Technical note

A Classical Least Squares (CLS) Approach for Protein Quantification in Downstream Processing Using Raman Spectroscopy

Authors

Nimesh Khadka[1], Ph.D.

Kristina Pleitt[2], Ph.D.

Michelle Nolasco[2]

[1]Analytical Instrument Group, Thermo Fisher Scientific, Tewksbury, Massachusetts USA

[2]Bioproduction Group, Thermo Fisher Scientific, St. Louis, Missouri USA

Industry/Application

Biopharma PAT / Downstream processing

Products used

Thermo Scientific™ MarqMetrix™ All-In-One Process Raman Analyzer, Thermo Scientific™ MarqMetrix™ FlowCell™ Sampling Optic

Goals

Demonstrate quick and easy chemometric strategies to develop an accurate protein quantification model for downstream applications using only a single spectrum of known concentration. Highlight the developed strategy's performance and transferability across quantification of other types of monoclonal antibodies (mAbs), different matrices, and processes.

Key analytes

Protein (mAb) quantification

Key benefits

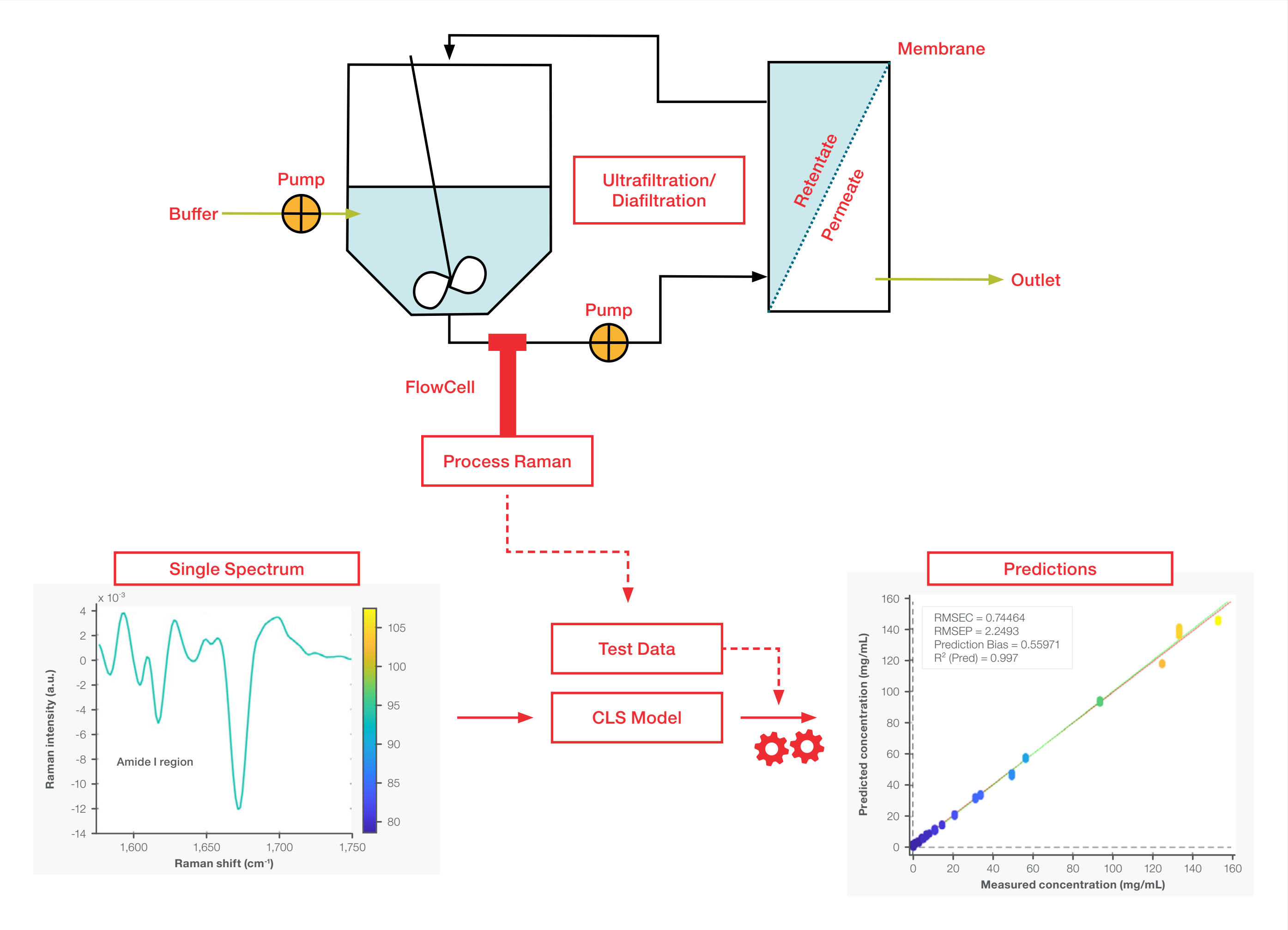
- Cost and time savings from eliminating the need for extensive training data collection and laboratory analytics
- Enable immediate deployment into process monitoring and control, allowing customers to leverage the numerous benefits of process Raman spectroscopy

The successful deployment of Raman spectroscopy to monitor complex processes relies on the robustness of chemometric models, which typically require large datasets and sophisticated algorithms. The time and cost associated with generating extensive and reliable training datasets often hinder the adoption and integration of Raman technology in biopharmaceutical applications. In this work, we introduce a chemical information-based approach to develop a robust and accurate chemometric model with a minimal training dataset. We employed the classical least squares (CLS) algorithm using a selected region of a single spectrum of a protein with a known concentration to develop a protein quantification model for monitoring ultrafiltration/diafiltration (UF/DF) in downstream processing. The performance of the CLS model and its transferability across different monoclonal antibodies are discussed. This approach can also be leveraged to build quick and accurate chemometric models for various other applications, making Raman technology more accessible for adoption and utilization.

Data acquisition and chemometric modelling

A monoclonal antibody (mAb; protein) at a concentration of 93.56 mg/mL in a formulation aqueous buffer containing histidine, arginine, and sucrose was passed through a Thermo Scientific™ MarqMetrix™ FlowCell Sampling Optic at a flow rate of 100 mL/min. Raman data were acquired during the dynamic flow using a Thermo Scientific™ MarqMetrix™ All-In-One Process Raman Analyzer with acquisition parameters of 450 mW power, 3000 ms integration time, and 3 averages. Ten spectra were acquired and preprocessed to remove any cosmic ray interference. The ten spectra were then averaged into a single spectrum that was used to build the CLS model.

Two spectral regions were selected before developing the CLS model: 3100 to 3230 cm^{-1} (water band) and 1570 to 1750 cm^{-1} (protein amide I region). The infinity norm for the 3100 to 3230 cm^{-1} spectral region was calculated and used as a weight to normalize the entire spectrum, correcting spectral path differences. Baseline features were removed by applying the Savitsky-Golay filter (2nd derivative, polynomial order = 2, window width = 13). The derivatized spectrum was then used to develop the CLS model, which in essence is a *ratio-metric model* that translates the ratio of Raman intensity of the amide I band to the water band into predicted concentrations.



Demonstrating a simpler chemometric approach for quantifying protein concentration in a downstream ultrafiltration/diafiltration (UF/DF) processing using process Raman.

Following the CLS model development, validation data were acquired using the same parameters over the mAb concentration range of 0 to 155 mg/mL in the same formulation buffer. The model's performance was further validated by applying it to the UF/DF process with different mAbs and by including tryptophan in the formulation buffer alongside histidine, arginine, and sucrose. A brief comparison of CLS and partial least square (PLS) models was also performed.

All data management, cosmic ray removal, averaging, and timestamp alignment were performed in Python™ programming language. The data were then processed in a commercially available chemometric package. All chemometric works were performed using software package SOLO 9.3.1 (2024), Eigenvector Research. Inc., Manson, WA USA 98831.

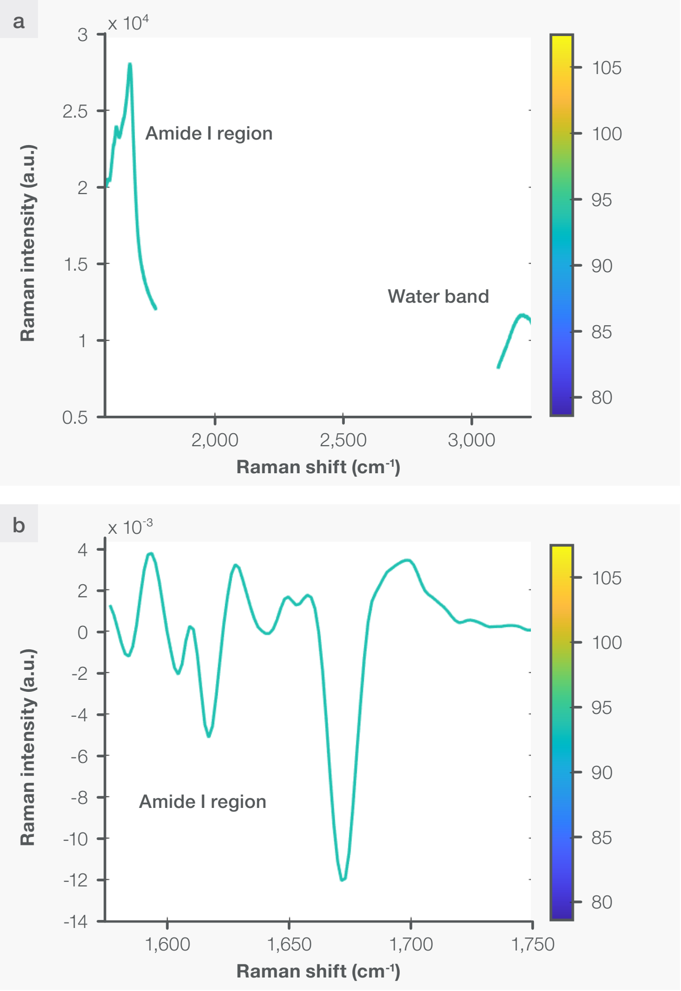


Figure 1. (a) Selected region used for chemometric model development. The amide I region provides specificity to the model for mAb while the water band is used for correction of spectral path length differences. (b) Showing the 2nd derivative preprocessed spectrum of amide I region used for model development.

Results and discussion

Standard normal variate (SNV) is a widely used normalization technique for spectroscopic data to correct for path length differences.[1] While effective when the primary contribution to variables is noise and they share the same overall signal, SNV may lead to non-linear responses if the overall signal changes significantly between samples.[2] This may especially be true in downstream processes where concentrations are high and dynamic changes result in rapid changes in spectral features and intensities. In this study, we opted to use the water band as an internal standard to normalize the spectra as the water concentration remains relatively constant throughout the bioprocesses.[3–5]

The region from 3100 to 3230 cm^{-1} includes the Raman band attributed to the symmetric stretching of the O-H vibrational bond in water molecules. The O-H stretching vibrational band is susceptible to changes in pH, ionic strength, and temperature; nevertheless these parameters are well-defined and controlled during process development. This ensures the reliability of this region for spectral normalization. Additionally, unlike the symmetric bending vibration of water molecules at 1640 cm^{-1}, the 3100 to 3230 cm^{-1} spectral region has minimal spectral overlap from the constituents commonly used in the downstream processes. This makes it a viable option for spectral normalization.[6]

One concern is the low quantum efficiency of silicon optical sensors in this high Raman shift region. We, and others, have previously used this region for modelling, and the region's performance has proven to be acceptable, likely because the high concentration of water (~55.5 M) compensates for the limitation in efficiency.[3–5]

Figure 1a shows the average spectrum used for building the training model, while Figure 1b presents the 2nd derivative plot of the Raman spectra in the amide I region of mAb. The amide I region spans the Raman shift from approximately 1630 to 1700 cm^{-1}, primarily influenced by the variations in the energy of C=O symmetric stretching vibrations in different secondary structures of mAb[7]. Our previous work has demonstrated that the amide I region is free of spectral interferences and can be effectively utilized to develop models with high specificity for mAb for downstream processes.[3,5] Consequently, the amide I region was selected for the CLS model. The Savitsky-Golay filter (2nd derivative) was used in the preprocessing step to remove the baseline shifts as well as the broad water band, ensuring that the Raman information from the mAb is used to train the model.

The CLS model is a quantitative analytical method that explains the observed spectrum of a given sample by using the linear combination of the spectra of the pure components present in the sample.[8] For the CLS model to perform accurately, it is essential to acquire the pure spectrum for each component of the mixture, which is practically challenging or impossible for a complex bioprocess. This problem is addressed in this study by using the mAb-specific amide I region that is free of spectral overlap. If a larger region is used to create the model, the CLS model showed a decrease in performance (data not shown).

The loading for the CLS model is shown in Figure 2. The model has high influence from the Raman shift at approximately 1670 cm^{-1} in the amide I region, which is assigned to the symmetric C=O stretching of the β-sheet secondary structure of the mAb, thus providing specificity to the model. As expected for the single-component CLS model, the loading and the preprocessed spectrum are similar (compare Figures 1b and 2).

The CLS model was then applied to the validation dataset (shown in Figure 3a) across a wide concentration range in different buffer matrices: 0 mg/mL in water; 1 to 33 mg/mL in tris buffer; and 33 to 155 mg/mL in histidine, arginine, and sucrose buffer. Across the concentration range in diverse buffer components, the root mean square error of prediction (RMSEP) was approximately 2.25 mg/mL as shown in the correlation plot of Figure 3b. The low RMSEP for the concentration range of 0 to 155 mg/mL demonstrates excellent model performance and transferability across buffer matrices.

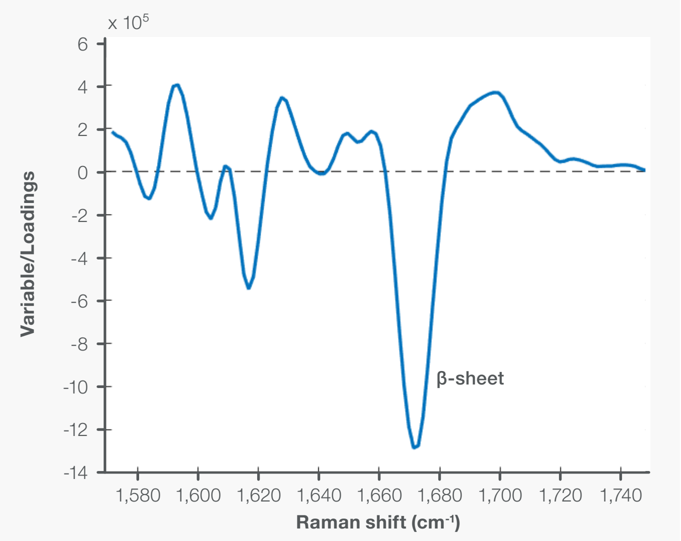


Figure 2. The regression vector for the CLS model demonstrating the influence of the approximately 1670 cm^{-1} Raman shift in the amide I region, which is assigned to the symmetric C=O stretching of the β-sheet secondary structure of the mAb.

Note that the spectra shown in Figure 3a exhibit differences in water band intensities at approximately 3240 cm^{-1}. The water concentration is relatively constant across the samples.

These differences in intensities within the water band indicate variations in optical path length during data acquisition, caused by turbidity and occasionally small air bubbles trapped inside the MarqMetrix FlowCell probe. The inclusion of water band normalization in the CLS model appropriately corrects for the path length differences and improves prediction accuracy. Additionally, no baseline was removed before water band normalization. Although the data is not shown, the CLS model demonstrated similar performance with or without baseline removal (using automatic Whittaker filter and automatic weighted least squares) before water band normalization.

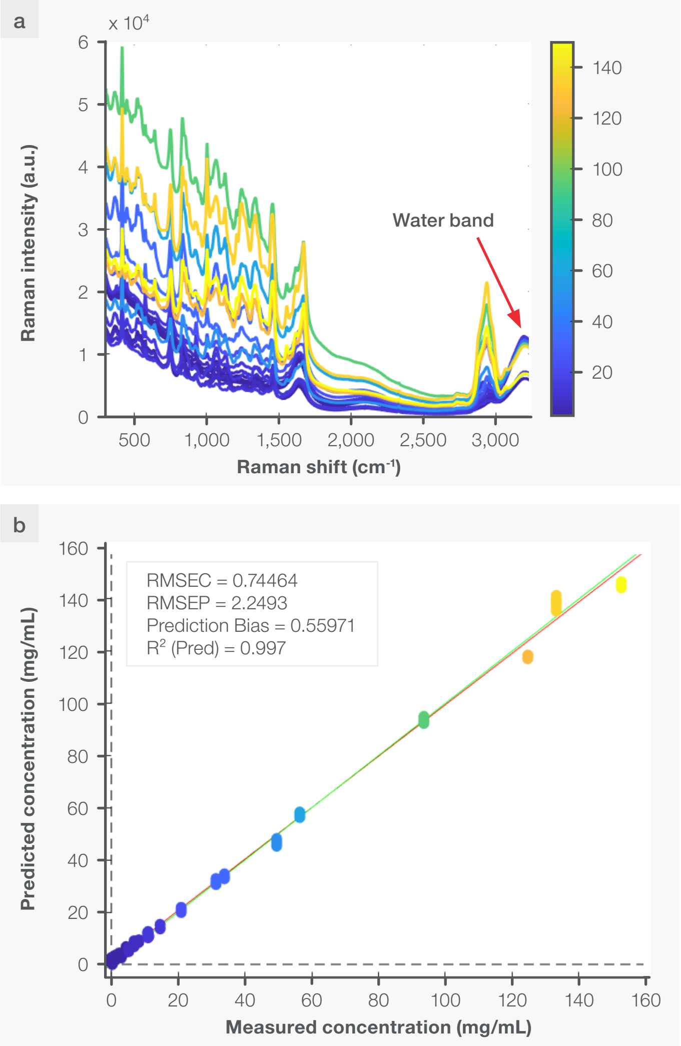


Figure 3. (a) Spectral data used for validation of the CLS model. Also, highlighted are the differences in the intensities of water band across spectra which substantiate the need of inclusion of the water band normalization step in the model. (b) Correlation plot for measured vs. predicted concentration along with performance statistics shown as inset.

Reference concentration (mg/mL)	PLS model predictions (mg/mL)	CLS model prediction (mg/mL)	PLS model average abs. % error	CLS model average abs. % error
30.72	30.6 ± 0.4	30.8 ± 0.1	1.12	0.36
155.9	146 ± 1	154 ± 1	6.15	0.67

Table 1. Performance of CLS model.

To validate the performance of the CLS model, the training and validation data used to develop the CLS model were combined into a single dataset. The combined data was fed into the PLS algorithm using the same spectral region and preprocessing as the CLS model. A one-latent-variable PLS model was selected based on the leave-one-out cross-validation (LOOCV) strategy, where each class was left out once. The root mean square error of cross-validation (RMSECV) was calculated, as shown in inset of Figure 4. The RMSECV for the PLS model was 2.88 mg/mL, while the RMSEP for the CLS model was 2.24 mg/mL. The RMSEP of CLS model and RMSECV of the PLS model is not a direct comparison, nonetheless, with some approximation, these results do indicate that the CLS model performed comparable to the PLS model.

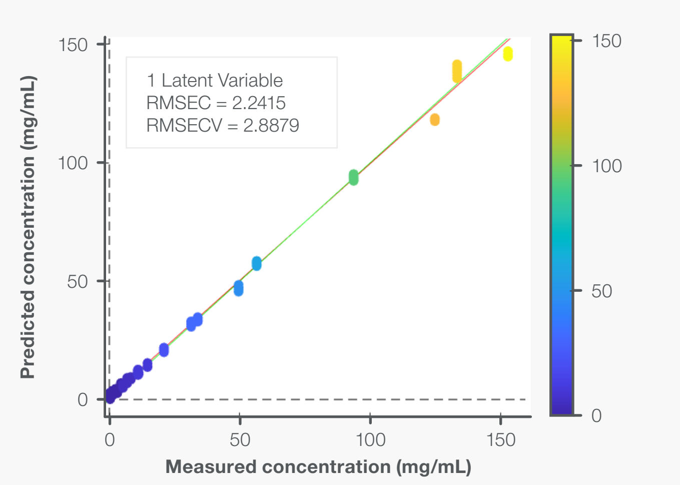


Figure 4. PLS model shown with model statistics in inset. The training and test dataset used for the CLS model were combined to develop the PLS model. The RMSECV of the PLS model calculated using leave-one-out cross validation is close to the RMSEP for the CLS model, indicating similar model performance.

To further test the model's performance, scalability, and transferability, we applied the lab-based PLS and CLS models to Raman data collected in-line during a UF/DF pilot run. This run used a different type of mAb and a formulation buffer. The buffer included tryptophan, histidine, arginine, and sucrose that were added during diafiltration. We have described the relevance of this experiment before.[3] The predicted protein concentrations from the CLS and PLS models showed a high correlation. This is shown in Figure 5 with the orange and blue traces. The pooled samples (marked by red stars) were measured using HPLC and UV-Vis spectroscopy. The absolute prediction errors for both models are shown in Table 1.

The predictions from the CLS model exhibited lower errors compared to the PLS model, however it *does not mean CLS is superior to PLS based on a statistically insignificant sample size of n=1 dataset*. In this study, the CLS model was built by selecting the amide I region that has high specificity for mAb and minimum spectral interferences If the entire spectral region was used with complex spectral overlap or in cases with ill-conditioning matrix and multicollinearity, PLS or other regression models are better choices with a multitude of other advantages. Additionally, using different regions of spectra and other combinations of preprocessing, the performance of the PLS model may be further improved in the above case. Here, PLS is used only as reference but not for comparison.

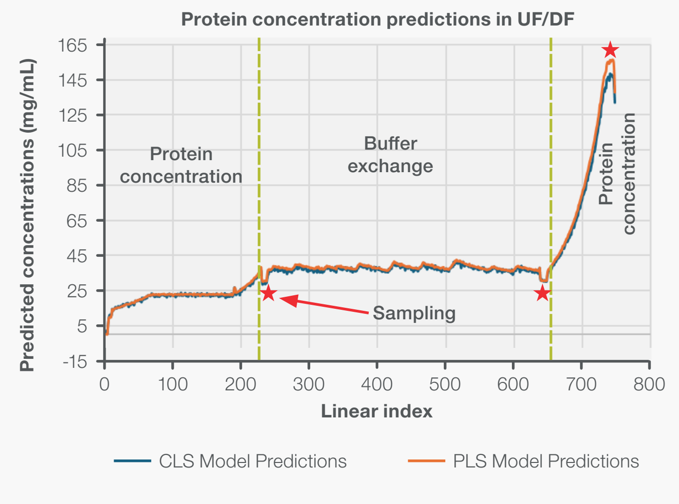


Figure 5. This plot shows the agreement in prediction of protein concentration from the PLS (orange) and CLS (blue) models for the UF/DF run.

Conclusion

This study demonstrated an alternative approach of a single-spectrum-based CLS model for protein quantification in downstream bioprocesses. The CLS model exhibited lower prediction errors than the broadly acceptable tolerance of < 5-10 % for process monitoring. It also showed that the single-spectrum-based CLS model has scalability from lab to pilot scale and transferability across different monoclonal antibody and buffer matrices.

A key factor in the successful implementation of the CLS model was appropriate region selection. We found that the unique Raman signature of the amide I region of monoclonal antibodies has minimal spectral interference from other constituents commonly used in downstream processes.[3,5] This allowed us to develop a mAb-specific CLS model using the Raman intensity of the amide I region that linearly scales with the concentration. Identifying similar unique regions in other applications can provide a rapid and straightforward method to build robust and accurate chemometric models. In addition, augmenting more data to the CLS model especially at the upper and lower concentration range will further improve the model.

Another important aspect involves normalizing the spectra using the O-H symmetric stretching Raman band of water molecules. Recent literature has also utilized a similar normalization strategy for developing chemometric models for upstream bioreactor monitoring.[4] This strategy appears to work across all modalities.

Finally, the ability to build the chemometric model using a minimal dataset not only facilitates the adoption of Raman technology but also broadens its applicability.

References:

1. Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra. *Appl Spectrosc* **1989**, *43* (5), 772–777. https://doi.org/10.1366/0003702894202201.
2. *Advanced Preprocessing: Sample Normalization - Eigenvector Research Documentation Wiki.* https://www.wiki.eigenvector.com/index.php?title=Advanced_Preprocessing:_Sample_Normalization.
3. Nolasco, M.; Pleitt, K.; Khadka, N. Using a Process Raman Analyzer as an In-Line Tool for Accurate Protein Quantification in Downstream Processes.
4. Pétillot, L.; Pewny, F.; Wolf, M.; Sanchez, C.; Thomas, F.; Sarrazin, J.; Fauland, K.; Katinger, H.; Javalet, C.; Bonneville, C. Calibration Transfer for Bioprocess Raman Monitoring Using Kennard Stone Piecewise Direct Standardization and Multivariate Algorithms. *Engineering Reports* **2020**, *2* (11), e12230. https://doi.org/10.1002/eng2.12230.
5. Nolasco, M.; Pleitt, K.; Khadka, N. Raman-Based Accurate Protein Quantification in a Matrix That Interferes with UV-Vis Measurement.
6. Palacký, J.; Mojzeš, P.; Bok, J. SVD-Based Method for Intensity Normalization, Background Correction and Solvent Subtraction in Raman Spectroscopy Exploiting the Properties of Water Stretching Vibrations. *Journal of Raman Spectroscopy* **2011**, *42* (7), 1528–1539. https://doi.org/10.1002/jrs.2896.
7. Peters, J.; Park, E.; Kalyanaraman, R.; Luczak, A.; Ganesh, V. Protein Secondary Structure Determination Using Drop Coat Deposition Confocal Raman Spectroscopy. **2016**, *31*, 31–39.
8. Lackey, H. E.; Sell, R. L.; Nelson, G. L.; Bryan, T. A.; Lines, A. M.; Bryan, S. A. Practical Guide to Chemometric Analysis of Optical Spectroscopic Data. *J. Chem. Educ.* **2023**, *100* (7), 2608–2626. https://doi.org/10.1021/acs.jchemed.2c01112.

Learn more at **thermofisher.com/marqmetrix**

For Research Use Only. Not for use in diagnostic procedures.