Machine learning-assisted Raman spectroscopy for pH and lactate sensing in body fluids

I.Olaetxea,[†,‡] A.Valero,[†] E.Lopez,[†] H.Lafuente,[¶] A.Izeta,[¶] I.Jaunarena,[§,‖] and A.Seifert[*,†,⊥]

†*Nanoengineering Group, CIC nanoGUNE BRTA, 20018 San Sebastián, Spain*

‡*PhD Student, Dept. of Communications Engineering, University of the Basque Country (UPV/EHU)*

¶*Tissue Engineering, Biodonostia Health Research Institute, 20014 San Sebastián, Spain*

§*Obstetrics and Gynaecology, Biodonostia Health Research Institute, 20014 San Sebastián, Spain*

‖*Donostia University Hospital, 20014 San Sebastián, Spain*

⊥*IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain*

E-mail: a.seifert@nanogune.eu

Phone: +34 943 574 000

Abstract

This study presents the combination of Raman spectroscopy with machine learning algorithms as a prospective diagnostic tool capable of detecting and monitoring relevant variations of pH and lactate as recognized biomarkers of several pathologies. The applicability of the method proposed here is tested both in vitro and ex vivo. In a first step, Raman spectra of aqueous solutions are evaluated for the identification of

characteristic patterns resulting from changes in pH or in the concentration of lactate. The method is further validated with blood and plasma samples. Principal Component Analysis is used to highlight the relevant features that differentiate the Raman spectra regarding their pH and concentration of lactate. Partial Least Squares regression models are developed to capture and model the spectral variability of the Raman spectra. The performance of these predictive regression models is demonstrated by clinically accurate predictions of pH and lactate from unknown samples in the physiologically relevant range. These results prove the potential of our method to develop a non-invasive technology, based on Raman spectroscopy, for continuous monitoring of pH and lactate in vivo.

Introduction

Quantification of pH and lactate is a common clinical procedure for evaluating the severity of diseases and pathologies, such as sepsis, hypoxia, fatigue, shock and cardiac arrest, and assessing appropriate therapeutic measures. Little variations of these physiological parameters might reflect an abnormal behaviour of a metabolic pathway associated with pathological findings.[1,2]

In the glycolysis, glucose is transformed to pyruvate, which in presence of oxygen, enters the mitochondria and is transformed into adenosine triphosphate (ATP), the energy-carrying molecule. The consumption of ATP by the cells generates hydrogen ions which are mainly consumed in the aerobic pathway, specifically in the oxidative phosphorylation process. When oxygen is not available, pyruvate is converted to lactate. At normal conditions, most of the energy comes from the aerobic pathway, but to some extent lactate is also produced, meaning that the anaerobic pathway is also active.[2,3] A decreased clearance or an overproduction of lactate, regardless of being caused by oxygen deprivation or not, or a combination of both, can generate high lactate levels that are related with morbidity and mortality in many cases. However, the pathogenesis differs among different diseases or

even patients. In this light, it would lead to incalculable risks taking a hasty diagnostic snapshot of elevated lactate, based on a single measurement at one specific moment, without considering the systemic clinical picture.[1,3]

Current clinical diagnostics, as blood gas analysis, voltammetric-based or liquid chromatography detection methods, require arterial blood sampling, which can only be done intermittently. Moreover, these techniques require significant time for analysis, are quite expensive, expose healthcare professionals to patients' blood, and result in iatrogenic blood loss.[4] Considering these limitations, an effort has been put on developing continuous pH and lactate monitoring systems to allow for robust decision-making and timely clinical interference in critical care medicine. The capabilities and strengths of biophotonic methods distinctly showcase the importance that spectroscopy gains in medical diagnostics.[5,6] Although some attractive methods based on microwave sensors[7] or near-infrared spectroscopy-based sensors[8] for non-invasive monitoring of pH and lactate have been proposed, so far none of them has shown enough predictive accuracy or reliability in clinical use.

Raman spectroscopy, as a promising and emerging photonic technique for diagnostics, has become of general interest in medical research. The inelastic scattering of the incident photons due to specific vibrational modes of individual molecules provides the chemical fingerprint of the sample being studied, hence, delivering highly specific and diagnostically valuable information. The great ability to non-invasively provide high molecular selectivity makes Raman spectroscopy an ideal method for chemical quantification in medical diagnostics.[9] Moreover, Raman instrumentation is sufficiently mature for studying the human organism and physiological parameters. In fact, several studies have already demonstrated that Raman spectroscopy could be used for rapid biomedical analysis ex vivo as well as for continuous monitoring in vivo.[10–12] The interpretation of Raman spectra is far from being straightforward. Due to the generally weak Raman signal and high fluorescence of biomolecules by nature, as well as the overlapping of spectral features from different constituents present in a sample, peak identification and accurate prediction become a serious

challenge. Nevertheless, the combination of Raman spectroscopy with a proper sequence of data preprocessing methods and machine learning algorithms results in a very powerful technology for unveiling hidden features. Machine learning algorithms are characterized by their capacity to identify relevant features and extract valuable knowledge from data that allows them to learn how to handle new situations. The potential of such algorithms for quantification of complex Raman spectroscopy signals has been amply demonstrated. [13] However, although such a combination has already been used to analyze pH and lactate for different purposes, such as food control [14] or sensing of aqueous solutions, [15,16] the lack of evidence to achieve clinical needs still generates doubts about its utility to quantify complex media.

In the present study, we consider the applicability of Raman spectroscopy together with machine learning for quantitative and qualitative analysis of both pH and lactate in complex body fluids, as blood and plasma. This approach represents a significant step towards diagnosis of related pathologies, as for example hypoxia,[17] where lactate $\geq 4.8\,\mathrm{mM}$ and pH ≤ 7.20 are defined as standards for intervention, or in case of sepsis,[18] where lactate $\geq 2\,\mathrm{mM}$ reflects physiological anomaly.

Methods

Sample preparation and acquisition of Raman spectra

In the current study both in vitro and ex vivo samples were analyzed. First, spectroscopic measurements were carried out with aqueous solutions, such as milliQ water and phosphate-buffered saline (PBS, 10010015, Thermo Fisher Scientific), to demonstrate proof-of-principle. For a total number of 50 samples, pH was varied from 6.80 to 7.60, by adding 1 M sodium hydroxide (NaOH, S5881, Sigma-Aldrich) or hydrochloric acid (HCl, H1758, Sigma-Aldrich) diluted at 10 %. For each sample, three consecutive spectra were collected; each spectrum was formed by accumulation of three measurements with an integration time of 10 s. Similarly, Sodium L-Lactate (71718, Sigma-Aldrich) was used to vary the concentration of lactate,

firstly for the range of 0 - 30 mM (70 samples) and secondly for the physiologically relevant range of 0 - 10 mM (60 samples).

In a second step, blood from domestic pigs was analyzed. Blood, stored at 4°C, was provided by Biodonostia Health Research Institute. pH values from 30 samples were studied in the range from 6.97 to 7.53. For lactate, based on the initial concentration in blood, further concentrations were prepared by adding sodium lactate. Due to the higher complexity in sample handling, a slightly broader range than the physiological range has been chosen (2 - 18 mM) and 32 samples prepared. For achieving a better signal-to-noise ratio (SNR) in blood, four consecutive spectra were collected per sample and each spectrum was obtained by accumulating five measurements with an integration time of 10 s.

Additionally to blood samples, the variation of lactate was also studied in plasma samples, as a second example for analyzing complex media. Plasma was extracted from blood samples by centrifugation at 1000 g for 30 minutes and collection of the supernatant. Measurement specifications were maintained except the integration time, which was reduced to 6 s to avoid saturation of the detector.

To perform the Raman measurements, 250 μL of the samples to be analyzed were deposited in a well on a fused silica substrate of a customized Raman spectroscopy system in inverted microscope configuration, consisting of: a continuous wave diode laser, emitting at 785 nm with an emission power of 130 mW, as Raman excitation source; and a grating spectrometer (EAGLE Raman-S with Andor iVac 316 detector, Ibsen Photonics), with a spectral resolution of 4 cm^{-1}, to collect the backscattered Raman light. As described in Figure S2, liquid samples were excited from below such that the effect of evaporation was minimized and unstable focusing and excitation was avoided. Measurements were performed at room temperature (23°C). All experimental procedures were conducted in strict compliance with European and Spanish regulations on the protection of animals used for scientific purposes (European Directive 2010/63/EU and Spanish Royal Legislative Decree 53/2013).

Pre-processing of Raman spectra

Analysis of Raman data in biological samples is extremely challenging due to their heterogeneous nature. Unwanted contributions caused by different background and noise sources, might hide weak Raman signals from the sample under test. Therefore, prior to develop any predictive model, preprocessing of raw Raman data is essential for the correct interpretation of the Raman spectra and for a reliable quantification of the samples. Hence, all Raman spectra were first trimmed to the range from 300 cm^{-1} to 1700 cm^{-1} and preprocessed in MATLAB (MathWorks®). An open toolbox[19] was used to apply a 6th order Extended Multiplicative Signal Correction (EMSC) to the raw spectra. Interfering additive and multiplicative artifacts were removed by scaling all the Raman spectra to the mean spectrum.[20] The median of consecutive spectra from each sample was taken to eliminate spike artifacts sporadically produced by cosmic rays. All Raman spectra were smoothed using a Savitzky-Golay filter with a window of 15 points and 3rd order polynomial. Finally, the asymmetric least squares (ALS) method was used to subtract a smoothed background produced by the intrinsic fluorescence of the molecules.[21] The integrity of the Raman bands is retained by giving higher relevance to positive residuals.[22]

Multivariate analysis

Simultaneous analysis and combination of different variables enables the development of efficient predictive models. However, the correct extraction of significant features within thousands of misleading, correlated and redundant variables to be handled is a major challenge.[22] The high number of variables, or pixels in this case, obtained by Raman spectroscopy might lead to overfitting and consequent incapacity of the model to generalize. Machine learning is used to reliably identify and model the spectral variability from preprocessed spectra.[13,22] Partial Least Squares (PLS) and Principal Component Analysis (PCA) are both dimensionality-reduction techniques used in machine learning. PCA was used to reveal any pattern and facilitate visualization and interpretation of the spectral data. The discrimination of

spectra as a function of their pH or concentration of lactate was evaluated by projecting the data onto the most relevant principal components.[22] Moreover, Partial Least Squares (PLS) was used to develop predictive regression models that define the correlation between the spectral variability and the pH and the concentration of lactate of the samples.[23,24] To evaluate the performance of predictive models, such as those resulting from PLS, machine learning algorithms consist of two stages. First, the prediction model is built and calibrated with a *training set* of measurement data and validated with a new *testing set* of unknown samples afterwards. For model calibration, leave-one-out cross-validation (LOOCV) method was used to determine an optimum set of latent vectors, and hence, avoiding overfitting. The root mean square error of prediction (RMSEP) and the coefficient of determination R^2 were used as indicators of the quality of the predictive model. Whereas RMSEP indicates the standard deviation of the predictions from the observed values, R^2 reflects the amount of variance in the dependent variable that is explained by the model, or in other words, how well is the data fitting the model (the goodness of fit). All the analysis was performed in MATLAB environment.

Results and discussion

To exclude secondary effects of complex media, our method was firstly validated with a set of aqueous samples. A protocol was developed for the determination of physiological pH and lactate, considering the entire process chain from sample preparation over Raman measurements to data evaluation. Following the protocol, the variation of pH and lactate in blood and plasma samples from domestic pigs was investigated.

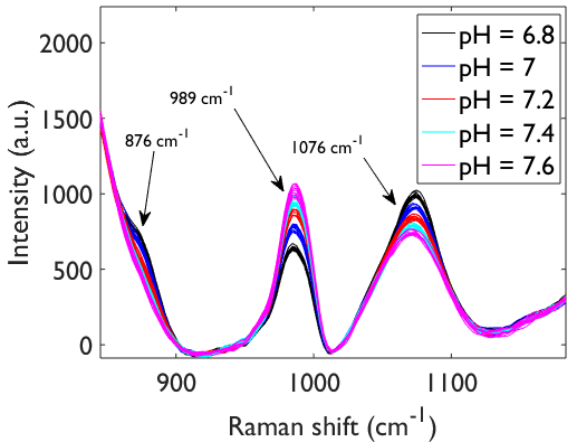

Figure 1: Preprocessed Raman spectra of PBS solutions with different pH values from 6.8 to 7.6. Although the spectral range from 500 to 1500 cm^{-1} is studied, only the range from 850 to 1180 cm^{-1} is shown for better visualization.

pH and lactate monitoring in aqueous solutions

pH analysis

Due to purity of milliQ water and its subsequent low buffering capacity, PBS has been used to avoid unwanted additional pH variations when exposed to air. As explained previously, Raman spectroscopy provides specific information about the vibrational modes of the molecules. Considering that pH is a metric to determine the acidity of a solution by measuring the concentration of hydrogen ions, the variation of pH, unlike lactate, is not directly quantifiable by Raman spectroscopy. However, the variation of pH can indeed induce changes either in the structure, the concentration or the chemical bonds of the molecules in the sample being measured, which are eventually detectable by Raman spectroscopy and can be quantified and correlated with pH. In the case of PBS, which is a buffered solution, pH is expected to provoke changes to the salts included. In Figure 1, Raman bands at 876 cm^{-1} and 1076 cm^{-1}, and the peak at 989 cm^{-1}, which have previously been related to H_2PO_4 and HPO_4 by Fontana et al.,[25] have shown to be pH-dependent. These spectral variations are highlighted by applying PCA. Figure 2a clearly separates the different samples as a function of their pH. The loading plot in Figure 2b substantiates that most of the spectral variation

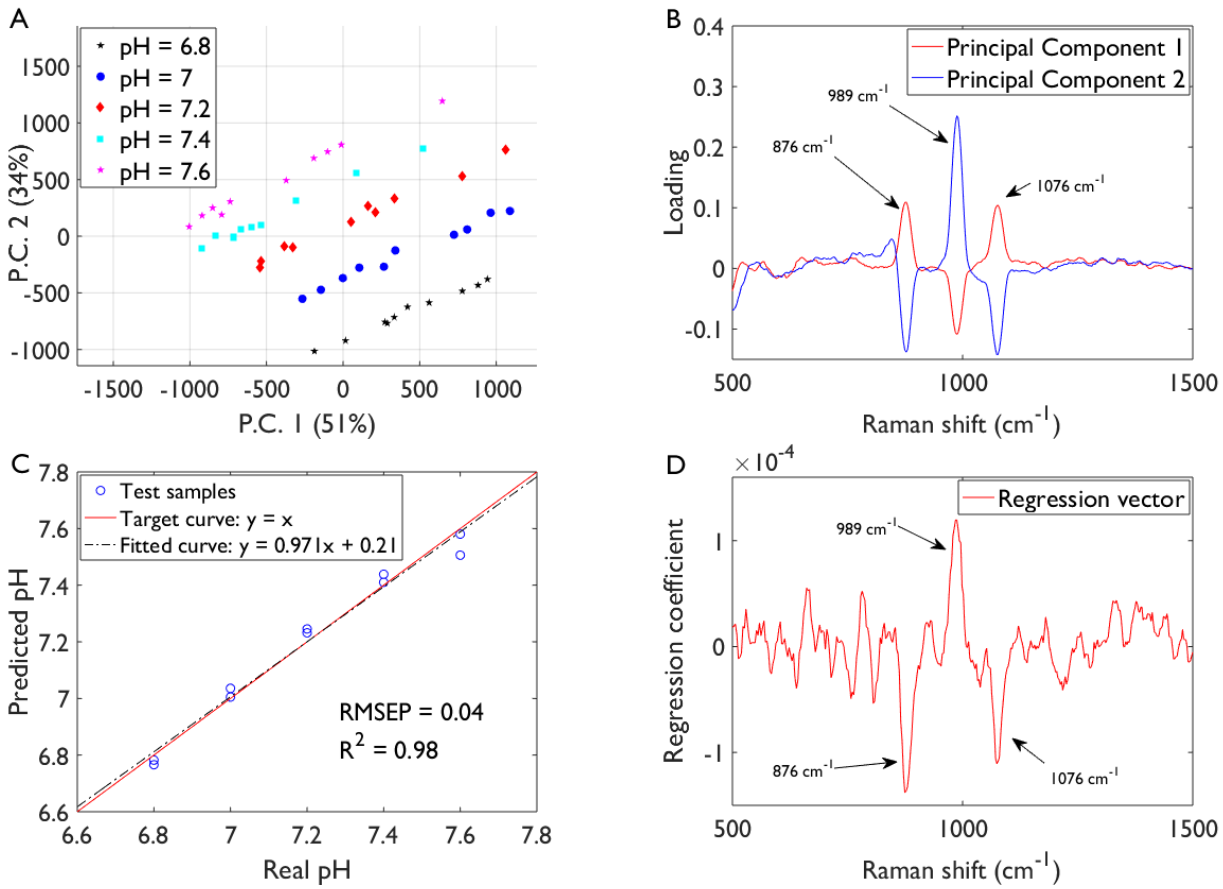

Figure 2: Multivariate analysis of preprocessed spectra from PBS solutions with varying values of pH in the physiological range of 6.8 - 7.6. **(A)** PCA score plot (PC-1 vs. PC-2), **(B)** PCA loading plot, **(C)** prediction of the PLS regression model with an optimum number of latent vectors of $N = 3$; the values of RMSEP and R^2 are characteristic of the model and are calculated with respect to the target curve, **(D)** and regression coefficient plot of the PLS regression model.

captured by the first two components (85 % of the variance) is indeed caused by the effect of pH in the vibrational modes of both H_2PO_4 and HPO_4.

A predictive PLS regression model has been developed for the prediction of pH in PBS samples. The total of 50 samples were divided into two sets, 40 training samples for calibration and 10 testing samples. In Figure 2c, the small RMSEP of 0.04 and R^2 of 0.98 obtained by the PLS predictive model for unknown samples demonstrates the quality of the model and the feasibility of the method to detect pH variations in aqueous solutions. Figure 2d shows the main contribution of H_2PO_4 and HPO_4 in the predictive model.

Lactate analysis

The fingerprint region of pure sodium-lactate powder is shown in Figure S1. Some of the characteristic Raman peaks of lactate, previously reported,[26,27] are indicated: a sharp predominant peak at 853 cm^{-1} associated with a C-C single bond vibration, at 543 cm^{-1} from CO_2 wagging, at 1040 cm^{-1} and 1053 cm^{-1} due to C-CH_3 stretching vibrations, at 1081 cm^{-1} caused by C–O vibrations, and at 1459 cm^{-1} from asymmetric CH_3 deformation modes. Some other visible Raman bands in Figure S1 have not been addressed, but are characteristic of sodium-lactate.

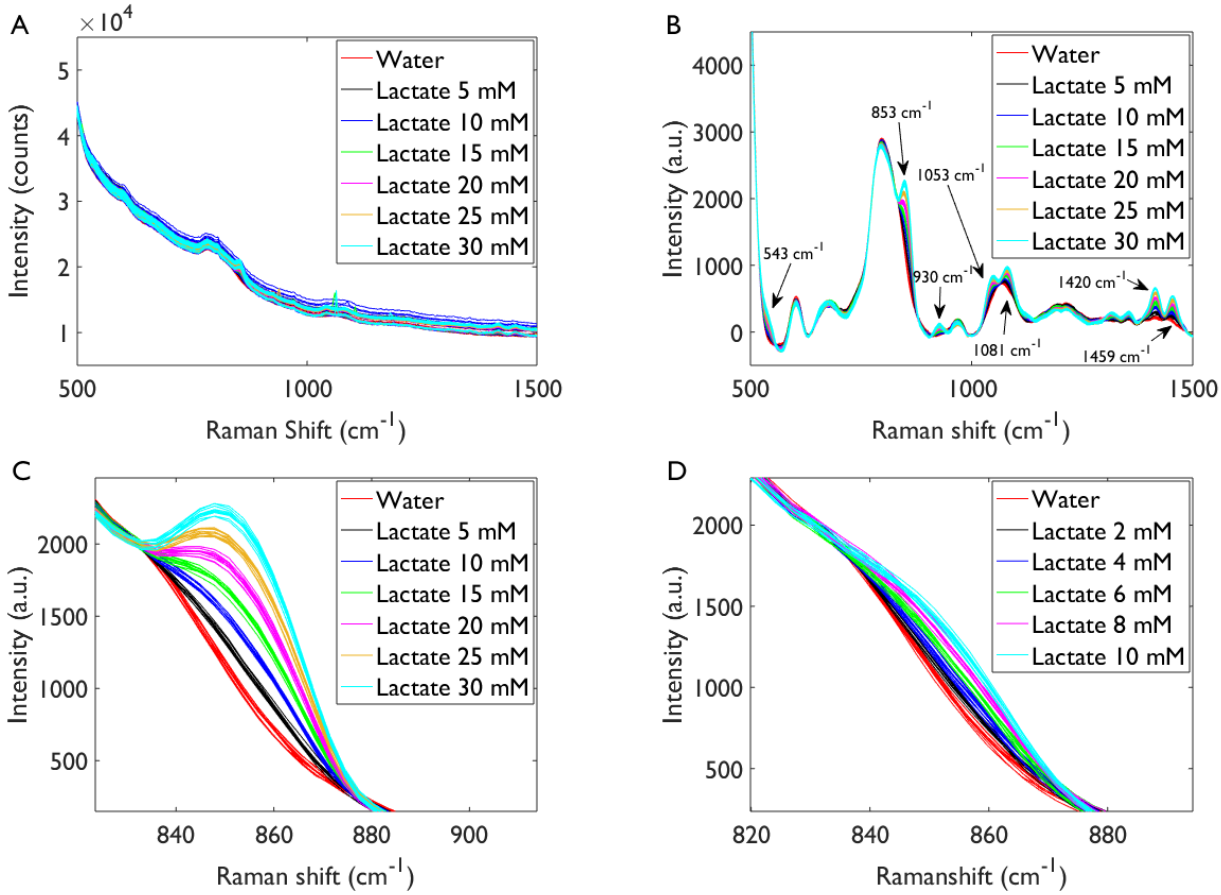

Figure 3: Aqueous solutions with different concentrations of sodium lactate. **(A)** Raw Raman spectra in the region from 500 cm^{-1} to 1500 cm^{-1} with varying concentrations of lactate in the range of 0 - 30 mM, **(B)** and its corresponding preprocessed spectra. **(D)** Peak at 853 cm^{-1}, associated with a C-C single bond vibration, for different concentrations of lactate in the range of 0 - 30 mM, **(D)** and in the physiological range of 0 - 10 mM.

As mentioned before, the first step was to identify specific Raman features of lactate in

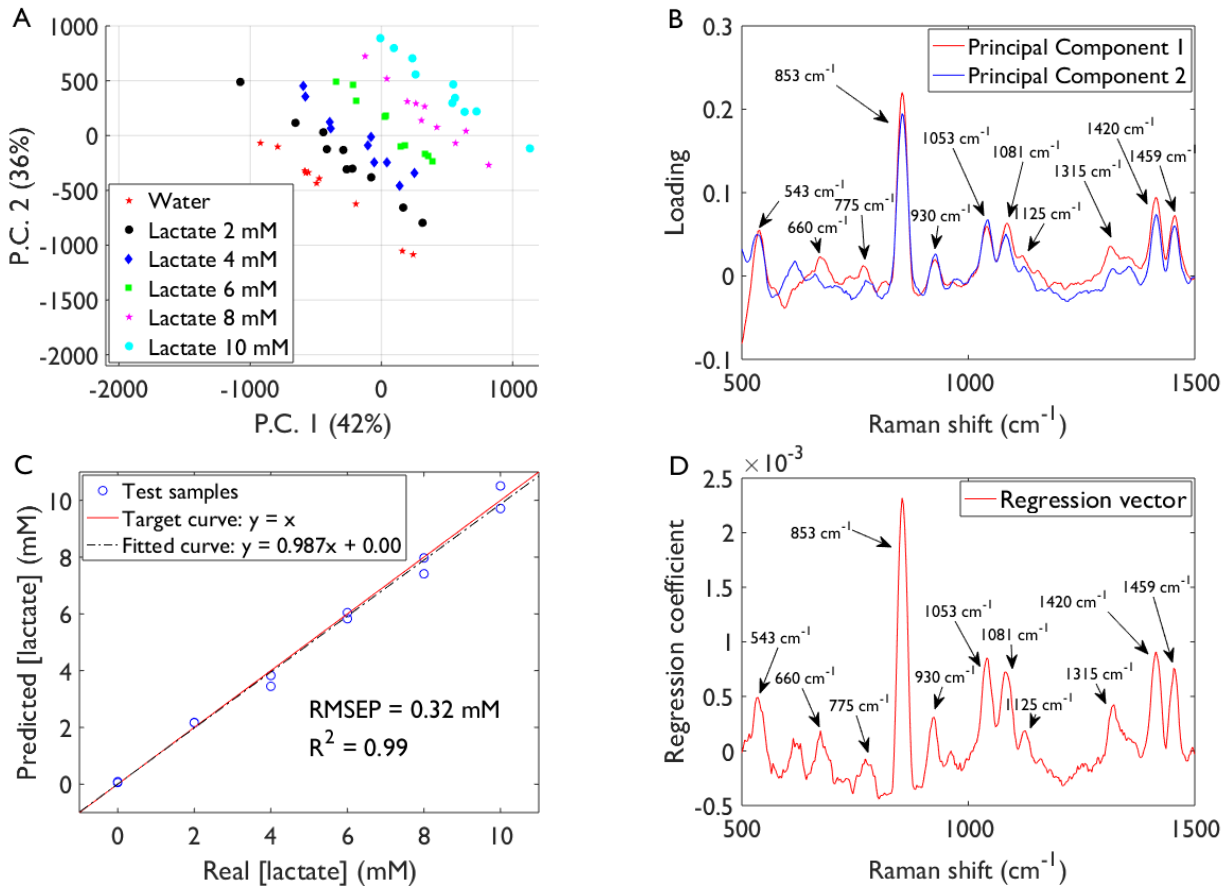

Figure 4: Multivariate analysis of preprocessed spectra from aqueous solutions with varying concentrations of lactate in the physiological range of 0 - 10 mM. **(A)** PCA score plot (PC-1 vs. PC-2), **(B)** PCA loading plot, **(C)** prediction of the PLS regression model with an optimum number of latent vectors of $N = 3$; RMSEP and R^2 defined as in Figure 2, **(D)** and regression coefficient plot of the PLS regression model.

milliQ water solutions. Figure 3a and Figure 3b present the raw data and its corresponding preprocessed spectra from the set of samples with varying concentrations of lactate in a range between 0 and 30 mM. In raw data, it is extremely difficult to recognize any pattern. However, a completely different picture emerges after applying the described preprocessing, which demonstrates the importance of this step. Several regions of the spectra undergo fluctuations in the intensity of the Raman features. In Figure 3b, Raman bands from pure sodium lactate as well as some additionally reported peaks,[28] such as 930 cm^{-1} or 1125 cm^{-1}, are observed. Figure 3c is a magnification of the predominant peak at 853 cm^{-1}, which clearly shows how the intensity of the peak varies in correlation with the concentration of

lactate; the higher the concentration of lactate, the higher the intensity of the peak. When the concentration of lactate is reduced to the common physiological range between 0 and 10 mM, lactate Raman features are still evident despite the little differences between the spectra, as shown in Figure 3d. The great advantage of using multivariate analysis is the possibility to take into account many features of the spectra instead of single peaks, thereby unveiling hidden information. Moreover, multivariate analysis omits redundant information so that reliable and accurate predictive models are developed.

To have a better visualization of the spectra, preprocessed spectra are projected onto the first two principal components, that explain 78 % of the spectral variation (Figure 4a). The loading plot of PC-1 and PC-2 unveils the influence of lactate in the spectral variation captured by these components (Figure 4b). Spectral features associated with lactate constitute a dominant part. Even Raman peaks, such as 660 cm^{-1}, 775 cm^{-1}, 1125 cm^{-1} or 1315 cm^{-1}, that were not perceptible in Figure 3b by eye, are now gathered by PCA.

A predictive PLS regression model has been developed based on preprocessed spectra. A total of 60 samples were divided into two sets, 48 training samples for calibration and 12 testing samples. The model was calibrated using the leave-one-out cross-validation (RMSECV) method and the optimum number of latent vectors was determined by Wold's criterion.[29] The predictive model has been successfully tested with unknown samples providing a very small RMSEP of 0.32 mM and R^2 of 0.99, which corroborates the quality of the model. Figure 4c compares the predicted and real values of lactate concentration. To understand how strongly each Raman band determines the model, the regression coefficients of the calibrated PLS model are shown in Figure 4d. The major contributing Raman features are associated with lactate.

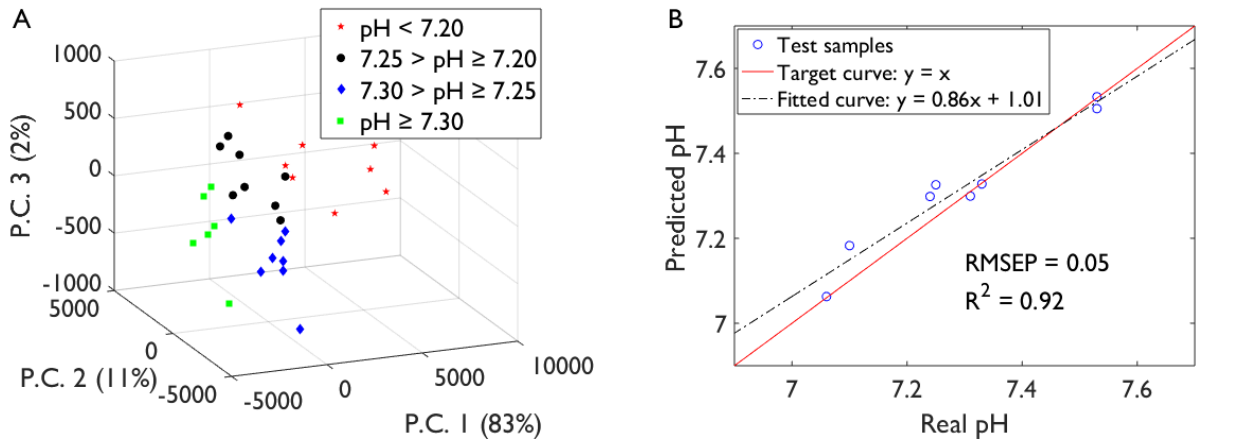

Figure 5: Multivariate analysis of preprocessed spectra from pig blood with varying values of pH in the range of 6.97 - 7.53. **(A)** PCA score plot for the first three principal components, **(B)** prediction of the PLS regression model with an optimum number of latent vectors of $N = 6$; RMSEP and R^2 defined as in Figure 2.

pH and lactate monitoring in blood

pH analysis

To further study the efficacy of our method, Raman spectra from blood samples from domestic pigs, as representatives for real physiological samples, have been analyzed at different pH values (Figure S3). As mentioned before, the objective is to identify and indirectly quantify variations of pH by detecting Raman bands from molecules that are sensitive to pH. Vibrational modes of hemoglobin have actually been recognized as the main responsible ones for the Raman spectrum of blood.[30]

In Figure 5a, the projection of spectra onto the first three principal components shows a clear clustering of the spectra with respect to their pH value. The loading plot in Figure S4a unveils some spectral variation in Raman bands, such as 420 cm^{-1}, 567 cm^{-1}, 754 cm^{-1}, 789 cm^{-1}, 1002 cm^{-1}, 1207 cm^{-1}, 1222 cm^{-1}, 1357 cm^{-1}, 1547 cm^{-1} and 1639 cm^{-1}, which have already been assigned to hemoglobin.[31] It has been reported that hemoglobin also exhibits characteristic active Raman bands in the region from 1300 cm^{-1} to 1650 cm^{-1}, however, it is interesting to notice that many of these reported peaks could not be precisely identified in Figure S4a. Nevertheless, the high variability of the spectra in this region indicates pH-

induced modifications in the vibrational modes of hemoglobin.

A PLS model has been implemented for the prediction of pH in blood from preprocessed spectra. The model was calibrated with 22 samples for future validation with 8 new samples. Figure 5b displays the excellent predictive ability of the model for unknown samples, which is reflected in the small RMSEP of 0.05 and the good R^2 of 0.92, when the reference and the predicted values of pH are compared. Each Raman band associated with hemoglobin is affected differently by the variation of pH, and hence, contributes individually to the model. For that reason, it is not straightforward to identify exact positions of the characteristic Raman peaks of hemoglobin in the regression vector in Figure S4b. Nonetheless, stronger contributions from regions around 420 cm^{-1}, 1000 cm^{-1}, 1200 cm^{-1} and 1600 cm^{-1} are observed.

Lactate analysis

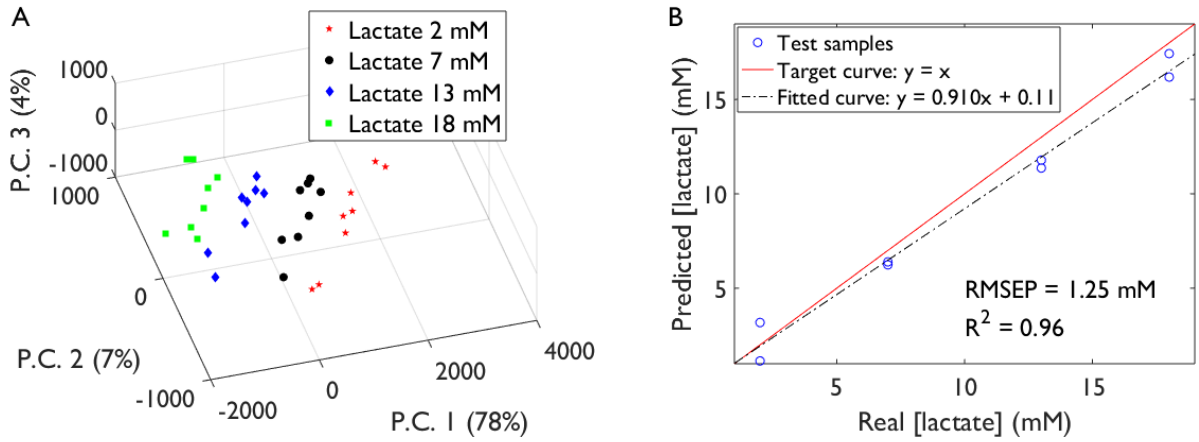

Figure 6: Multivariate analysis of preprocessed spectra from pig blood with varying concentrations of lactate in the range of 2 - 18 mM. **(A)** PCA score plot for the first three principal components, **(B)** prediction of the PLS regression model with an optimum number of latent vectors of $N = 4$; RMSEP and R^2 defined as in Figure 2.

Unlike in the case of pH, lactate is quantified by detecting characteristic Raman peaks in the preprocessed spectra from blood, which are displayed in Figure S5. The discriminating capability of PCA is demonstrated in Figure 6a, where the projection of preprocessed spectra in the first three principal components allows for visual separation of the samples into distinct groups. Looking at the loading plot in Figure S6a, the spectral variation captured by the

first two components (85 % of the variance) reveals some of the characteristic features of lactate, but also many other bands that are not related to it.

For developing a PLS regression model, preprocessed spectra of 32 samples have been used, 24 for training and 8 for testing. The prediction, when tested with unknown samples, demonstrates again the quality of the model with an RMSEP of 1.25 mM and $R^2 = 0.96$, as shown in Figure 6b. Two main messages can be drawn from the regression coefficients of the predictive model in Figure S6b: first, most of the spectral variation present in the data is in accordance with different concentrations of lactate, since most of its characteristic Raman bands are identified; and second, this spectral variation is not only due to lactate since other Raman bands, which are not associated with lactate, are also contributing to the model.

To study where these contributions are coming from, a gold standard method, based on electrochemical analysis, has been used to take a reference measurement of the main parameters commonly measured in clinical settings (i-STAT, Abbott Technologies). The values of four blood samples with different concentrations of lactate at two time intervals are displayed in Table 1. As expected, the concentration of lactate changes from sample to sample, but it can also be observed that many of these parameters are changing with time, meaning that sample preparation and therefore exposure of the samples to air is extremely determinant. Furthermore, Lemler et al.[32] already mentioned that hemoglobin experiences some chemical changes as soon as blood is extracted. Considering the fact that sample preparation was performed by groups at different times, the variation of the different parameters shown in Table 1 could explain the unexpected contributions present in the model (Figure 6b, Figure S6b). Moreover, it has been demonstrated that prolonged exposure to laser and high incident power are responsible for hemoglobin denaturation and heme aggregates.[32] This unwanted effect might hinder possible correlations induced by the addition of lactate.

Table 1: Clinical parameters, measured with i-STAT (Abbott Technologies), from pig blood with varying concentrations of lactate at two different periods of time.

Clinical parameters	Sample 1		Sample 2		Sample 3		Sample 4	
	t1	*t2*	*t1*	*t2*	*t1*	*t2*	*t1*	*t2*
pH	6.92	7.02	6.99	7.06	6.98	7.08	6.99	7.06
pCO_2 (mmHg)	125	95	103	84	101	76	95	79
pO_2 (mmHg)	175	186	178	166	150	157	133	145
BE (mM)	-7	-6	-7	-7	-8	-7	-8	-8
HCO_2 (mM)	25.6	24.7	24.4	23.5	23.8	22.6	23.0	22.7
TCO_2	29	27	27	26	27	25	26	25
sO_2 (%)	98	99	98	98	97	98	96	98
Lactate (mM)	1.78	2.07	7.51	7.68	13.08	13.42	18.67	18.98

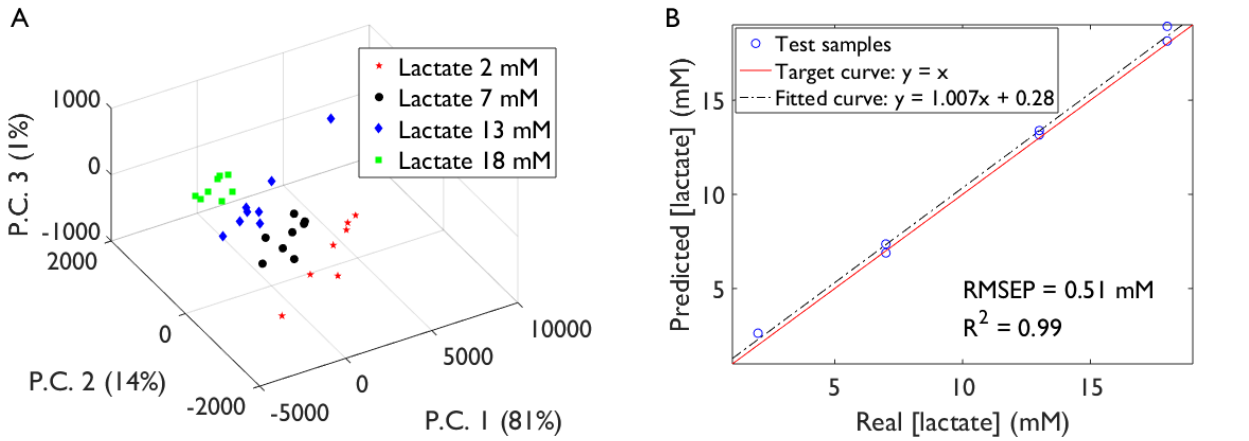

Figure 7: Multivariate analysis of preprocessed spectra from pig blood plasma with varying concentrations of lactate in the range of 2 - 18 mM. **(A)** PCA score plot for the first three principal components, **(B)** prediction of the PLS regression model with an optimum number of latent vectors of $N = 5$; RMSEP and R^2 defined as in Figure 2.

Lactate monitoring in blood plasma

To minimize the effect induced by sample preparation, blood cells had been removed and the experiment was repeated in plasma samples. Visual inspection of PCA reveals that plasma spectra still form clusters according to their concentration of lactate (Figure 7a) and that lactate Raman features are still captured by the first two components, which is reflected in the loading plot in Figure S7a.

However, labelling the data and developing a PLS regression model, a more robust and accurate model is obtained. In comparison with blood samples, the prediction error has been reduced to a value of RMSEP = 0.51 mM (Figure 7b), which demonstrates a much stronger contribution of lactate (Figure S7b).

Evaluation of predictive models

PLS regression models have been implemented and used for evaluating the predictive power of pH and lactate in aqueous solutions and body fluids by Raman spectroscopy. Quantification errors are summarized in Table 2. It has been shown that data preprocessing plays a crucial role in the development of models with good performance. The capability of our method for sensing lactate and pH variations has been demonstrated by this study. Moreover, it has been shown that blood, especially hemoglobin, is extremely sensitive to measurement conditions, such as exposure to air or the power of the laser, which introduce uncontrolled additional variability to the predictive model that should be considered.

Table 2: Prediction errors of pH and lactate in aqueous solutions and body fluids.

	PBS/Water	Blood	Plasma
pH	0.04	0.05	-
Lactate (mM)	0.32	1.25	0.51

Conclusions

The present study has demonstrated that the combination of Raman spectroscopy with machine learning represents a suitable tool for monitoring lactate and pH values from body fluids. PCA has proved clear discrimination of the spectra as a function of the pH value and concentration of lactate. For future clinical applications, this constitutes a powerful tool for both binary and multiclass classification in a variety of fields, such as fatigue analysis, sepsis classification or identification of a hypoxia state. In addition, PLS predictive models have been successfully tested with unknown samples, providing clinically promising errors that verify the reliability, the accuracy and hence, the quality of our models. In conclusion, we have verified a method for the development of a new, non-invasive and real-time technology that could be used for continuous monitoring in vivo of pH and lactate related clinical diagnostics.

Acknowledgement

This work was supported by the Spanish Ministry of Economy, Industry and Competitiveness under the Maria de Maeztu Units of Excellence Programme - MDM-2016-0618; further financial support by the Basque Government, Department of Health, in the program R&D Projects for Health 2019 - Ref.Nr. 2019222008, and by the Basque Government, Department of Education, in the Predoctoral Education Program.

References

(1) Andersen, L. W.; Mackenhauer, J.; Roberts, J. C.; Berg, K. M.; Cocchi, M. N.; Donnino, M. W. Etiology and therapeutic approach to elevated lactate levels. Mayo Clinic Proceedings. 2013; pp 1127–1140.

(2) Suetrong, B.; Walley, K. R. Lactic acidosis in sepsis: it's not all anaerobic: implications for diagnosis and management. *Chest* 2016, *149*, 252–261.

(3) Bakker, J.; Nijsten, M. W.; Jansen, T. C. Clinical use of lactate monitoring in critically ill patients. *Annals of intensive care* 2013, *3*, 12.

(4) Ube, T.; Yoneyama, Y.; Ishiguro, T. In situ Measurement of the pH-dependent Transmission Infrared Spectra of Aqueous Lactic Acid Solutions. *Analytical Sciences* 2017, *33*, 1395–1400.

(5) Marcu, L.; Boppart, S. A.; Hutchinson, M. R.; Popp, J.; Wilson, B. C. Biophotonics: the big picture. *Journal of biomedical optics* 2017, *23*, 021103.

(6) Jerjes, W. K.; Upile, T.; Wong, B. J.; Betz, C. S.; Sterenborg, H. J.; Witjes, M. J.; Berg, K.; Van Veen, R.; Biel, M. A.; El-Naggar, A. K., et al. The future of medical diagnostics. *Head & Neck Oncology* 2011, *3*, 38.

(7) Mason, A.; Korostynska, O.; Louis, J.; Cordova-Lopez, L. E.; Abdullah, B.; Greene, J.; Connell, R.; Hopkins, J. Noninvasive In-Situ Measurement of Blood Lactate Using Microwave Sensors. *IEEE Transactions on Biomedical Engineering* 2018, *65*, 698–705.

(8) Ellerby, G.; Smith, C.; Zou, F.; Scott, P.; Soller, B. Validation of a spectroscopic sensor for the continuous, noninvasive measurement of muscle oxygen saturation and pH. *Physiological measurement* 2013, *34*, 859.

(9) Butler, H. J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N. J.; Gardner, B.; Martin-Hirsch, P. L., et al. Using Raman spectroscopy to characterize biological materials. *Nature protocols* 2016, *11*, 664.

(10) Ashok, P. C.; Giardini, M. E.; Dholakia, K.; Sibbett, W. A Raman spectroscopy biosensor for tissue discrimination in surgical robotics. *Journal of biophotonics* 2014, *7*, 103–109.

(11) Khan, S.; Ullah, R.; Shahzad, S.; Anbreen, N.; Bilal, M.; Khan, A. Analysis of tuberculosis disease through Raman spectroscopy and machine learning. *Photodiagnosis and Photodynamic Therapy* 2018, *24*, 286–291.

(12) Pandey, R.; Paidi, S. K.; Valdez, T. A.; Zhang, C.; Spegazzini, N.; Dasari, R. R.; Barman, I. Noninvasive monitoring of blood glucose with raman spectroscopy. *Accounts of chemical research* 2017, *50*, 264–272.

(13) Lussier, F.; Thibault, V.; Charron, B.; Wallace, G. Q.; Masson, J.-F. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC Trends in Analytical Chemistry* 2020, *124*, 115796.

(14) Nache, M.; Scheier, R.; Schmidt, H.; Hitzmann, B. Non-invasive lactate-and pH-monitoring in porcine meat using Raman spectroscopy and chemometrics. *Chemometrics and Intelligent Laboratory Systems* 2015, *142*, 197–205.

(15) Ren, M.; Arnold, M. A. Comparison of multivariate calibration models for glucose, urea, and lactate from near-infrared and Raman spectra. *Analytical and bioanalytical chemistry* 2007, *387*, 879–888.

(16) Shah, N. C.; Lyandres, O.; Walsh, J. T.; Glucksberg, M. R.; Van Duyne, R. P. Lactate and sequential lactate- glucose sensing using surface-enhanced Raman spectroscopy. *Analytical chemistry* 2007, *79*, 6927–6932.

(17) Cummins, G.; Kremer, J.; Bernassau, A.; Brown, A.; Bridle, H.; Schulze, H.; Bachmann, T.; Crichton, M.; Denison, F.; Desmulliez, M. Sensors for Fetal Hypoxia and Metabolic Acidosis: A Review. *Sensors* 2018, *18*, 2648.

(18) Lee, S. M.; An, W. S. New clinical criteria for septic shock: serum lactate level as new emerging vital sign. *Journal of thoracic disease* 2016, *8*, 1388.

(19) Liland, K. H.; Kohler, A.; Afseth, N. K. Model-based pre-processing in Raman spectroscopy of biological samples. *Journal of Raman Spectroscopy* 2016, *47*, 643–650.

(20) Fearn, T. Extended multiplicative scatter correction. *NIR news* 2005, *16*, 3–5.

(21) Eilers, P. H.; Boelens, H. F. Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report* 2005, *1*, 5.

(22) Gautam, R.; Vanga, S.; Ariese, F.; Umapathy, S. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation* 2015, *2*, 1–38.

(23) Tobias, R. D., et al. An introduction to partial least squares regression. Proceedings of the twentieth annual SAS users group international conference. 1995; pp 1250–1257.

(24) Abdi, H. Partial least square regression (PLS regression). *Encyclopedia for research methods for the social sciences* 2003, *6*, 792–795.

(25) Fontana, M. D.; Mabrouk, K. B.; Kauffmann, T. H. Raman spectroscopic sensors for inorganic salts. *Spectroscopic properties of inorganic and organometallic compounds* 2013, *44*, 40–67.

(26) Atkins, C. G.; Buckley, K.; Chen, D.; Schulze, H. G.; Devine, D. V.; Blades, M. W.; Turner, R. F. Raman spectroscopy as a novel tool for monitoring biochemical changes and inter-donor variability in stored red blood cell units. *Analyst* 2016, *141*, 3319–3327.

(27) Valpapuram, I.; Candeloro, P.; Coluccio, M. L.; Parrotta, E. I.; Giugni, A.; Das, G.; Cuda, G.; Di Fabrizio, E.; Perozziello, G. Waveguiding and SERS Simplified Raman Spectroscopy on Biological Samples. *Biosensors* 2019, *9*, 37.

(28) Cassanas, G.; Morssli, M.; Fabregue, E.; Bardet, L. Vibrational spectra of lactic acid and lactates. *Journal of Raman spectroscopy* 1991, *22*, 409–413.

(29) Gómez-Carracedo, M.; Andrade, J.; Rutledge, D.; Faber, N. Selecting the optimum number of partial least squares components for the calibration of attenuated total reflectance-mid-infrared spectra of undesigned kerosene samples. *Analytica chimica acta* 2007, *585*, 253–265.

(30) Doty, K. C.; Lednev, I. K. Differentiation of human blood from animal blood using Raman spectroscopy: A survey of forensically relevant species. *Forensic science international* 2018, *282*, 204–210.

(31) Atkins, C. G.; Buckley, K.; Blades, M. W.; Turner, R. F. Raman spectroscopy of blood and blood components. *Applied spectroscopy* 2017, *71*, 767–793.

(32) Lemler, P.; Premasiri, W.; DelMonaco, A.; Ziegler, L. NIR Raman spectra of whole human blood: effects of laser-induced and in vitro hemoglobin denaturation. *Analytical and bioanalytical chemistry* 2014, *406*, 193–200.