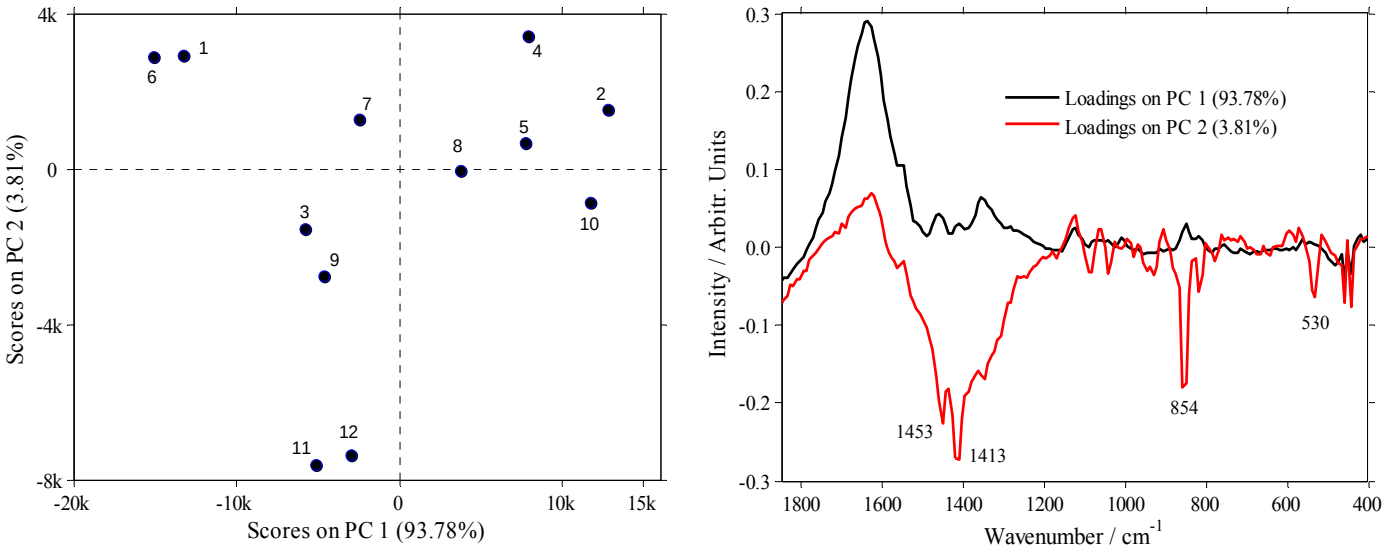

Figure S-11: Scores and loadings plots generated from the PCA of the Raman spectra (1853–400 cm^{-1} range) of samples (DS1 to DS12) from one single production run. Spectra were not normalised.

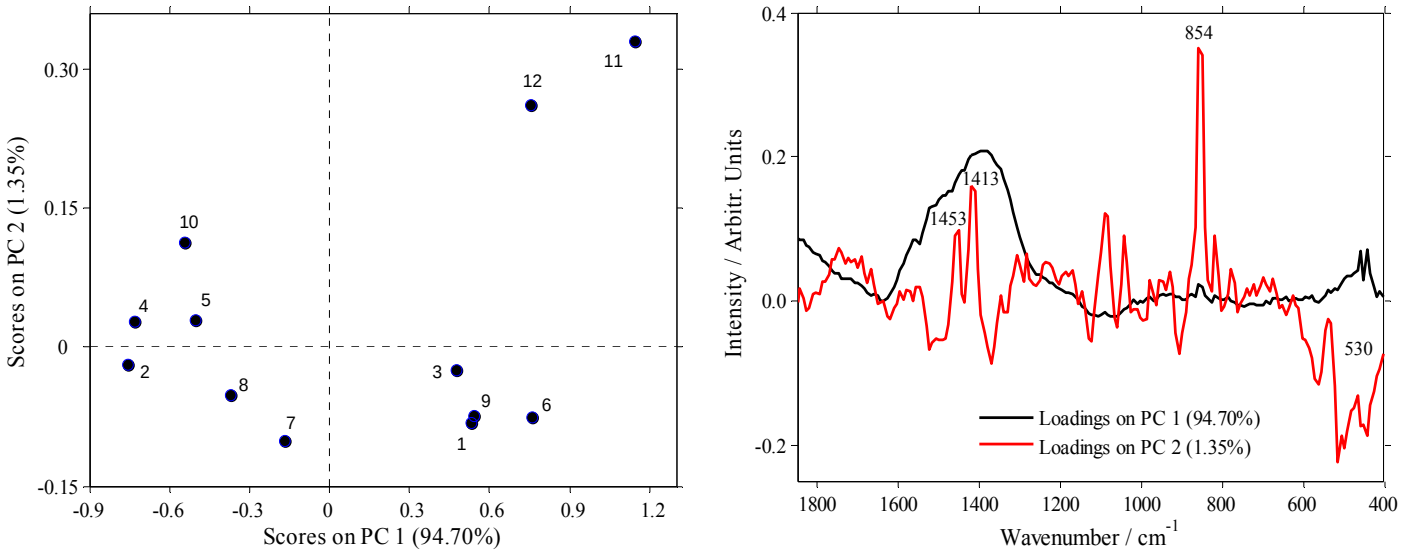

Figure S-12: Scores and loadings plots resulted from the PCA of the Raman spectra (1853–400 cm^{-1} range) of samples (DS1 to DS12) from one single production run, after the spectra being first baseline-corrected and then normalized to the water band at 1636 cm^{-1}.

S7. VARIABLE SELECTION TABLES

Table S-4: Common variables selected by CoAdReS and ACO for the different sample sets. The wavenumber values which are in bold highlight the five variables with the highest histogram scores.

Data set	Common var.	Wavenumbers (cm^{-1})
DS4	14	482, 522, **619**, **979**, **1155**, 1163, 1196, **1204**, 1524, **1572**, 1580, 1620, 1676, 1789
DS6	10	426, 827, **1147**, 1396, **1412**, **1524**, 1556, **1628**, 1668, **1789**
DS9	18	466, **506**, 522, 530, 538, **563**, 579, **587**, **667**, 899, 995, 1003, 1147, 1155, 1188, 1260, **1268**, 1348
DS5	10	506, 563, **619**, 939, **971**, 1035, **1059**, **1075**, 1412, **1764**
DS7	12	**482**, 619, 651, 923, 1027, 1043, **1075**, **1220**, **1308**, **1404**, 1596, 1660
DS8	9	**498**, **771**, **979**, **1083**, 1236, **1340**, 1708, 1805, 1845
DS10	11	466, 506, **619**, **659**, 763, **771**, 779, 883, **923**, **1781**, 1829
DS11	8	**595**, **643**, **1131**, 1556, **1652**, 1668, **1724**, 1740
DS12	17	**530**, 923, 931, **1043**, 1051, **1099**, 1123, 1147, 1155, 1212, 1244, 1308, **1380**, **1572**, 1604, 1652, 1829

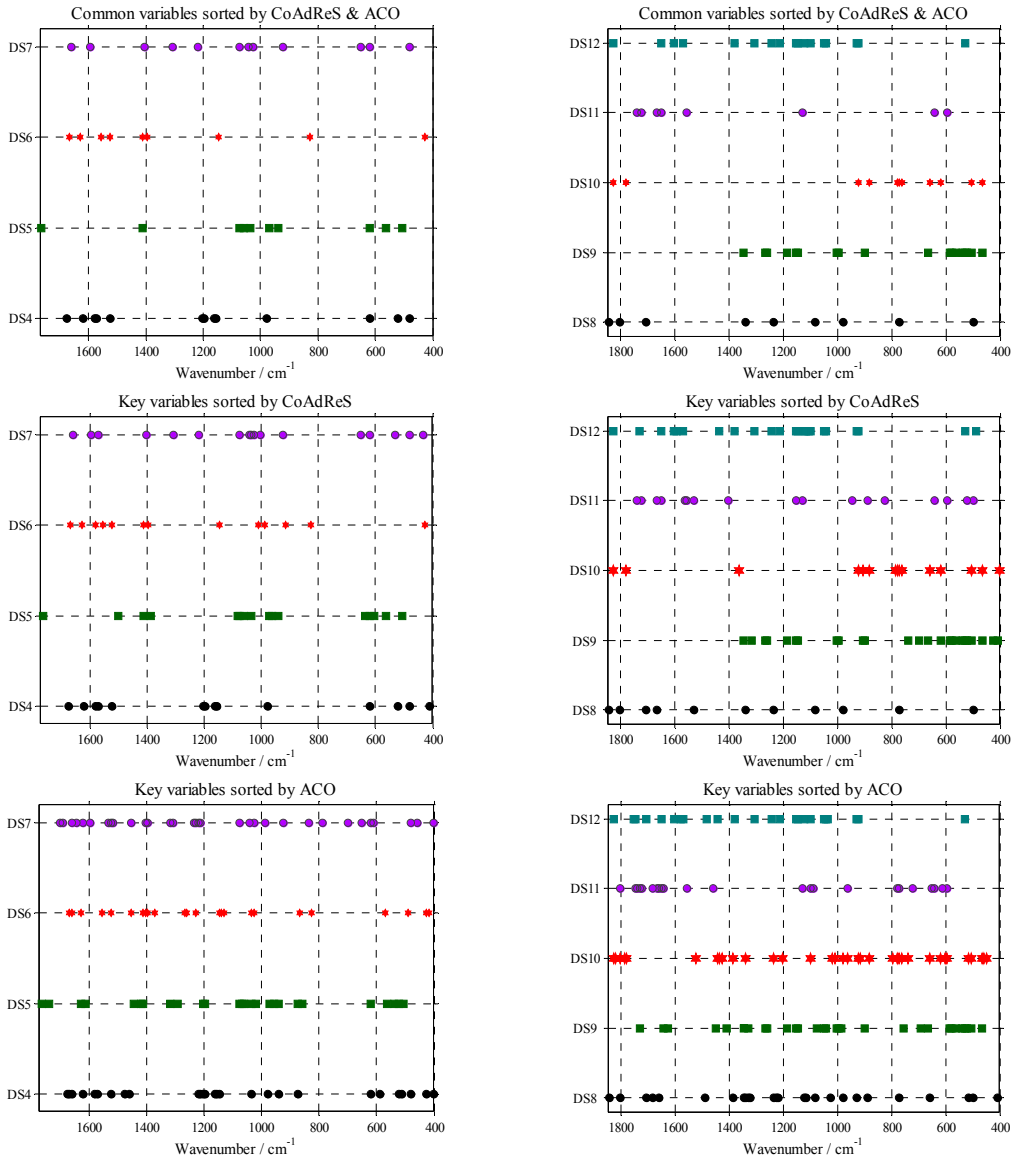

Figure S-13: LEFT, DS4 to DS7 sample sets: (Top) Common variables selected by both CoAdReS and ACO; (Middle) CoAdReS selected variables; (Bottom) ACO selected variables. **RIGHT, DS8 to DS12 sample sets:** (Top) Common variables selected by both CoAdReS and ACO; (Middle) CoAdReS selected variables; (Bottom) ACO selected variables. An arbitrary vertical scale was used throughout.

REFERENCES

[1] A.G. Ryder, J. De Vincentis, B.Y. Li, P.W. Ryan, N.M.S. Sirimuthu, K.J. Leister, *J. Raman Spectrosc.*, 41 (2010) 1266.
[2] H. Martens, T. Naes, *Multivariate Calibration*, New York, 1989.
[3] D.M. Haaland, R.G. Easterling, *Appl. Spectrosc.*, 36 (1982) 665.
[4] A. Lorber, *Anal. Chem.*, 58 (1986) 1167.
[5] A. Savitzky, M.J.E. Golay, *Anal. Chem.*, 36 (1964) 1627.
[6] Q.S. Xu, Y.Z. Liang, *Chemometr. Intell. Lab. Syst.*, 56 (2001) 1.

Performance Monitoring of a Mammalian Cell Based Bioprocess using Raman spectroscopy. B. Li, B.H. Ray, K.J. Leister, and A.G. Ryder. *Analytica Chimica Acta*, 796, 84-91, (2013). DOI: http://dx.doi.org/10.1016/j.aca.2013.07.058

[7] S. Wiklund, D. Nilsson, L. Eriksson, M. Sjostrom, S. Wold, K. Faber, *J. Chemometr.*, 21 (2007) 427.