WHITE PAPER

De novo Approaches for Bioprocess Parameter Estimation

Introduction

Raman spectroscopy has enjoyed substantial growth in pharmaceutical applications over the last 20 years. Polymorph analysis was an early and unique capability that Raman offered in pharmaceutical analysis in analytical labs[1], along with superb spectral-microscopy capability for particle, substrate, and surface analysis[2]. Handheld Raman systems in pharma proliferated starting in the late 2010s. These systems were configured with purpose-build operating systems for excipient and API pharmacopeial identity testing in GMP environments,[3,4] solid dosage form authentication & anti-counterfeit analysis[5] and are now the *de facto* standard for high-efficiency GMP raw material identity testing.

Bioprocess monitoring was a very early area of interest for spectroscopic platforms. Near-IR and mid-IR systems had been investigated for bioprocess metabolite monitoring applications as early as the late 1990s, (*e.g.* references 6-8) but the profound absorption of water in the IR region severely limits the pathlengths usable for absorption measurements without excessive detector noise. Raman spectroscopy benefits from a comparatively weak water scattering cross-section, and so it was unsurprising that Raman began to be investigated for this application in the very early 2000s as well.[9-11] Raman technology also offers considerable flexibility in terms of optical sampling geometries given the minimal interference of plastics, glasses, and minerals as sampling interfaces.

The key focus areas for this early Raman bioprocess work were cellular metabolites in a variety of biological systems, and this application has continued with rapidly expanding interest. Authors have also published on the possibility of assessing product quality attributes, such as protein post-translational modifications,[12-13] and aggregation[14] among others. Citations related to "*Raman + bioprocess*" have exponentially risen according to Google Scholar over the last 10 years (Figure 1) and appear to be poised to surpass 4000 citations in 2023.

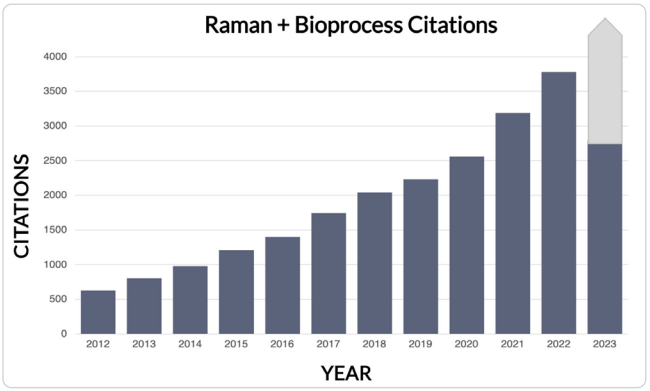

Figure 1. Google Scholar-reported citations to journal articles, patents & proceedings including keywords Raman + bioprocess.

Challenge:

Empirical Calibration Of Bioprocesses

The analysis of Raman data in complex biosystems requires computational assistance. A wide variety of chemometric/multivariate tools can be employed in this endeavor as reviewed by Ryder.[15] With respect to modeling of critical process parameters and critical quality attributes (CPPs and CQAs), overwhelmingly partial least-squares (PLS) regression is noted in the literature. PLS is one of a very large class of latent variable / regularized empirical linear calibration methods. The reasons for its apparent domination on chemical applications are largely historical and commercial, as it has no more favorable performance

than other methods.[16,17] But all empirical methods do enjoy the advantage that little to no detailed understanding is necessary of the underlying cell culture environment, chemistry and physics of the measurement instrument.

But there are several significant challenges in modeling bioprocess data using these empirical calibration approaches, as enumerated below.

1. Nonstationarity & homoscedasticity: In mathematics and statistics "stationarity" is a term implying that each datum (spectrum in this case) is drawn from a random distribution with fixed distributional properties. The empirical methods such as PLS in most commercial software are only theoretically accurate and optimized with 'stationary' data. This would imply that every bioprocess must run the same, with consistent correlations among chemical species. It also implies that the measurement variance in the instrument is always the same in time and channels (homoscedasticity). This is never true with Raman (or NIR or MIR absorption), and particularly not in bioprocesses when significant biomass can contribute to very large fluorescence differences between/within bioprocess runs, and therefore orders of magnitude difference in shot noise.

2. Extreme covariates: Almost by definition there are extreme time-domain correlations between many species over the course of a bioprocess. The empirical methods that are widely used are designed to leverage those empirical time-domain correlations, but these are extremely prone to non-specific associations that have limited predictive accuracy or generalization.[17-18]

3. Exchangeability & cross-validation: Related to the above, cross-validation is often done as a quasi-validation evaluation of an empirical model in data modeling efforts. For cross-validation results to be valid and representative, the data must be 'exchangeable',[19] but for the reasons noted with respect to extreme covariates, this condition is typically grossly violated with bioprocess data.

4. Trial-and-Error: Most of these empirical methods include a bevy of options for variable selection, preprocessing treatments, normalization, and correction methods. The recommended approach is 'try it and see what seems to work', as there is often little theoretical justification to guide the choice of one approach over another.

5. Figures of Merit: Related to the above, the primary metric reported in most commercial software is "RMSEC/RMSECV/RMSEP": root-mean-squared-error-of-[calibration/cross-validation/prediction]. Compendial analytical standards usually expect estimates of selectivity, linearity, precision, limit of detection, and sensitivity, but unfortunately, empirical modeling approaches don't provide direct estimates of the central figures of merit.[20] Users can do experimental work to evaluate these, but it is quite challenging and typically requires custom programming/analysis.

6. Spectrometer variation: When empirical methods are developed, their covariance also captures properties and non-idealities of the individual spectrometer.[21] When spectrometers are exchanged, or sources/detectors replaced, frequently the multivariate model needs to be corrected for relevance on the new spectrometer properties. A broad range of mathematical methods are used to perform this 'calibration transfer'.

7. Regulatory overhang: The black-box nature of empirical calibration methods requires extensive empirical validation efforts to demonstrate sensitivity, selectivity, linearity, and robustness. A few general guidelines have been offered in regulator documents (*e.g.*, ICH Q14 10.3), but they are not particularly clear-cut or grounded in the mathematical basis of these methods.

Given these challenges, it is little wonder that robust Raman method development and deployment has been a particularly vexing challenge in bioprocess applications.

There have been numerous efforts to overcome several of these impediments. Intentionally perturbed and designed experiments (*e.g.* 22, and references there) can be used to try to 'break' the extreme covariates that are intrinsically present and expand the range

of the empirical data available for modeling. Several groups have reported success building 'generic' models using PLS with various pre-treatment methods and have reported reasonable success for defined platform methods,[23-26] but often upwards of 25-30 process runs were involved in these efforts at very considerable expense—several years of process time—and that excludes deployment and maintenance activities that follow. These literature results are consistent with reports from groups in industry reporting at technical conferences.

It was our aim to ameliorate challenges with Raman implementation for bioprocess monitoring, initially for mammalian processes employing CHO and HEK293 cell lines which are widely used for protein/monoclonal antibodies and viral vector manufacturing, but with a scalable framework for future development opportunities.

A *de novo* Model

It is difficult to circumvent many of the afore-mentioned challenges with purely empirical modeling/calibration. Hybrid models are of increasing interest in the biology and bioprocess domains.[27] To date, these approaches have largely combined knowledge of fundamental biological mechanisms, chemical engineering knowledge, computational fluid dynamics, and other elements, along with some empirically measured or observed data for increased process understanding. The more mechanistic elements of the model constrain the empirical optimization in such a way as to reduce the risk of overfitting / local minima and guide the overall model to an interpretable and robust approximation. The use of first principles or building-block information to predict complex outcomes is sometimes referred to as *de novo* methods, such as *de novo* protein structure modeling,[28] and that is the terminology we have adopted to describe MAVERICK's computational framework.

MAVERICK's *de novo* model is derived from work dating as far back as the 70s on explicit probabilistic frameworks for multivariate calibration (MVC) such as the early work of Morgan and others.[29-31] It is contrasted with the usual empirical multivariate calibration construct in Figure 2.

The empirical MVC approach estimates a predictor $\hat{\mathbf{b}}$ from an approximation of observed spectral data, $\mathbf{X}$ ($\tilde{\mathbf{X}}$), and paired reference data ($\mathbf{y}$), in the presence of some reference error, $\mathbf{e}$. The calculation of $\hat{\mathbf{b}}$ itself is elementary; the challenges 1-7 noted above largely manifest in the approximation of '$\mathbf{X}$' in each domain—what experiments should be done, on what hardware, across which conditions, how should the raw data be manipulated/processed prior to calculating $\hat{\mathbf{b}}$, and how does the resultant model perform in truly prospective conditions.

The approximation of $\mathbf{X}$ is essential to control the risk of overfitting with empirical methods, and there are many, many, many different possible 'approximators' of $\mathbf{X}$ ($\tilde{\mathbf{X}}$) that may be useful in practice. PLS (partial least-squares)

Traditional Empirical MVC

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \qquad \mathbf{X} \in \mathbb{R}^{nxm}, \mathbf{y} \in \mathbb{R}^{n}$$
$$\hat{\mathbf{b}} = \tilde{\mathbf{X}}^{+}\mathbf{y}$$

where $\tilde{\mathbf{X}}$ is a reduced approximation of $\mathbf{X}$, *e.g.* via subspace methods such as PLS, PCA, and '+' denotes a pseudoinverse.

$$\hat{y} = \tilde{\mathbf{x}}_{pred}\hat{\mathbf{b}}$$

MAVERICK *de novo* Model

$$\hat{\mathbf{B}}_t = f(\mathbf{K}, \boldsymbol{\Psi}, \mathbf{F}_t, \hat{\boldsymbol{\Sigma}}_t, \hat{\mathbf{E}}_t, \boldsymbol{\Lambda})$$
$$\hat{y} = \mathbf{x}_{pred}\hat{\mathbf{B}}_t$$

$\mathbf{K}$ is a basis set spanning the reference spectra in the formulary
$\boldsymbol{\Psi}$ is a multivariate prior represented expected variance-covariance of the formulary concentrations
$\mathbf{F}$ is an optimal filter function, specific to each MAVERICK, and its measurement error & environmental state at time t
$\boldsymbol{\Sigma}$ is the estimated error covariance at time t
$\mathbf{E}$ is an environmental function at time t
$\boldsymbol{\Lambda}$ is an autoregressive-with-innovation function tracking time-domain behavior of the estimator

Figure 2. High level comparison of traditional multivariate calibration approach versus MAVERICK's 'de novo' model.

is one of many subspace projection methods, and it is particularly widespread owing to its early inclusion in several commercial software packages. It is also common to also eliminate wavelength ranges or apply other linear or nonlinear transformations in the creation of $\tilde{\mathbf{X}}$. The myriad options available for this 'approximation' step of modeling are a significant secondary source of overfitting as sometimes hundreds or thousands of options are evaluated, ravenously consuming generalized degrees of freedom.[32]

In contrast, MAVERICK's *de novo* model doesn't use any empirically observed $\mathbf{X}$ or $\mathbf{y}$ data. Instead, it uses the terms noted in Figure 2—some static and some dynamic—to create a 'best linear predictor' at time $\mathbf{t}$ for the system under *active* measurement. While the heart of this model is probabilistic, several of its critical terms are directly derivable from first principles with sufficient knowledge of the optical train, electronics, and multivariate statistics. Since these effects are dynamic in a Raman system observing a bioprocess, several of the model terms are (unsurprisingly) dynamic.

The formulary elements $\mathbf{K}$,$\mathbf{\Psi}$ represent a 'master list' of *possible* chemical/biochemical contributors to the observable Raman spectrum and an associated prior probability density function, from which concentration estimates are produced *a posteriori*. One might wonder how it is possible to cover all possibilities in the formulary, but there are a few helpful boundaries. While it is very likely that the number of chemical/biochemical species in an active bioprocess may number in the thousands,[33] the limited sensitivity of Raman spectroscopy implies that one really needs only to consider the major components above approximately 0.01 g/L. At these limits in mammalian cultures, we have found that a few hundred intrinsic species are relevant, along with a cross-section of non-biologic additives (*e.g.* surfactants, anti-foaming agents). *Deconvolving* an observed Raman spectrum with that many degrees of freedom is generally an ill-posed problem, but using the *de novo* framework the solution is sufficiently self-conditioned to produce low-variance estimates of concentration.

The remaining terms are both device and time dependent. $\mathbf{F}$ is an optimal filter function derived from a multidimensional factory characterization of each MAVERICK system and is adapted in real-time for changing sample and system conditions. Many of the largest sources of error in Raman systems are fundamental to the system's optical design and electronics. MAVERICK's internal system model allows it to estimate $\mathbf{\Sigma}_t$ the measurement error covariance in real-time. Related, the system model also allows for $\mathbf{E}_t$ to adapt to, for example, changing lighting. temperature, and turbidity conditions. Finally, since in a bioprocess the system state at time t is related to the state at time $\mathbf{t\text{-}1}$, environmental and autoregressive components ($\mathbf{\Lambda}$) are included in the model for inertia.

Figures of Merit

Several properties of this estimator have been previously discussed, such as closed-form expressions for mean-squared error of prediction (MSEP).[34]

$$\widehat{MSEP} = \mathbf{W}^T\mathbf{\Psi}\mathbf{W} + \hat{\mathbf{B}}^T\mathbf{K}^T\mathbf{\Sigma}\mathbf{F}\hat{\mathbf{B}}$$

where

$$\mathbf{W} = \left(\mathbf{KF}\hat{\mathbf{B}} - \mathbf{I}_n\right)$$

(The index 't' for $\mathbf{B}$ and $\mathbf{\Sigma}$ has been dropped for simplicity.)

As noted above, one consistent challenge in empirical model development is the opacity of the resulting model properties. It is quite rare to find publications demonstrating bioprocess Raman application citing standard analytical figures-of-merit—sensitivity, selectivity, LOD for example—for the resulting models, because literature definitions are complicated for multivariate models. Sensitivity and selectivity factors consistent with the IUPAC definitions can be directly estimated from the *de novo* model following the processes noted in 34. Lastly, other model diagnostics can also be inferred, such as in-plane and out-of-plane conformity, analogous to Hoteling or leverage statistics and F-ratios[35]:

$$d_{\parallel} = \mathbf{x}_{pred}\mathbf{F}(\mathbf{\Sigma}^{-1}\mathbf{K}^{\mathbf{T}}\mathbf{F}^{\mathbf{T}}(\mathbf{KF}\mathbf{\Sigma}^{-1}\mathbf{K}^{\mathbf{T}}\mathbf{F}^{\mathbf{T}})^{-1}\mathbf{KF}) + \mathbf{f}(\mathbf{E})$$
$$d_{\perp} = \boldsymbol{x}_{pred}\mathbf{F}(\mathbf{I}_{\mathrm{n}} - \mathbf{\Sigma}^{-1}\mathbf{K}^{\mathbf{T}}\mathbf{F}^{\mathbf{T}}(\mathbf{KF}\mathbf{\Sigma}^{-1}\mathbf{K}^{\mathbf{T}}\mathbf{F}^{\mathbf{T}})^{-1}\mathbf{KF}) + \mathbf{f}(\mathbf{E})$$

Quick Calibration

The *de novo* approach of the MAVERICK system relieves a substantial burden from the end user but doesn't make it *completely* free of all forms of 'calibration'. Since MAVERICK systems are designed to plug and play across measurement modules, probe adapters and probes, there is one preparatory step that is required to confirm quantitative system suitability before bioprocess analysis can begin. This is a 3-step process, guided on screen by the MAVERICK's software:

1. Immerse probe into "LOW" check solution, press go (wait approximately 4 minutes)
2. Immerse probe into "HIGH" check solution, press go (wait approximately 4 minutes)
3. Autoclave the probe for immersion in the actual bioprocess.

Steps 1+2 check that several properties of the MAVERICK + probe are conforming with the *de novo* model, and a minor scalar correction is made to the *de novo* model outputs for the particular combination of MAVERICK measurement model, probe adapter and probe. This information also allows for automated & audited performance qualification and tracking with the serialized/microchipped probe. MAVERICK also supports a single point 'live' reference which can be helpful eliminate any small observed biases that may be consistent between a particular offline reference analyzer and MAVERICK's *de novo* outputs.

Illustratory Data

Figure 3 illustrates example running performance of MAVERICK on CHO and HEK293 process using the turn-key *de novo* model, compared to some common off-line reference analyzers (enzymatic).

Figure 4 illustrates some of the behind-the-scenes diagnostic information that is afforded by the *de novo* model. This information was extracted from a CHO process running in a laboratory with large windows to the outdoors. In the upper figure, the small undulations observable in the estimated RMSE (g/L) are precisely as expected—the *de novo* model is tracking fundamental

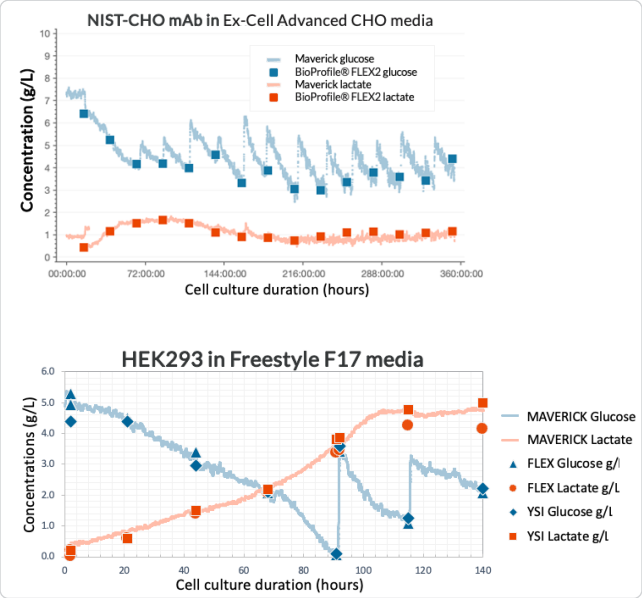

Figure 3. Example data from MAVERICK's de novo Model running on various CHO (top) & HEK (bottom) processes.

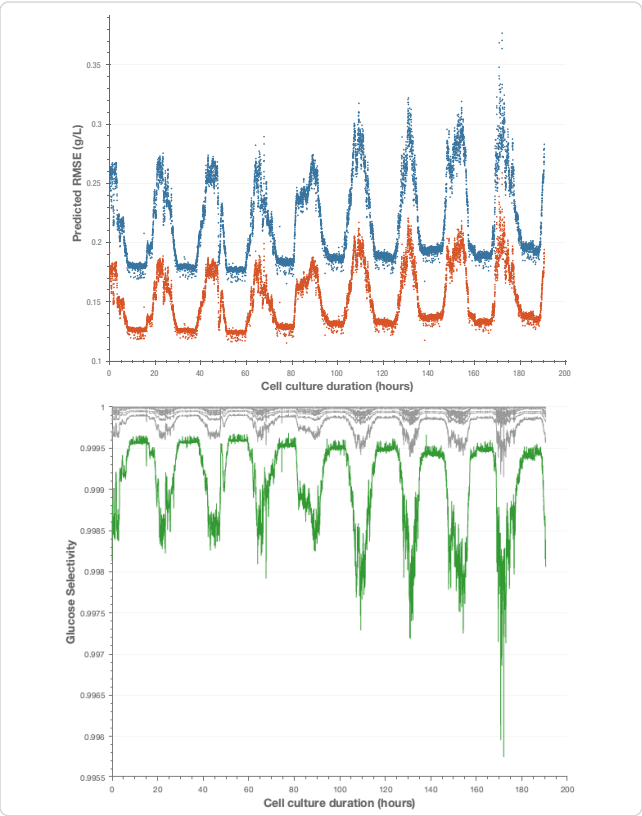

Figure 4. Example of behind-the-scenes figure of merit information from the 'de novo' model for forecasted RMSEP (gL-1) for glucose, lactate, and model selectivity (gL^{-1}/gL^{-1}) for lactate. Selectivity values close to one indicate excellent selectivity against the specified interferent.

shot noise changes throughout the day/night cycle, affecting $\hat{\Sigma}_t$. The same effect is propagating through to selectivity for glucose in the lower figure, which plots glucose selectivity against the top 20 other cell culture media components: as ambient lighting increases the *de novo* model adjusts and regularizes to preserve selectivity despite changes in ambient light. Glutathione is highlighted in green. While it happens to be the 'least' glucose-selective species in this bioprocess run, as the y-axis indicates, glucose selectivity is still excellent (>0.99).

The proliferation of cellular/proteinaceous material in the later stages of the bioprocess process can induce moderate to severe autofluorescence, which is well known to cause substantial difficulties with empirical calibration models. The *de novo* model's figures of merit reflect this effect, observable as a slow general trend upward of the RMSE's, but since the *de novo* model is tracking and compensating for the increase in shot noise from fluorescence in the measurement error model, this effect is handled quite seamlessly.

Boundaries and Opportunities

The key advantages of the *de novo* approach—namely transparency and the avoidance of the pitfalls of empirically derived models—can also be considered its key limitation. As noted above, if optically active components of the bioprocess are not represented in the formulary, measurement error model, or environmental model, the results reported by the *de novo* model are apt to be biased. The degree to which they are biased depends heavily on how optically active they are: trace metals in the low microgram/liter level will have no impact because a) they are optically inactive and b) the concentrations are far too low to be observed with Raman in solution. In general, only covalently bonded chemical species in the 0.01 g/L range and above are considered relevant.

The *de novo* construct is also unable to support so-called 'soft sensors'—virtual parameters that may be *inferred* from empirically observed data, even if there are no direct spectral effects (*e.g.* pH). Without an aetiologic spectral effect for formulary inclusion, the *de novo* model cannot be applied. For those interested in soft-sensor modeling or extended prediction models, customers may choose to take advantage of MAVERICK's full spectral exports, which can be accessed in real-time via OPCUA, or as a consolidated data file at the conclusion of a measurement session.

There are further opportunities to exploit hybrid modeling approaches for $\boldsymbol{\Psi}$ and $\mathbf{K}$. At present a single $\boldsymbol{\Psi}$ seems to be adequate for mammalian bioprocesses, but we are exploring an adaptive $\boldsymbol{\Psi}$ for even more diverse media systems (*e.g.* non-CHO or HEK293 mammalian, avian, insect *etc.*), or alternatively, dynamic constraints on $\mathbf{K}$ if it is apparent from the data that particular formulary components are absent, *e.g.*, via an L1-type regularization.[36] It doesn't escape our notice that dynamical systems models such as so-called digital twins may also directly interface with the *de novo* model for continuous time-domain updates.

Summary

There are exciting opportunities to continue to expand the reporting capabilities of MAVERICK as we validate performance across other analytes and other cell/media processes. It also appears that the de novo model's flexibility should improve robustness across scales/geometries as processes transition from early phase process development to pilot-scale and production. For numerous examples of MAVERICK's current validated performance in a variety of culture systems, please refer to the application notes on the 908 Devices' website at www.908devices.com

References

1. Z. Sun *et al.*, Review of the Application of Raman Spectroscopy in Qualitative and Quantitative Analysis of Drug Polymorphism, *Current Topics in Medicinal Chemistry* 2023, 23:1340-1351
2. M.E. Andersen, R.Z. Muggli, Microscopical techniques in the use of the molecular optics laser examiner Raman microprobe, *Analytical Chemistry* 1981, 53(12):1772–1777
3. R.L.Green, C.D. Brown, Raw-material authentication using a handheld Raman spectrometer, *Pharmaceutical Technology* 2008, 32(3) https://www.pharmtech.com/view/raw-material-authentication-using-handheld-Raman-spectrometer
4. W. Jalenak, R.C. Brush, R.L. Green, C.D. Brown, Verification methods for 198 common raw materials using a handheld Raman spectrometer, *Pharmaceutical Technology* 2009, 33(10) https://www.pharmtech.com/view/verification-methods-198-common-raw-materials-using-handheld-Raman-spectrometer
5. C. Ricci, L. Nyadong, F. Yang, F.M. Fernandez, C.D. Brown, P.N. Newton, S.G. Kazarian, Assessment of hand-held Raman instrumentation for in-situ screening for potentially counterfeit artesunate antimalarial tablets

by FT-Raman spectroscopy and direct ionization mass spectrometry, *Analytica Chimica Acta* 2008, 623(2):178-186

6. A. Hashimoto, A. Yamanaka, M. Kanou, K. Nakanishi, T. Kameoka, Simple and rapid determination of metabolite content in plant cell culture medium using an FT-IR/ATR method. Bioprocess & Biosystems Engineering 2005, 27:115–123
7. M.R. Riley, M. Rhiel , X.J. Zhou, M.A. Arnold, D.W. Murhammer, Simultaneous measurement of glucose and glutamine in insect cell culture media by near infrared spectroscopy. *Biotechnology and Bioengineering* 1997, 55: 11–15
8. M. Rhiel, P. Ducommun, I. Bolzonella, I. Marison, U. von Stockar, Real-time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements. *Biotechnology and Bioengineering* 2002, 77:174–185
9. A. McGovern, D. Broadhurst, J. Taylor, N. Kaderbhai, Winson M, Small D, Rowland J, Kell D, Goodacre R. Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: Application to gibberellic acid production. *Biotechnol Bioeng* 2002, 78:527–538
10. Cannizzaro C, Rhiel M, Marison I, von Stockar U. On-line monitoring of Phaffia rhodozyma fed-batch process with in situ dispersive Raman spectroscopy. *Biotechnology & Bioengineering* 2003, 83:668–680
11. Boyan Li, Paul W. Ryan, Bryan H. Ray, Kirk J. Leister, Narayana M.S. Sirimuthu, Alan G. Ryder Rapid Characterization and Quality Control of Complex Cell Culture Media Solutions Using Raman Spectroscopy and Chemometrics. *Biotechnology and Bioengineering* 2010, (107):290-301
12. McAvan BS, Bowsher LA, Powell T, O'Hara JF, Spitali M, Goodacre R, Doig AJ. Raman spectroscopy to monitor post-translational modifications and degradation in monoclonal antibody therapeutics. *Analytical Chemistry* 2020, 92(15):10381–89.
13. Wei B, Woon N, Dai L, Fish R, Tai M, Handagama W, Yin A, Sun J, Maier A, McDaniel D, et al. Multi-attribute Raman spectroscopy (MARS) for monitoring product quality attributes in formulated monoclonal antibody therapeutics. MAbs 2021, 14(1):e2007564
14. Zhang C, Springall JS, Wang X, Barman I. Rapid, quantitative determination of aggregation and particle formation for antibody drug conjugate therapeutics with label-free Raman spectroscopy. *Anal Chim Acta* 2019, 1081:138–45.
15. Kevin Buckley and Alan G. Ryder, Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing: A Short Review. *Applied Spectroscopy* 2017, Vol. 71(6): 1085–1116
16. P.D. Wentzell, L.V. Montoto, Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures, Chemometrics and Intelligent Laboratory Systems 2003, (65):257-279
17. C.D. Brown, R.L. Green, 2009 Critical factors limiting the interpretation of regression vectors in multivariate calibration, *Trends in Analytical Chemistry*, Vol. 28, No. 4: 506-514
18. C.E. Eskildsen, S.B. Engelsen, K.R. Dankel, L.E. Solberg, T.Naes, Diagnosing indirect relationships in multivariate calibration models, *Journal of Chemometrics* 2021, 35:e3366
19. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2nd Ed.; Springer 2009
20. A.C. Olivieri, Analytical figures of merit: from univariate to multiway calibration, *Chemical Reviews* 2014, 114(10): 5358–5378
21. Jerome J. Workman, A Review of Calibration Transfer Practices and Instrument Differences in Spectroscopy, *Applied Spectroscopy* 2017, 72(3): 340-365
22. Jiarui Wang, Jingyi Chen, Joey Studts & Gang Wang In-line product quality monitoring during biopharmaceutical manufacturing using computational Raman spectroscopy, mAbs 2023, 15:1, 2220149,
23. Yousefi-Darani, A.; Paquet-Durand, O.; von Wrochem, A.; Classen, J.; Tränkle, J.; Mertens, M.; Snelders, J.; Chotteau, V.; Mäkinen, M.; Handl, A.; et al. Generic Chemometric Models for Metabolite Concentration Prediction Based on Raman Spectra. *Sensors* 2022, 22, 5581.
24. Tulsyan A, Wang T, Schorner G, Khodabandehlou H, Coufal M, Undey C. Automatic real-time calibration, assessment, and maintenance of generic Raman models for online monitoring of cell culture processes. Biotechnol Bioeng. 2020, 117(2):406–16.
25. Andre S, Lagresle S, Da Sliva A, Heimendinger P, Hannas Z, Calvosa E, Duponchel L. Developing global regression models for metabolite concentration prediction regardless of cell line. *Biotechnology & Bioengineering* 2017, 114:2550–2559
26. Thaddaeus A. Webster, Brian C. Hadley, William Hilliard, Colin Jaques, Carrie Mason, Development of generic Raman models for a GS-KO ™ CHO platform process, *Biotechnology Progress* 2018, 34(3): 730-737
27. A. Tsopanaglou, I. Jimenez del Val, Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses, *Current Opinion in Chemical Engineering* 2021, 32:100691
28. https://en.wikipedia.org/wiki/De_novo_protein_structure_prediction
29. D. R. Morgan, Spectral absorption pattern detection and estimation. I Analytical techniques, *Applied Spectroscopy* 1977, 31: 404-415
30. P.J. Brown, Multivariate Calibration, Journal of the Royal Statistical Society B 1982, 44(3):287-308
31. C.D. Brown, Discordance between net analyte signal theory and practical multivariate calibration, *Analytical Chemistry* 2004, 76:4364-4373
32. J. Ye, On Measuring and Correcting the Effects of Data Mining and Model Selection, Journal of the American Statistical Association 1998, 93(441): 120-131.
33. C.G. Enke, L.J. Nagels, Undetected components in natural mixtures: how many? What concentrations? Do they account for chemical noise? What is needed to detect them? *Analytical Chemistry* 2011, 83: 2539-2546
34. C.D. Brown, T.D. Ridder, Framework for Multivariate Selectivity Analysis, Part I: Theoretical and Practical Merits, *Applied Spectroscopy* 2005, 59(6):787-803 1
35. E. Andries, J. H. Kalivas, Multivariate calibration leverages and spectral F-ratios via the filter factor representation, *Journal of Chemometrics* 2010, 24(5): 249-260.
36. R. Tibshirani, Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society B* 1996, 58(1):267-288