



Reliable and Transferable Raman “Core” Models for Monitoring Bioreactor Using the MarqMetrix All-In-One Process Raman Analyzer

Authors

Juan Villa¹, Matthew Zustiak¹, Lin Zhang²,
Abe Martin³, Nimesh Khadka²

¹ BioProduction Group, Thermo Fisher Scientific, St. Louis, Missouri, USA

² Analytical Instrument Group, Thermo Fisher Scientific, Tewksbury, Massachusetts USA

³ BioProduction Group, Thermo Fisher Scientific, Logan, Utah, USA

Introduction

Bioreactors have emerged as indispensable tools in bioprocessing, revolutionizing the production of many biotechnological products.^{1,2} They provide controlled and optimized environments for enhancing cell cultivation and are thus extensively used in various industries, ranging from biopharmaceuticals to food production. The consistency and high quality of products derived from bioreactors depend on the proper control of the critical process parameters (CPPs) like pH, temperature, dissolved oxygen, nutrient availability, cell density/viability, and others. Recognizing the importance of real-time monitoring and control of CPPs, the U.S. Food and Drug Administration (FDA) has emphasized the integration of process analytical technology (PAT) into pharmaceutical manufacturing following the principles of quality by design (QbD).³ The PAT framework promotes the use of advanced analytical technologies in an integrated synergy with chemical, physical, microbial, mathematical, and risk analyses to enable understanding and control of CPPs. As outlined in the FDA guidance, the overall objectives of PAT are as follows:

- Control process variability and ensure product consistency and quality.
- Reduce production cycle time.
- Prevent rejection, scrap, and reprocessing.
- Introduce the potential for real-time release.
- Increase automation to improve safety and reduce human error.
- Facilitate continuous processing to improve efficiency, manage variability, and improve capacity.

Process Raman is one such PAT solution that is rapidly being adopted by industries.⁴ Process Raman has several competitive advantages over other technologies for bioprocess monitoring. It offers benefits such as low water background interference, high molecular specificity, simple spectral interpretation due to fewer overtones and combination bands, the ability to probe biomolecules in their native state, non-destructive analysis, rapid measurements, and easy integration with automation.⁵ All these attributes make Process Raman a powerful PAT tool for monitoring cell cultures in a bioreactor. By integrating process Raman into bioreactors, operators can gain rapid, actionable, and accurate insights on CPPs, facilitating timely adjustments and optimizing bioprocess parameters. This real-time monitoring enables efficient process control, reduces batch-to-batch variability, and ensures uniform product quality throughout the manufactured batches.⁶

The success of process Raman for monitoring and controlling bioreactors depends on the performance of chemometric models.⁷ These models translate the spectral data into meaningful and actionable information. The chemometric models are developed in a multi-stage and iterative process (shown in Figure 1) that requires significant time, cost, resources, and technical expertise. The chemical compositions inside bioreactors are highly complex due to the presence of multiple components, each contributing its own Raman signature. This results in a mixture of overlapping spectral features. Furthermore, the dynamic nature of bioreactor systems, with continuous changes in component concentrations and interactions (e.g., different feeding materials and schedules), further adds complexity to the Raman spectra. Thus, a large set of training data with associated offline reference values is required to build robust, reliable, and accurate chemometric models. These models undergo multiple rounds of optimization and validation before they can be confidently deployed for process monitoring. Even after deployment, the performance of models needs to be monitored. Whenever there is a change in processes, the models may require updates or re-optimization using additional data, as shown in Figure 1. Thus, developing accurate and reliable chemometric models is often an obstacle to the widespread adoption of process Raman technology.

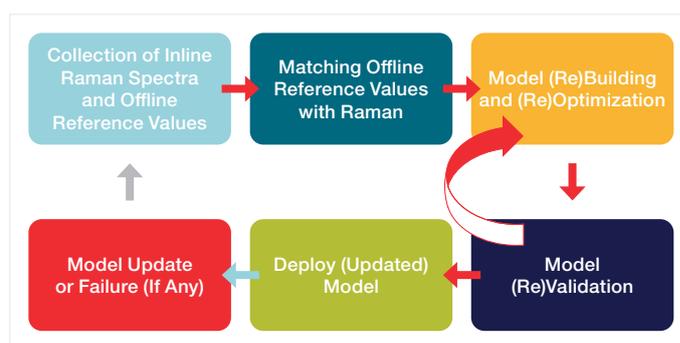


Figure 1. A general workflow of chemometric model building and implementation.

To address this bottleneck, we gathered the Raman data from various bioreactor runs and used it to develop the chemometric “core” models for real-time monitoring of glucose and lactate. These core models were tested across multiple cell lines, media, scales, and different processes and instruments. As demonstrated previously, the core model showed excellent transferability. In this note, we provide detailed information on these core models to facilitate easy integration of process Raman for bioreactor monitoring.

Experimental details

1. Data collection

The Raman data were collected in the “*in-line*” mode using a Thermo Scientific™ MarqMetrix™ All-In-One Process Raman Analyzer integrated with an immersible probe (Figure 2). The signal acquisition parameters were initially optimized and set to a power level of 450 mW, exposure time of 3000 ms, and 20 averages for all runs. Data were continuously acquired throughout the 2-week run, with one Raman spectrum recorded every 2 minutes.



Figure 2. MarqMetrix All-In-One Process Raman Analyzer with immersible probe.

The training data were collected in different bioreactors conducted using a variety of cell lines and different initial and feed media, as specified in Table 1. The bioreactors were operated in various modes, including fed-batch, perfusion, and hybrid perfusion, with different feeding strategies such as bolus, continuous, or a combination of both. Additionally, the data were collected from bioreactors of various scales, including 3 L and 5 L glass bioreactors and a 500 L Thermo Scientific™ DynaDrive™ Single-Use Bioreactor (S.U.B.). Finally, a total of six different instruments were used for data collection. All these variations in the training dataset were intentionally introduced to capture the variation with changes in process, scales, and instruments.

Run #	Cell Line	Initial Media	Feed Media 1	Feed Media 2	Feed Media Type	Glucose Feeding Type	Run Mode	Reactor Type
1	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Bolus	Bolus	Fed Batch	5L Glass
2	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Bolus	Bolus	Fed Batch	5L Glass
3	CHO-M	Balance CD CHO growth A + 6 mM L-glutamine + 1 g/L pluronic	Cell Boost 7a	Cell Boost 7b	Bolus	Bolus	Fed Batch	500L Dyna Drive
4	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Bolus	Bolus	Fed Batch	5L Glass
5	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	None	Continuous	Bolus	Hybrid Perfusion	500L Dyna Drive
6	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Bolus	Bolus	Fed Batch	500L Dyna Drive
7	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Bolus	Bolus	Fed Batch	5L Glass
8	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	Continuous, EFC 2X 3% Weight/Day	None	Continuous	Continuous/Bolus	Fed Batch	5L Glass
9	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Bolus	Bolus	Fed Batch	500L Dyna Drive
10	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	None	Bolus	Bolus	Perfusion	500L Dyna Drive
11	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	Continuous, EFC 2X 3% Weight/Day	None	Continuous	Continuous/Bolus	Fed Batch	5L Glass
12	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	Continuous, EFC 2X, VCV Model	None	Continuous	Continuous	Fed Batch	500L Dyna Drive
13	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	Continuous, EFC 2X, VCV Model	None	Continuous	Continuous/Bolus	Intensified Fed Batch	5L Glass
14	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	Continuous, EFC 2X Glucose free, 3% Weight/Day	None	Continuous	Continuous/Continuous	Fed Batch	5L Glass
15	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	None	Continuous	Continuous	Perfusion	5L Glass
16	CHO K1	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 51 ml	None	Bolus	Bolus	Fed Batch	3L TruBio6
17	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	None	Continuous	Continuous	Perfusion	5L Glass
18	NistCHO	EX-CELL® Advanced™ CHO Fed Batch medium	NISTCHO is EX-CELL® Advanced™ Feed	None	Bolus	Bolus	Fed Batch	3L TruBio6
19	HEK293	Expi293 expression media	Expi293 expression media	None	Continuous	Continuous	Perfusion	5L Glass
20	ExpiCHO 1C6	HipCHO media	HipCHO media	None	Continuous	Continuous	Perfusion	5L Glass
21	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Continuous	Continuous	Fed Batch	500L Dyna Drive
22	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	None	Continuous	Continuous	Perfusion	5L Glass
23	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Continuous	Continuous	Fed Batch	3L TruBio6
24	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Bolus/Continuous	Bolus/Continuous	Fed Batch	5L Glass
25	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	None	Continuous	Continuous	Perfusion	3L TruBio6
26	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X, Capacitance VCV model	None	Continuous	Continuous	Fed Batch	500L Dyna Drive
27	ExpiCHO 1C6	HipCHO media	HipCHO media	None	Continuous	Continuous	Perfusion	3L TruBio6
28	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X 3% Weight/Day	None	Continuous	Continuous	Fed Batch	3L TruBio6
29	CHO-K1	HyCell CHO Medium	CellBoost 7A	Cell Boost 7b	Bolus	Bolus	Fed Batch	5L Glass
30	ExpiCHO 1C6	HipCHO media	HipCHO media	None	Continuous	Continuous	Perfusion	500L Dyna Drive
31	ExpiCHO 1C6	ExpiCHO SPM + 6 mM L-Glutamine + 2 g/L pluronic	EFC 2X spiked Met and Phe, Capacitance VCV model	None	Continuous	Continuous	Fed Batch	500L Dyna Drive
32	ExpiCHO 1C6	HipCHO media	HipCHO media	None	Continuous	Continuous	Perfusion	5L Glass
33	ExpiCHO 1C6	HipCHO media	HipCHO media	None	Continuous	Continuous	Perfusion	5L Glass
34	CHO-K1	Efficient PRO	Efficient PRO Feed 2	None	Continuous	Continuous	Fed Batch	500L Dyna Drive
35	ExpiCHO 1C6	Efficient PRO	Efficient PRO	None	Continuous	Continuous	Perfusion	5L Glass

Table 1.

During the bioreactor run, samples were aseptically pooled and sent for at-line or offline reference value analysis. The timestamp for pooled samples was recorded for data analysis.

Analyte	Model Type	No. of Latent Variables	Region Selection cm^{-1}	Preprocessing
Glucose	PLS	5	1065-1232; 1595-1863; 2704-3078	Sav-Gol filter (1 st Derivative; order = 2; Window width = 13) + SNV + Mean Center
Lactate	PLS	5	800 - 1750	Sav-Gol filter (1 st Derivative; order = 2; Window width = 11) + L-1 Norm (Area = 1 for 1540-1750 cm^{-1}) + Mean Center

Table 2.

2. Data management

The initial data management was performed in the Python platform. More than 9000 Raman spectra were collected in .spc file format for each bioreactor run. The timestamp for each spectrum was extracted from its metadata and matched with the timestamp of the reference values for the target analytes. To improve the signal-to-noise ratio (SNR), five Raman data with timestamps centered around the time of each reference were selected and averaged. This resulted in each averaged Raman spectrum having a total exposure time of 5 minutes. The averaged Raman spectra were used for chemometric modeling. The chemometric models were developed using Python and Eigenvector SOLO software. The procedure was repeated for all target analytes.



Thermo Scientific High-Intensity Perfusion CHO Medium

3. Chemometric calibration model development

Before building the chemometric models, the quality of the spectra was assessed visually. The bioreactor Raman data consists of both Raman scattering and background signal from fluorescence, and baseline shift due to Mie scattering. Occasionally, the detector is saturated or reaches saturation due to a high background signal. This adds non-linear noise to Raman spectra. To avoid the inclusion of Raman spectra with low SNR in the training data, the Raman spectra were plotted and visually inspected. Any spectra with an intensity exceeding 50,000 counts in the 800 to 3250 cm^{-1} spectral range were excluded from the training set. The data were then preprocessed by removing the baseline using the Savitzky-Golay (Sav-Gol) filter, followed by normalization either by standard normal variate (SNV) or using area or peak intensity of the water band, as shown in Table 2. All the data were mean centered before developing the partial least square (PLS) regression models.

The PLS regression models were developed using only the selected regions as specified in Table 2. A leave-one-out (leave-a-run-out) cross-validation (LOOCV) strategy was used for the internal validation of the models. Initially, each model was built using different numbers of latent variables (LVs) ranging from 1 to 20. The root mean square error of calibration and cross-validation (RMSEC and RMSECV) were calculated for each model with different numbers of latent variables. The PLS model with the optimum number of latent variables was selected such that, one, adding more latent variables did not significantly improve the RMSECV; and (two) the ratio of RMSEC and RMSECV was close to 1.



Gibco™ Efficient-Pro™ Feed 2



Gibco™ Efficient-Pro™ Medium



Gibco™ Expi293™ Expression Medium

Chemometric models development

Glucose “Core” Model:

The PLS regression model for glucose was developed using the region of the Raman spectra corresponding to specific vibrational modes of the glucose, as shown in Figure 3. The Raman spectra of glucose in water and in the bioreactor after applying the SavGol filter (second derivative, order =2, window width =13) are shown in Figures 3A and 3B, respectively. The negative peak at $\sim 1125\text{ cm}^{-1}$ is assigned to the stretching vibrational mode of CO and CC and in-plane bending of COH bonds ($\nu(\text{CO})$, $\nu(\text{CC})$, $\beta(\text{COH})$) of a glucose molecule. Thus, as listed in Table 2 and Figure 4A, the glucose model was developed using three spectral regions:

$1065\text{--}1232\text{ cm}^{-1}$, $1595\text{--}1863\text{ cm}^{-1}$, and $2704\text{--}3078\text{ cm}^{-1}$.

The $1065\text{--}1232\text{ cm}^{-1}$ spectral region includes the characteristic Raman glucose peak. The spectral region of $1595\text{--}1863\text{ cm}^{-1}$ includes the Raman water peak. The concentration of water remains unchanged throughout the bioreactor run. Thus, the $1595\text{--}1863\text{ cm}^{-1}$ spectral region is used as a reference for spectral normalization. During the preprocessing, the mathematical operation of SNV mainly utilizes the dominant intensity from the spectral region of $1595\text{--}1863\text{ cm}^{-1}$ to normalize each spectrum by correcting the path length differences. Similarly, the spectral region of $2704\text{--}3078\text{ cm}^{-1}$ includes a Raman peak assigned to the symmetric and antisymmetric stretching vibration modes of CH_2 and CH bonds of glucose. Including these regions in the model ensured the model's accuracy and selectivity for glucose concentration prediction.

The glucose PLS model was developed using five latent variables for the 0 to 12 g/L concentration range, as shown in Figure 4B. The RMSECV did not improve after five latent variables, as shown in Figure 4C. Thus, a five-latent-variable PLS model was selected. The loadings for these latent variables that contain glucose information are shown in Figure 4D. The overall model statistics are shown in Table 3.

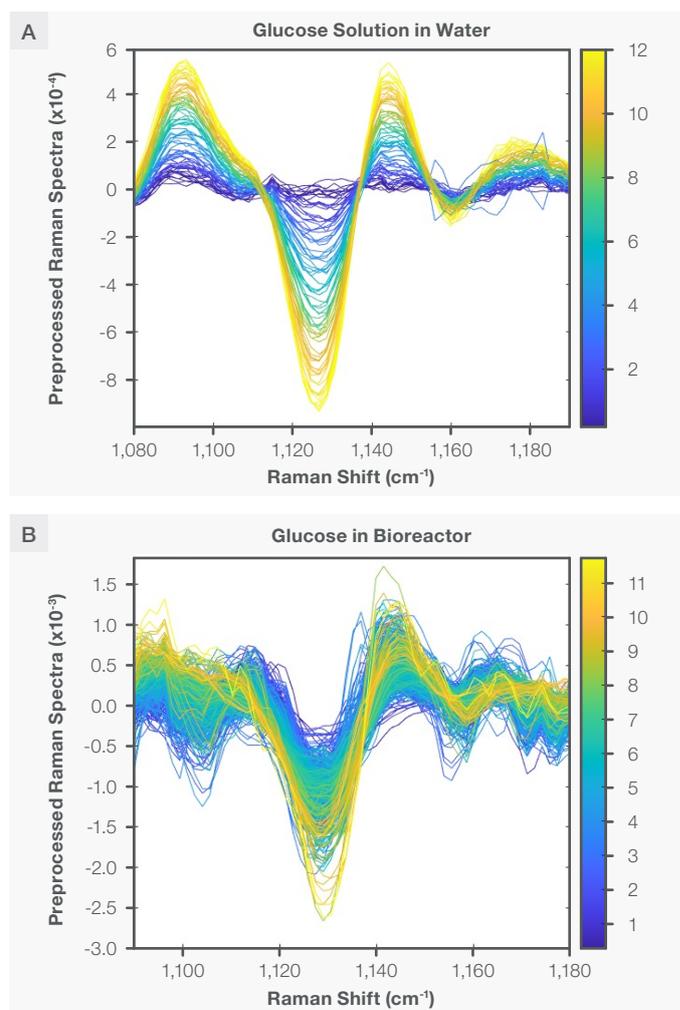


Figure 3. Characteristic Raman peaks of glucose at $\sim 1125\text{ cm}^{-1}$ in aqueous solution (plot A) and bioreactor (plot B) as second derivative (negative peaks) spectra. The spectra are color coded by the concentration shown as the vertical bar.

Glucose Core Model Parameters	Statistics
Model Range	0-12 g/L
Number of Latent Variables	5
RMSEC	0.43 g/L
RMSECV	0.49 g/L
R ² CV	0.94
CV Bias	-0.013 g/L

Table 3.

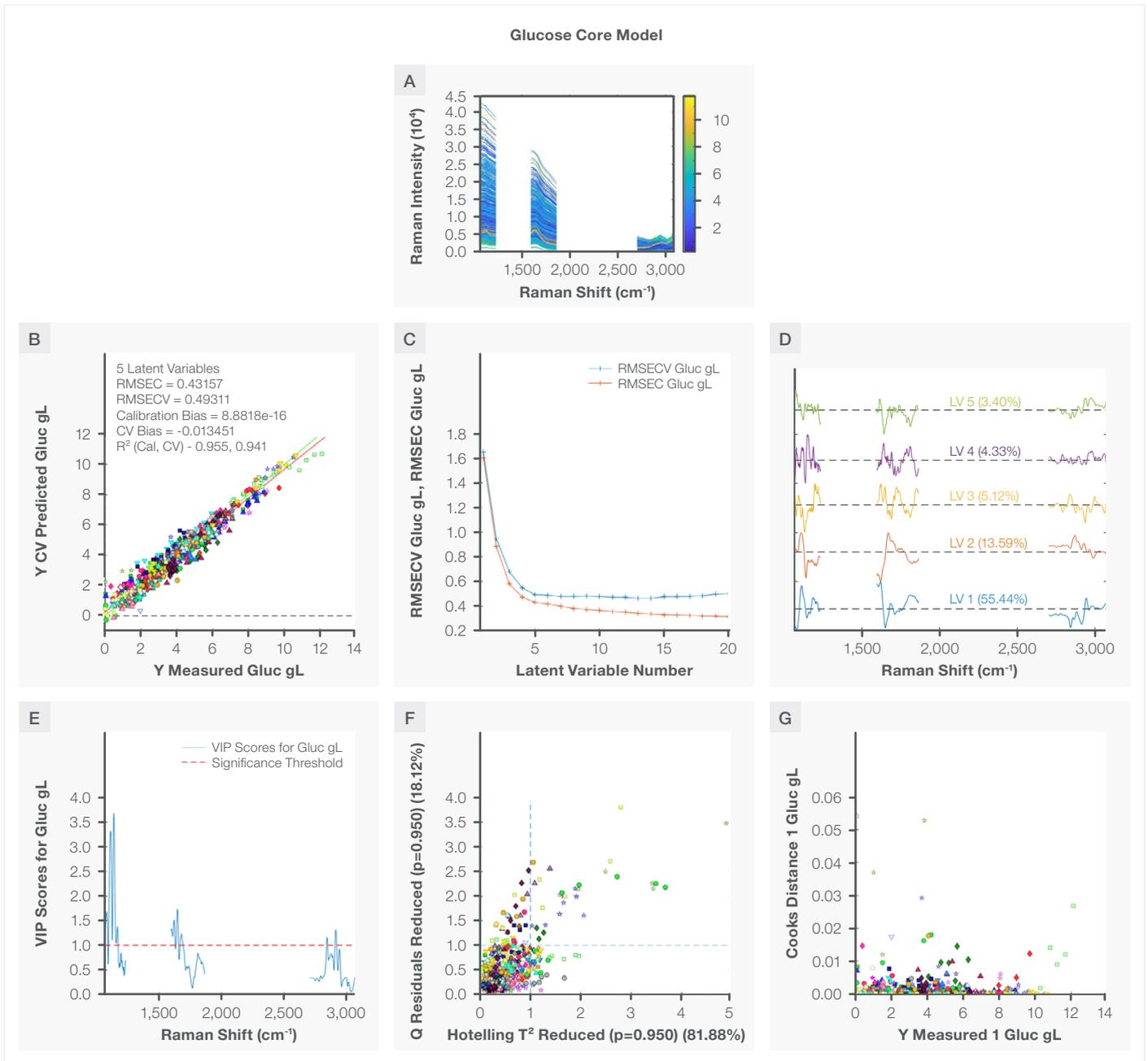


Figure 4. Glucose model. Plot A shows the region used to develop the model, plot B is the correlation plot between the measured and the prediction during cross validation, plot C shows RMSEC and RMSECV vs number of latent variables, plot D shows loading with percent variance captured, plot E shows VIP scores, plot F shows the reduced Q residual vs reduced Hotelling T^2 plot, and plot G is the Cook distance for all training samples.

Evaluation of the Glucose “Core” Model

i. Specificity: The specificity for the glucose core model was evaluated by calculating the variable importance in projection (VIP) scores. The VIP score plot shows the importance of each Raman shift in explaining the response variable, glucose concentration. The VIP scores of the glucose core model are shown in Figure 4E. The red dotted horizontal line represents the commonly used significance threshold (VIP score =1). Any Raman shift with a score higher than 1 is considered important for the model. As shown in the VIP score plot for the glucose model, the spectral region $\sim 1125\text{ cm}^{-1}$ has significantly higher scores than those from other regions. This spectral region is specific to the vibrational mode of glucose, as explained above in Figure 3. Thus, the VIP score plot is one of the mathematical assessments that indicate the model's specificity for glucose. In this study, we preemptively developed the model by selecting a specific region to improve model specificity.

ii. Accuracy/precision: The glucose core model was developed with a glucose concentration of 0 to 12 g/L in the training dataset. Its accuracy and precision were evaluated using RMSECV and CV bias. As listed in Table 3, the low RMSECV of 0.49 g/L and low CV bias of -0.013 g/L demonstrate the model's high accuracy for the concentration range of 0 to 12 g/L. The RMSECV can be treated as the normalized standard deviation across the concentration range.

The total measurement error is the square root of the sum of the squares of bias error and precision error.⁸ Using RMSECV as total error and CV bias as bias error, the precision error was calculated to be 0.48 g/L. The RMSECV of 0.49 g/L and the precision of 0.48 g/L also indicate that the model is statistically more accurate and reliable above the glucose concentration of 1.5 g/L ($3 \times \text{RMSECV}$). It can monitor or control glucose concentration in the process with a tolerance limit of 0.5 g/L.

iii. Linearity: The R^2 CV for the glucose core model is 0.94, demonstrating the linearity of spectral change across the range of 0 to 12 g/L.

iv. Q vs T^2 analysis and recommendations: The Q residual and Hotelling T^2 are other model evaluation criteria. The Q residual is the lack-of-fit statistic that measures the residual after projecting the data into the model space. It measures how well each sample conforms with the model. Hotelling T^2 measures the sample variation within the model by estimating how far the sample is from the center of distribution of the model. The Q residual and Hotelling T^2 statistics are sensitive to the total number of variables in the data, the number of latent variables, and the particular preprocessing used for a model. In contrast, normalized statistics called reduced Q residual and reduced Hotelling T^2 are calculated by dividing the Q residual and Hotelling T^2 with the corresponding confidence interval. The distribution of the reduced Q residual and reduced Hotelling T^2 , with a 95% confidence interval for the core glucose model, is shown in Figure 4F. The majority of the data resides within the boundaries limit (dotted blue line) of reduced Q residual < 1 and reduced Hotelling $T^2 < 1$. Few data points are out of the 95% confidence boundaries, which is expected for a model developed with such wide varieties of training data (different cell lines, media, and operational modes); the distribution of variance is not equivalence.

When the model is applied to new processes or instruments, the reduced Q residual and reduced Hotelling T^2 distribution plot should be used as an assessment criterion. The prediction from model can only be trusted with high confidence if the test samples fall with the model space. Under the scenario where the test samples have high values of reduced Q residual and reduced Hotelling T^2 , it is recommended that the predictions be validated using orthogonal reference techniques. It is also recommended that the limits for setting the boundaries of reduced Q residual and reduced Hotelling T^2 should be determined by the user only after applying the model to the statistically significant number of process runs.

v. Outlier analysis: To calculate if any of the data points are outliers or are highly influential in the model, especially the ones that lie out of 95% confidence boundaries in the reduced Q residual and reduced Hotelling T^2 plot, the Cook's distance was calculated. This measures the change in the regression estimates when a particular observation is removed from the dataset. It quantifies the effect of that observation on the overall model fit. A high Cook's distance for a specific observation indicates that its removal from the dataset would substantially impact the estimated regression coefficients and, therefore, might be an outlier or influential data. As shown in Figure 4G, all data have low Cook's distance, demonstrating no obvious outliers or influential data in the training set.

Lactate “Core” Model

The lactate core model was developed using the same strategy described above for the glucose core model. Figure 5 shows the characteristic Raman peak for lactate at $\sim 860 \text{ cm}^{-1}$ in water and in the bioreactor after applying the SavGol filter (second derivative, order =2, window width=13). To provide specificity toward lactate, the model was developed using a single spectral region (800 to 1750 cm^{-1}) that includes the characteristic Raman peak for lactate and water band for normalization, as shown in Figure 6A.

The lactate model was developed using five latent variables (Figure 6B). The choice of five latent variables was based on the result of the predicted residual error sum of squares (PRESS) against the number of latent variables, as shown in Figure 6C. There was no significant improvement in RMSECV after five latent variables, so the five latent variables PLS model was selected as the optimal one. The loadings for the model are shown in Figure 6D. The model statistics are summarized in Table 4.

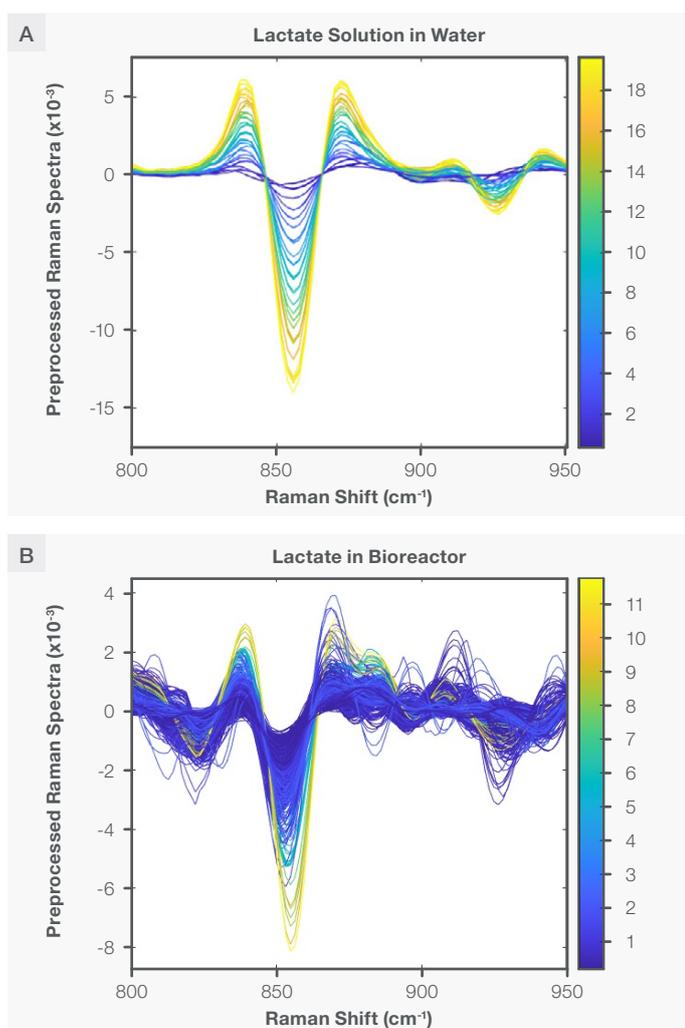


Figure 5. Characteristic Raman peaks of lactate at $\sim 860 \text{ cm}^{-1}$ in aqueous solution (plot A) and bioreactor (plot B) as second derivative (negative peaks) spectra. The spectra are color-coded by the concentration shown as the vertical bar.

Lactate Core Model Parameters	Statistics
Model Range	0-12 g/L
Number of Latent Variables	5
RMSEC	0.23 g/L
RMSECV	0.31 g/L
R ² CV	0.92
CV Bias	-0.0174 g/L

Table 4.

Evaluation of Lactate “Core” Model

- i. Specificity: The VIP score plot (Figure 6E) demonstrates that the region from ~ 820 to 880 cm^{-1} is important in the lactate model. This region is associated with the stretching vibrational mode of the C-COO⁻ bond of lactate, as shown in Figure 5. Thus, the VIP score plot demonstrates the model specificity for lactate.
- ii. Accuracy/precision: As explained above for the glucose core model, the RMSECV of 0.31 g/L, CV bias of -0.0174 g/L, and precision error of 0.31 g/L for the concentration range of 0 to 12 g/L demonstrate that the model's accuracy and precision are within the acceptable tolerance for typical bioreactor process monitoring.
- iii. Linearity: The linearity of the lactate core model is demonstrated by CV R² of ~ 0.94 for the concentration range of 0 to 12 g/L. Considering the diverse bioprocess conditions in bioreactors where training sets were collected, the CV R² of ~ 0.94 represents the high correlation between spectral information and measured concentration. Note that this metric from lactate is not as good as that from glucose, and this is because the “net analyte signal” of lactate is less than that of glucose.
- iv. Q vs. T² analysis and recommendations: The distribution of the training data in the Q vs. T² biplot related to 95% confidence boundaries is shown in Figure 6F. The 95% confidence boundaries of raw Q and T² values from the training data set are used in the denominators to normalize reduced Q residual and reduced Hotelling T² to 1, respectively. The dotted blue line represents these boundaries. The predictions from the model are trusted with high confidence if the test samples fall within the 95% confidence boundaries. When a test sample generates reduced Q residual or reduced Hotelling T² significantly higher than 1, an additional reference test is recommended to validate the prediction.
- v. Outlier analysis: The low Cook's distance for the training data indicates that there was no outlier in the training data, as shown in Figure 6G.

Lactate Core Model

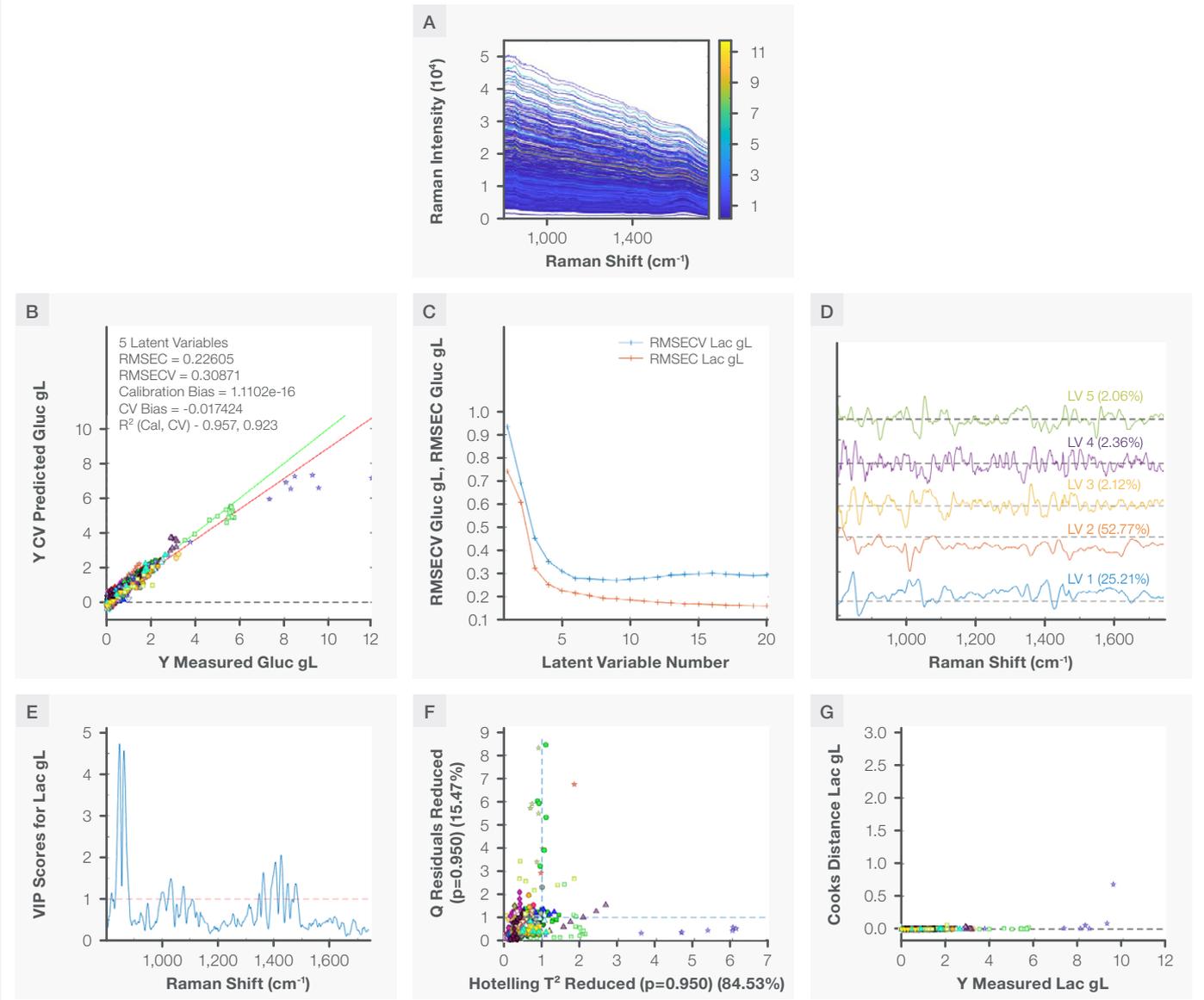


Figure 6. Lactate model. Plot A shows the region used to develop the model, plot B is the correlation plot between the measured and the prediction during cross validation, plot C shows RMSEC and RMSECV vs. number of latent variables, plot D shows loading with percent variance captured, plot E shows VIP scores, plot F shows the reduced Q residual vs. reduced Hotelling T^2 plot, and plot G is the Cooks distance for all training samples.

Performance of Glucose and Lactate “Core” Model

The performance of glucose and lactate core models described in this note was tested on five different cell lines/media, different scales of bioreactors, and also for the automated feedback control.^{9,10} The average root mean square of prediction (RMSEP) for glucose was ~ 0.5 g/L, while the average RMSEP for lactate was ~ 0.2 g/L. The low RMSEP for CHO and HEK cell lines in different cell media demonstrates the models' accuracy, reliability, and transferability across different processes and scales.

Conclusion

The accurate, reliable, and transferable glucose and lactate Raman “core” models were developed for bioreactor monitoring and control. By utilizing these models, users can save valuable time, costs, and resources that would otherwise be required for developing chemometric models. With the availability of these models, users can immediately initiate automated glucose feeding for their bioreactor while also taking advantage of the additional benefits offered by Raman as an in-line Process Analytical Technology (PAT) solution. This includes enhancing and regulating the process, minimizing batch-to-batch variability, and ensuring consistent product quality and quantity.

References

1. *Bioreactors: Design, Operation and Novel Applications* | Wiley. Wiley.com.
2. Zhong, J.-J. Recent Advances in Bioreactor Engineering. *Korean J. Chem. Eng.* **2010**, *27* (4), 1035–1041. <https://doi.org/10.1007/s11814-010-0277-5>.
3. Research, C. for D. E. and. PAT — *A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance*. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/pat-framework-innovative-pharmaceutical-development-manufacturing-and-quality-assurance> (accessed 2024-06-19).
4. Buckley, K.; Ryder, A. G. Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing: A Short Review. *Appl. Spectrosc.* **2017**, *71* (6), 1085–1116. <https://doi.org/10.1177/0003702817703270>.
5. Esmonde-White, K. A.; Cuellar, M.; Lewis, I. R. The Role of Raman Spectroscopy in Biopharmaceuticals from Development to Manufacturing. *Anal. Bioanal. Chem.* **2022**, *414* (2), 969–991. <https://doi.org/10.1007/s00216-021-03727-4>.
6. Webster, T. A.; Hadley, B. C.; Dickson, M.; Hodgkins, J.; Olin, M.; Wolnick, N.; Armstrong, J.; Mason, C.; Downey, B. Automated Raman Feed-Back Control of Multiple Supplemental Feeds to Enable an Intensified High Inoculation Density Fed-Batch Platform Process. *Bioprocess Biosyst. Eng.* **2023**, *46* (10), 1457–1470. <https://doi.org/10.1007/s00449-023-02912-2>.
7. Guo, S.; Popp, J.; Bocklitz, T. Chemometric Analysis in Raman Spectroscopy from Experimental Design to Machine Learning–Based Modeling. *Nat. Protoc.* **2021**, *16* (12), 5426–5459. <https://doi.org/10.1038/s41596-021-00620-3>.
8. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.-M.; McBratney, A. Critical Review of Chemometric Indicators Commonly Used for Assessing the Quality of the Prediction of Soil Attributes by NIR Spectroscopy. *TrAC Trends Anal. Chem.* **2010**, *29* (9), 1073–1081. <https://doi.org/10.1016/j.trac.2010.05.006>.
9. Villa, J.; Zustiak, M.; Kuntz, D.; Zhang, L.; Khadka, N.; Broadbelt, K.; Woods, S. Use of Lykos and TruBio Software Programs for Automated Feedback Control to Monitor and Maintain Glucose Concentrations in Real Time.
10. Villa, J.; Zustiak, M.; Ramirez, D.; Kruger, J.; Kuntz, D.; Zhang, L.; Khadka, N.; Broadbelt, K.; Woods, S. Demonstrating Chemometric Model Transferability for 5 Mammalian Cell Lines and 5 Media Types Using the Thermo Scientific MarqMetrix All-In-One Process Raman Analyzer to Monitor Upstream Bioprocesses.

Learn more at thermofisher.com/marqmetrixAIO