

Screening the low molecular weight fraction of human serum using ATR-IR spectroscopy

Franck Bonnier^{,1}, Guillaume Brachet², Romain Duong¹, Tobiloba Sojinrin³, Renaud Respaud⁴, Nicolas Aubrey^{5,6}, Matthew J. Baker⁶, Hugh J. Byrne³, Igor Chourpa¹*

¹ Université François-Rabelais de Tours, Faculty of Pharmacy, EA 6295 Nanomédicaments et Nanosondes, 31 avenue Monge, 37200 Tours, France.

² Université François Rabelais de Tours, UMR CNRS 7292 Génétique, Immunothérapie, Chimie et Cancer, Faculté de Médecine, 10 Bd Tonnellé, 37032 Tours Cedex

³ FOCAS Research Institute, Dublin Institute of Technology (DIT), Camden Row, Dublin 8, Ireland

⁴ Université François-Rabelais de Tours, UMR 1100, CHRU de Tours, Service de Pharmacie, F-37032 Tours, France

⁵ Université de Tours, UMR1282 Infectiologie et Santé Publique, 37200 Tours, France

⁶ Institut National de la Recherche Agronomique, UMR1282 Infectiologie et Santé Publique, 37380 Nouzilly, France

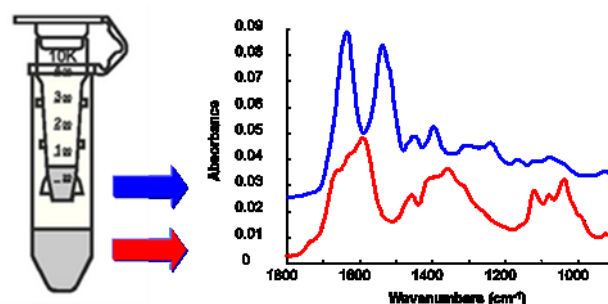
⁷ WestCHEM, Technology and Innovation Centre, Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street, Glasgow G1 1XL, UK

Received zzz, revised zzz, accepted zzz

Published online zzz

Key words: Human serum, centrifugal filtration, IR spectroscopy, Attenuated Total Reflection (ATR), Principal Component Analysis, Low Molecular Weight Fraction (LMWF)

Vibrational spectroscopic techniques can detect small variations in molecular content, linked with disease, showing promise for screening and early diagnosis. Biological fluids, particularly blood serum, are potentially valuable for diagnosis purposes. The so-called Low Molecular Weight Fraction (LMWF) contains the associated peptidome and metabolome and has been identified as potentially the most relevant molecular population for disease-associated biomarker research. Although vibrational spectroscopy can deliver a specific chemical fingerprint of the samples, the High Molecular Weight Fraction (HMWF), composed of the most abundant serum proteins, strongly dominates the response and ultimately makes the detection of minor spectral variations a challenging task. Spectroscopic detection of potential serum biomarkers present at relatively low concentrations can be improved using pre-analytical depletion of the HMWF. In the present study, human serum fractionation by centrifugal filtration was used prior to analysis by Attenuated Total Reflection infrared spectroscopy. Using a model sample based on glycine spiked serum, it is demonstrated that the screening of the LMWF can be applied to quantify blinded concentrations up to 50 times lower. Moreover, the approach is easily transferable to different bodily fluids which would support the development of more efficient and suitable clinical protocols exploring vibrational spectroscopy based *ex-vivo* diagnostic tools.



Revealing serum LMWF for spectral serological diagnostic applications

1. Introduction

Vibrational spectroscopic techniques have been extensively evaluated for their capacity to deliver a specific molecular fingerprint closely linked to the disease state [1–5]. Ranging from tissue to single cell analysis, both infrared and Raman spectroscopy are promising techniques with potential clinical applications [6]. Spectral imaging of biopsies using FTIR or Raman confocal micro-spectroscopy has proven the potential of these approaches for the identification of pathological areas delivering additional histological information for diagnosis based on the molecular profiles of the samples. Applied to single cell analysis, the specificity of the techniques allows *in vitro* screening of toxicological effects of ionising radiation, nanoparticles or the action of chemotherapeutic agents at the subcellular level [7–12]. Recently, vibrational spectroscopy in the field of spectral cytology performed on cervical samples for the detection of abnormal cells based on their chemical composition has attracted more attention [13, 14]. However, diseases such as cancer could remain asymptomatic until they reach a stage of development too advanced for effective treatment, and therefore routine screening for early detection is crucial to improve patient prognosis. In this context, bodily fluids are of great interest [15], and, although urine or saliva could be attractive candidates for patients screening and monitoring [16, 17], blood serum remains potentially the most valuable alternative to identify reliable biomarkers for diagnosis purposes [18]. The development of proteomic technologies with medical perspectives needs to align with clinical requirements such as compatibility with large patient cohort analysis, speed and cost effectiveness to enable clinical translation of these promising technologies. Analysis of human serum for the identification of disease-associated biomarkers is a challenging task, with numerous obstacles to overcome. For instance, in recent studies investigating the feasibility to develop methods for proteomic investigation, the high protein content (i.e. 60–80 mg/ml) could be interpreted as a beneficial attribute of the sample [19]. Indeed, with about 25% of all proteins encoded by the genome predicted to be secreted, serum is a rich source of biochemical products that can act as biomarkers of disease or physiological status of a patient [20]. It is expected that, by measuring concentrations of specific molecular species in the serum, the presence of disease associated proteins being secreted and shed from cells and tissues can be determined. For example, increased serum levels of prostate-specific antigen (PSA) and CA125 are routinely used for the *in vitro* detection of cancer in the prostate and ovary in clinic routines [21]. However, the main limiting characteristic of the serum is the dynamic range of concentrations observed between the abundant, high molecular weight and the sparse, low molecular weight proteins. As a part of the high molecular weight fraction (HMWF), proteins such as human serum albumin

(HSA) (57–71%) and globulins (8–26%), completely dominate the composition of the serum quantitatively but only account for 0.1% of the total number of protein species. Consequently, many potentially important molecules and markers can escape detection, due to difficulties in monitoring the smaller, less abundant serum constituents. For example, proteolysis within the tissue or deregulated post-translational events participate in fragmentation of proteins produced in the tumors that diffuse into the circulation [22]. Disease biomarkers are often low-molecular-weight proteins/peptides, exhibiting far greater tumor and vascular permeability due to their molecular size range, and are thus secreted into the bloodstream as a result of the disease process. The Low molecular weight fraction (LMWF) (≤ 25 kDa) plasma or serum proteome has been the focus of recent attempts to identify novel diagnostic and prognostic cancer biomarkers [23, 24]. The LMWF holds much promise for identification of diagnostics or prognostic markers, further supported by recent associations of pathological conditions with small protein and peptide profiles in serum, notably for diabetes [25], cardiovascular or infectious diseases [26]. The analytical capabilities of traditional proteomic methods are limited given the large dynamic range of concentrations in the serum. Therefore, depletion of the highly abundant proteins is now recognized as the first step, yet to be optimized, in the analysis of the serum proteome. A key challenge in identifying constituents in the LMWF in the serum is to overcome the 10 orders of magnitude concentration difference between the least abundant (interleukins, cytokines < 5 pg/mL) and the most abundant (albumin: up to 60 mg/mL).

Infrared and Raman spectroscopy have great potential for serum based detection of disease biomarkers in serum [27, 28]. However, the techniques suffer from similar limitations to those encountered with chromatography and/or mass spectroscopy based approaches. Highly abundant proteins can strongly dominate the data collected, making it difficult to detect small variations in the features related to the presence of the informative low molecular weight proteins/peptides [29]. This present study details a proof of concept approach, combining centrifugal filtration with Attenuated Total Reflection infrared (ATR-IR) spectroscopy for the detection of glycine (selected as a small model biomolecule/biomarker) within the complex human serum mixture. The filtration approach is an attractive alternative to time consuming separation techniques such as chromatography, offering a rapid and specific fractionation of proteins according to molecular weight without the need for chemical extraction that could affect the spectral profiles recorded. The spectroscopic approach is demonstrated as a highly sensitive, label free method of detecting and quantifying the small molecule within the LMWF.

2. Experimental

2.1 Sample preparation

Sterile filtered human serum from normal mixed pool (off the clot) was donated by the University Hospital CHU Angers (France) as part of the Cancéropôle du Grand-Ouest consortium. Commercially available centrifugal filtering devices, Amicon Ultra-0.5ml (Millipore - Merck, Germany), with cut-off points at 10 kDa, were employed in this study, as an illustrative example. As a result, 2 fractions were obtained following filtration; the first representing proteins with a molecular weight higher than the cut off point of the filter used (concentrate); the second corresponding to the fraction passed by the membrane and collected in the vial (filtrate). The serum was 1:10 diluted in distilled H₂O prior to processing, 0.5 mL of the solution being placed in the centrifugal filter for spinning, in order to minimize any potential effects of mechanical strain and clogging on the filter membrane. The procedure for washing the centrifugal devices prior to serum analysis was adapted from [30], based on manufacturer's guidelines, and performed as follows: The Amicon Ultra-0.5ml filter was spun thrice with a solution of NaOH (0.1M), followed by 3 rinses with Milli-Q water (Millipore Elix S). For both washing and rinsing, 0.5 mL of the respective liquid was added to the filters and the centrifugation was applied for 10 mins at 14 000g followed by a spinning with the devices upside down at 1000g for 2 mins in order to remove any residual solution contained in the filter.

Additionally, Albumin from human serum (Sigma-Aldrich, Ireland) and glycine (Cooper, France) were analysed as reference chemical compounds for the evaluation of the centrifugal filtration process. Notably, glycine has been selected because it is the smallest amino acid, having a molecular weight of 75 Da, which results in free diffusion through the membrane of the centrifugal filters and can serve as a model for LMWF peptides. Firstly, the whole human serum sample has been supplemented with known concentrations of glycine; 0.0 mg/mL (control), 0.5 mg/mL, 1 mg/mL, 2.5 mg/mL, 5 mg/mL, 7.5 mg/mL, 10 mg/mL, 17.5 mg/mL, 25 mg/mL and 50mg/mL, in order to explore the dynamic range and sensitivities of the ATR-IR measurement of unprocessed samples. Therefore, the concentrations has been selected to cover a wide range of concentrations from extremely high, non-clinically relevant, to relatively low in order to better understand and explore the capabilities and behavior of the ATR-IR instruments in humans serum monitoring. Secondly, a stock solution corresponding to the LMWF, prepared following centrifugal filtration with the 10kDa filters, has been supplemented with known concentrations of glycine of 0.01 mg/mL, 0.05 mg/mL, 0.1 mg/mL, 0.25 mg/mL, 0.5 mg/mL, 1 mg/mL, 2 mg/mL and 2.5 mg/mL. The serum stock solution without

addition of glycine has been used as a reference to build the regression models.

2.2 Dry mass measurements

In order to calculate total dry mass of the samples, it is necessary to freeze-dry (lyophilize) the samples. Thus, freeze drying was performed using a Labonco lyophilizer (Freezone, USA), whereby the vials have been placed overnight at -51°C and under vacuum (<0.05 mBar). This procedure has been applied on sets of whole human serum (n=3) and following ultrafiltration with a 10kDa filter (n=3) to estimate to overall dry mass but also determine the amount of biological material recovered in the filtrated fractions, corresponding to the LMWF.

2.3 Data collection using the ATR-IR

ATR-IR spectra were recorded using a Bruker vector 22 equipped with a UATR module. Sample penetration is both wavenumber and sample dependent, but is typically on the order of 1µm. Spectral data were the result of 16 scans, with a spectral resolution of 4 cm⁻¹ covering the spectral window 4000 – 600 cm⁻¹. A background spectrum was also recorded (64 scans) and automatically rationed with the sample spectrum by the software.

Whole human serum: The spectroscopic analysis of samples was performed directly after deposition of a drop on the crystal (0.5 µL, to ensure sample coverage is less than the full area of the crystal, minimizing the effects of inhomogeneities), while still in liquid form and also in a dry state following 10 min air drying prior to recording. 10 replicates have been performed for each serum sample.

Filtered fraction of the human serum: The relatively weak concentrations of the filtrates (LMWF) following centrifugal filtrations using a 10kDa filter have been found to greatly impact on the spectral data collected. The freeze drying process was therefore used to re-concentrate the molecules present in the LMWF. Based on a protocol detailed above, 400 µL of the filtrate were processed and the resulting lyophilized material was re-suspended in distilled water (20 µL). However, the recording of IR spectra from the LMWF in the liquid form has still been found not relevant due to the signal to noise ratio exhibited and only air drying of the drops deposited on the crystal has allowed to generate relevant data. Similarly to whole human, 0.5 µL has been deposited on the crystal and 10 replicates have been performed for each sample.

2.4 Size exclusion chromatography (SEC)

SEC was performed on an Alliance e2695 (Waters) equipped with both Wyatt miniDawn Treos and Optilab T-rEx (Wyatt Technology Europe GmbH) detectors. 50 μ L of each serum sample was injected onto a SHODEX KW803 column (300 mm x 7.8 mm) and separated at a flow rate of 0.7 mL/minute using PBS (pH 7.2) as elution buffer. The PDA 2998 apparatus (Waters) was employed for the UV detection at 280 nm, while the MALLS detection was performed with a three-angle (43.6°; 90.0°; 136.4°) miniDawn Treos detector (Wyatt Technology Europe) operating with a 60-mW solid-state laser at 658 nm. For consistency, a total elution time of 25 mins was used for all samples. The whole serum being highly concentrated, a 1:50 dilution was prepared prior to SEC analysis. In comparison, the LMW filtrates were analyzed without any further dilution, since they were obtained after centrifugal filtration performed on diluted serum (1:10).

2.5 Data pre-processing and analysis

The different pre-processing and data analysis steps were performed using Matlab (Mathworks, USA) with the unsupervised attribute of Principal Component Analysis (PCA) exploited as an approach to highlight and quantify the spectral variability generated by the addition of known concentrations of glycine to the human serum.

The analysis of the spectra collected has been restricted to the finger print region where the water contribution is less pronounced. While the spectra collected from liquid samples have been min-max normalised on the water band at 1636 cm^{-1} , the spectra collected from the air dried whole serum and the filtered samples (LMWF) were processed using baseline correction (rubber-band) followed by vector normalization. This enabled the observation and evaluation of the systemic variations in the data following the increased of glycine with the PCA analysis. **Vector normalization has been found to be particularly suitable for this study in order to equalize the contribution of the different serum physiological constituents in the spectra collected between sets and further enhance the systematic variations in the glycine features. Ultimately the PCA analysis performed on the whole finger print region reflects the evolution of band ratios in the data, such that the features of the mixed pool serum act as internal standards such that the added glycine can be quantified.** It should be noted that, for each tested condition the lowest concentration of glycine prepared (0.5 mg/mL or 0.01 mg/mL) has been excluded from the regression model and kept as a blinded sample to verify and validate the robustness of the method.

3. Results and discussion

3.1 ATR-IR detection of a single model

biomolecule

Analysis of molecules using infrared spectroscopy can be performed following a number of different set ups, depending on the instrumentation available. Most traditionally, the spectra are recorded using the transmission mode, involving the IR beam passing through the sample. However, the ATR (Attenuated Total Reflectance) mode is particularly adapted for the analysis of highly concentrated liquid samples [31]. Notably, the approach does not require expensive spectroscopic substrates and the planar, easily washable surface offered by the crystal is an advantage for repeated measurements. Therefore, a drop can be directly deposited onto the crystal, delivering high quality spectra without saturation that can be encountered with transmission mode measurements [32]. Liquid samples can be analysed and characterised using ATR-IR spectra, but the sensitivity of the information collected directly depends on the concentration of the solution. Figure 1 presents IR spectra collected from aqueous solutions of glycine (200 mg/mL) using the ATR mode of solution ('wet' sample) (Figure 1A) and following air drying (Figure 1B). The limited spectral region of 1800-650 cm^{-1} was chosen to avoid the regions of strong water absorption in the liquid samples, and to minimise any effects of incomplete drying in dried samples, thus enabling better comparison between liquid and dried samples. While the spectrum obtained from the solution exhibits the specific features of glycine, the contribution of the spectrum of deionised water can still be observed (dotted line). However, specific features of glycine such as CH_2 rocking (914 cm^{-1}), C-N stretching (1039 cm^{-1}), N-H₃ rocking (1117 cm^{-1}), C-H₂ wagging (1333 cm^{-1}) or O-C=O symmetric stretching (1412 cm^{-1}) are easily identifiable. The band at 1333 cm^{-1} is particularly interesting due to its strong relative intensity but also its positioning remote from the main water band (1636 cm^{-1}), and is thus used for further analysis in Figure 2. The spectrum collected from the solution of glycine following air drying (Figure 1B) exhibits notable modifications that could result from aggregation and/or crystallization effects (e.g. bands between ~1100 – 1150 cm^{-1}). The main spectral features are, however, preserved.

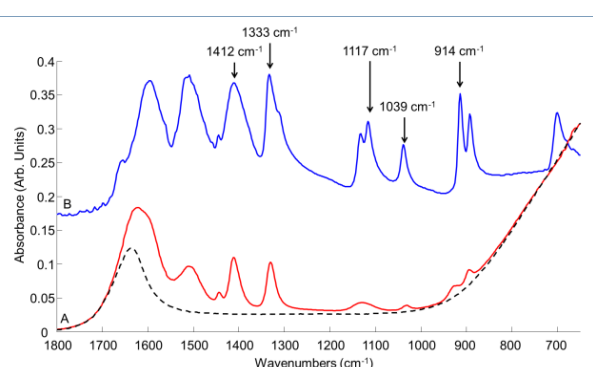


Figure 1: Typical ATR IR spectrum collected from solution ('wet' samples) of glycine at 200 mg/mL deposited on the ATR crystal (A) and after air-drying (B). No preprocessing has been applied. Additionally the dotted line represents the contribution of the water to the signal collected from the solution. The spectra have been offset for clarity and the main features of interest are highlighted in grey.

Figure 2 presents the concentration dependence of the area under the curve observed in raw data (AUC – between baseline and peak maximum) calculated for the band 1360-1280 cm^{-1} from solution of glycine (figure 2 – blue) and following air drying (figure 2 - red). Note, the spectra have not been normalized. In the case of the liquid samples, the thickness of the droplet exceeds the penetration depth of the ATR evanescent wave, and the systematic increase of the solution concentration results in a linear correlation between the concentration of glycine and the intensity of the signal ($R^2 = 0.9999$), in accordance with the Beer-Lambert law. In the dry sample, all the dissolved glycine is deposited on the crystal surface and therefore is sampled by the evanescent wave, resulting in a substantially larger absorbance for all concentrations and a steeper initial increase in the concentration dependence. However, the linearity of the measurement response is lost when the thickness of the dried sample exceeds the depth of penetration of the evanescent wave. Therefore, as seen in figure 2 (red dots), above 25 mg/mL there is a saturation of the IR absorbance. Indeed, this effect can only be clearly seen in unprocessed data collected from dry samples, as the commonly employed normalization steps of the preprocessing protocol would compensate for the loss of intensity in the spectra. However, the observations highlight the fact that, following air drying, the Beer-Lambert type dependence of the absorbance on concentration and length of the optical path are both partially lost and the critical parameter is the amount of matter deposited. Thus, while a linear dependence of the absorbance is observed over a wider range for the liquid samples, for the dried samples, increased sensitivity is achievable at the lower end of the concentration range.

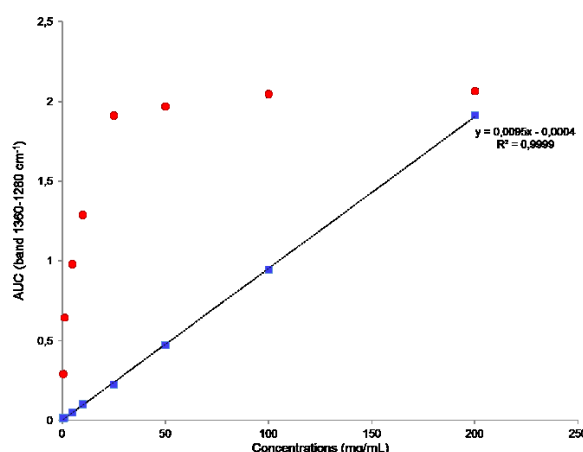


Figure 2: Evolution of Area Under the Curve (AUC) of the band 1360-1280 cm^{-1} as a function of glycine concentrations. The blue squares represent the measurement performed from the sample in liquid form while the red dots correspond to the samples deposited after air-drying. The dotted line represents a linear fit to the liquid form data.

3.2 Analysis of human blood serum

Although recording reproducible IR spectra from human serum could appear an easy task due to the particularly high concentration of protein content, specificity towards detection of minority species could be severely limited because of the complexity of the signal collected. Although human serum is naturally highly concentrated, spectra collected for the liquid form are still dominated by the water contribution (Figure 3A). In contrast, the water free spectrum collected after air drying (Figure 3B) displays numerous well defined peaks within the 4000-650 cm^{-1} spectral window, allowing clearly identification of specific features of the human serum, for example those highlighted in grey at 3280 cm^{-1} (H–O–H stretching), 2957 cm^{-1} (Asymmetric CH_3 stretching), 2920 cm^{-1} (Asymmetric CH_2 stretching), 2872 cm^{-1} (Symmetric CH_3 stretching), 1536 cm^{-1} (Amide II of proteins), 1453 cm^{-1} (CH_2 scissoring), 1394 cm^{-1} ($\text{C}=\text{O}$ stretch of COO^-), 1242 cm^{-1} (Asymmetric PO_2 stretch), 1171 cm^{-1} (Ester C–O asymmetric stretch) and 1080 cm^{-1} (C–O stretch) [33-35]. However, the dried serum spectrum is very similar to the spectrum collected from similarly dried human serum albumin (HSA), as a result of its high concentration in human serum (Figure 3D). This example perfectly illustrates the spectral dominance that can be observed from the contributions of the abundant proteins of the HMWF (Figure 3C). The spectra of these proteins swamp the contribution of less represented biomolecules, raising the question to whether or not minor variations of minority components can be detected and quantified.

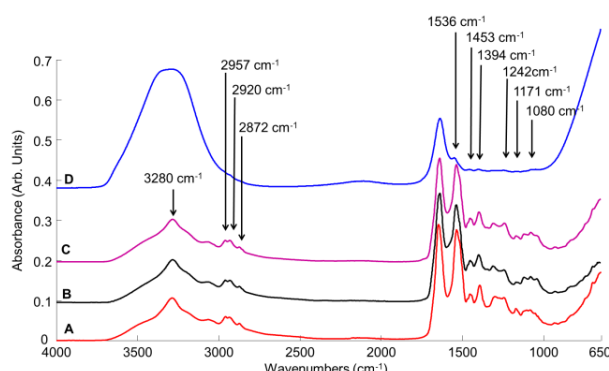


Figure 3: Typical ATR IR spectrum of a 5 mg/mL air dried human serum albumin solution (A) compared to the HMWF concentrated with a 10kDa filter (B); the human unfiltered serum following air drying (C) and in its liquid form (D). Spectra have been offset for clarity with the main features highlighted in grey. No spectral preprocessing has been applied.

3.3 Detection of Molecules in human serum

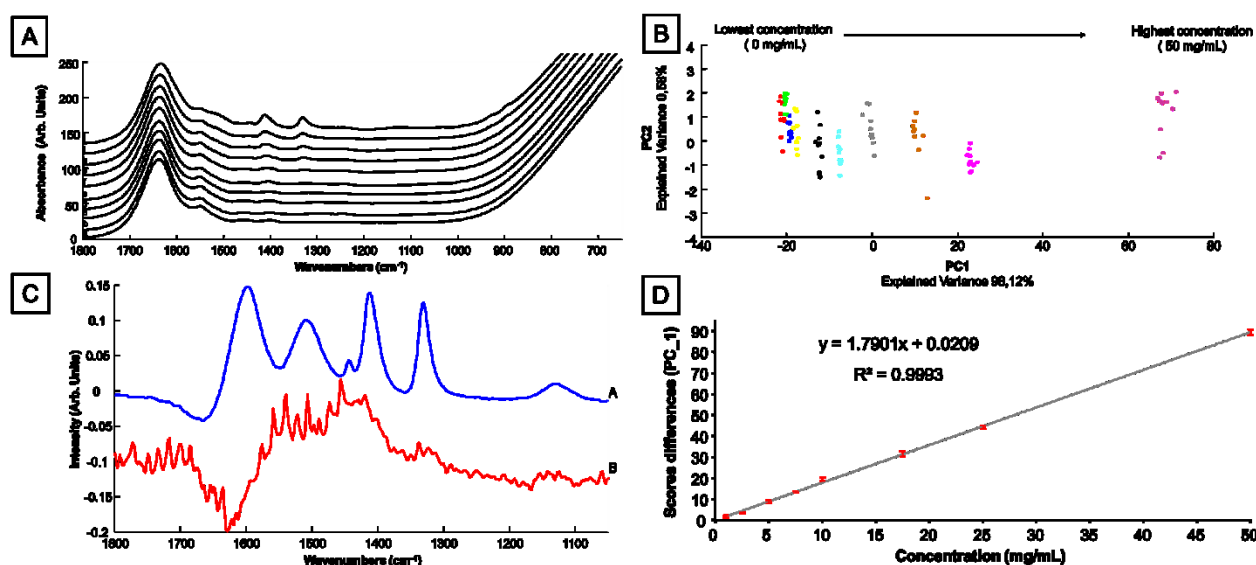


Figure 4: A: Mean ATR-IR spectra collected from unprocessed whole human serum (a) and supplemented with 0.5 mg/mL, 1 mg/mL, 2.5 mg/mL, 5 mg/mL, 7.5 mg/mL, 10 mg/mL, 17.5 mg/mL, 25 mg/mL and 50 mg/mL of glycine (respectively b-j). Spectra offset for clarity; B: Scatter plot from the PCA performed on ATR-IR spectra collected from the different solutions. Colour coding has been used to better identify the different groups of spectra. The top arrow indicates the distribution of the data across the scatter plot according to concentrations of glycine added; C: Loadings 1 and 2 corresponding to the PCA analysis performed on serum solutions (respectively blue and red); D: Linear regression fit of ATR-IR absorption versus concentration of glycine in whole human serum in solution. The plot represents the distance between each groups and the un-supplemented serum (0 mg/mL) calculated using the PC₁ score from the PCA.

In order to build a model to mimic variations that could naturally occur in the LMWF, pure lyophilized glycine has been added at known and variable concentrations to human serum. Following the collection of ATR-IR spectra, the capabilities of the technique have been evaluated through the monitoring of systematic spectral modifications in the data sets, despite the presence of dominant HMW proteins. Thus, a wide range of concentrations has been used for the analyses which have been performed in both liquid and dried forms, in order to evaluate the sensitivity and determine the specificity offered by the experimental set up used. Concentrations covering the range 0.5 mg/mL to 50 mg/mL have been found most suited to illustrate the capabilities of the ATR-IR toward whole human serum analysis. Although the concentrations used for illustration are not necessarily physiologically relevant, the limits of the tested range have been purposely extended in order to offer better appreciation of the spectral behavior for various amounts of glycine.

The mean IR spectra collected from the different samples prepared (liquid) are presented in figure 4A, organized from bottom to top according to the concentrations of glycine added to the whole human serum, and off-set for clarity. A stock solution of human mixed pool serum has been used in this study in order to have a consistent basis for the preparation of the samples and ensuring that the chemical composition only varies based on the added quantities of glycine.

The data collected from the samples have been analyzed by Principal Component Analysis (PCA), delivering, in an unsupervised fashion, the variance between the data sets, either graphically with the scatter plot or through the loadings expressing the discriminating spectral features [36]. The results corresponding to the PCA performed on

spectra collected from the liquid samples supplemented with glycine compared to the un-supplemented serum stock solution are presented in figure 4B. The distribution along the PC 1 axis is directly related to the variation of concentration of glycine added to the human serum, an observation confirmed by the loading of PC1, accounting for 98.12% of the variance and displayed in figure 4C.

While a clear discrimination between the different clusters can be seen at the highest concentrations (figure 4B top arrow), the degree of separation is strongly reduced when the concentrations of glycine added are below 10 mg/mL, especially for the 2.5 mg/mL (yellow dots), 1 mg/mL (blue dots) and 0.5 mg/mL (green dots).

PC 2 only accounts for 0.58% of variance which indicates quite minor variations in the spectra collected. The spectral profile of PC2 (Figure 4B) does not clearly correlate with the glycine spectrum. The positions of the different clusters formed by the spectra along the PC 2 axis are, however, not correlated with the concentration of glycine and the inter and intra cluster variability could thus be attributed to instrumental or experimental variability.

While the loading of PC 1 exhibits a spectral profile identical to the pure solution of glycine (figure 1A), the loading of PC 2 (figure 4Cb) is particularly noisy and does not exhibit any clear features which can be unambiguously related to spectral features of glycine. This is further supported by the 0.58% of variance expressed by PC 2, clearly highlighting the absence of relevant discriminative information along this PC. It should be noted that, in order to remove the tail effect visible in figure 4A due to the presence of water in the liquid samples, the PCA analysis has been performed on the range 1800-1000 cm⁻¹.

PCA remains a powerful tool, widely employed for data mining and in particular for studies aiming to discriminate data groups for diagnostic purposes [36]. The chemical information obtained from the loadings allows correlation of the distribution of the data points across the scatter plot with specific variations in the spectra, and, in this instance, to confirm the contribution of the glycine signature to the observed variance. Moreover, the location of each spectrum in the scatterplot is defined by scores along both PC 1 and PC 2 (figure 4B) which can be assimilated to coordinates positioning the groups with respect to each other. Therefore, the most different spectra, according to a specific PC, will be separated by a greater distance than those exhibiting stronger similarities. Ultimately, when the limit of specificity of the information contained in the spectra has been reached, no more discrimination would be discernible on the scatter plot, meaning that, score wise, the spectra would be characterized by similar values along the relevant PC.

To represent this, firstly, the mean score values for each tested concentration of glycine are evaluated for PC 1

using $M_i = \sum_{\text{scores}}/n_i$ (with n_i = number of spectra for a given concentration). Secondly, the distance between each group and the un-supplemented serum (0 mg/mL) has been calculated ($\Delta_{PC} = M_i - M_{0 \text{ mg/mL}}$). Finally, the degree of separation has been expressed using Eq.1:

$$xi = (\Delta_{PC_1i})^{1/2} \quad (1)$$

(i: different concentrations tested)

Considering the case of liquid human serum, PC 2, accounting for only 0.58% of the variance, can be disregarded, as there is no clear correlation with the systematic change in concentration. Finally, the values calculated can be plotted against the different concentrations of glycine added in the whole serum, as presented in figure 4D. The results for the analysis of liquid samples display acceptable linearity ($R^2 = 0.9993$). Note that the regression model has been built without taking into account the lowest concentration, 0.5 mg/mL, which has been then used as a blinded sample to evaluate the sensitivity of the methods. Accordingly, the 10 replicates from the 0.5 mg/mL sample have been inputted into the regression model giving an average predicted value of 0.45 mg/mL +/- 0.16 mg/mL.

In order to better evaluate the impact of working from liquid serum samples, the drops deposited on the ATR crystal have been air dried before recording in order to collect water free spectra. The resulting mean spectra are displayed in figure 5A, and the corresponding results from the PCA are presented in figure 5B, 5C and 5D. The absence of water bands greatly improves the visibility of the different features resulting from both the human serum constituents and the added glycine. The distribution of the different groups of spectra on the scatter plot follows a different pattern, both PC 1 and PC 2 being influenced by the presence of the glycine. As a consequence, the loadings 1 and 2 exhibit similar spectral features of this model amino acid and they respectively account for 98.2% and 1.4% of the explained variance. However, the spectra collected from the samples supplemented with 50 mg/mL glycine have been considered as outliers and excluded from the data set to build the regression model. It appears that such high concentration generated a quite different spectral profile, potentially due to aggregation, leading to difficulties in interpreting the peak position and intensity compared to lower concentrations (figure 5A).

Therefore, although the PC 2 expresses 1.4% of the explained variance only the PC 1 was used for the calculation of the degree of separation between the different clusters formed on the scatter plot, as described in equation (1).

The curve obtained from the PC 1 score according to Equation 1 exhibits quite a different profile compared to the liquid samples, as seen in figure 5D, with a deviation from linearity apparent above $\sim 10\text{mg/mL}$, similar to the case for dried glycine in Figure 1. The relationship between the glycine concentration in the serum and the variations observed in the spectra can be expressed by means of a polynomial fitting, giving a R^2 value of 0.9978. Polynomial expansions are common in the mathematical description of many physical phenomena, including perturbation theory of, for example an anharmonic oscillator, or non-linear optics. In the case of concentration dependent effects, higher order terms are commonly employed to account for intermolecular interactions, e.g. aggregation [37]. Similar to the liquid samples, the spectra collected from the 0.5 mg/mL samples have been inputted into the regression model, delivering a predicted value of $0.383 \pm 0.007 \text{ mg/mL}$.

The ATR-IR measurement setup used in this study is perfectly suited for the analysis of complex mixtures. The direct deposition of solutions onto the crystal would offer the two options to collect the data from the liquid samples or following air drying. Ultimately, high quality spectra capturing the overall chemical composition can be collected, as highlighted with the unprocessed data displayed in figure 1 and figure 3. Both approaches deliver predicted values in the range of the blinded samples tested, although they underestimate the real concentration.

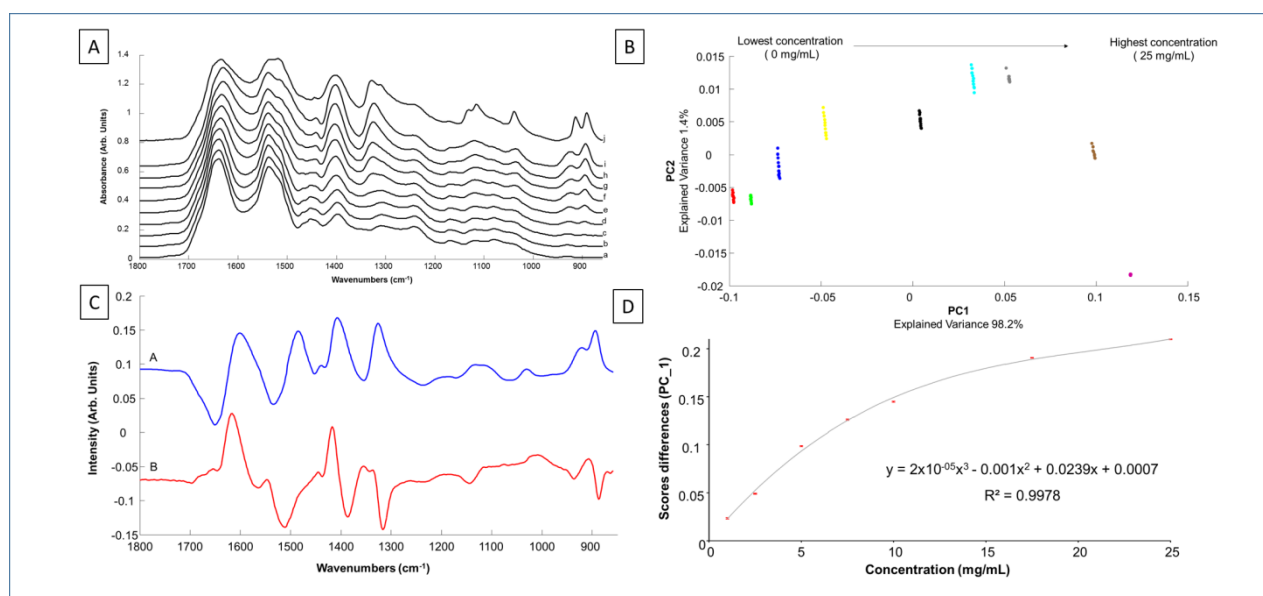


Figure 5: **A:** Mean ATR IR spectra collected from unprocessed whole human serum (a) and supplemented with 0.5 mg/mL, 1 mg/mL, 2.5 mg/mL, 5 mg/mL, 7.5 mg/mL, 10 mg/mL, 17.5 mg/mL, 25 mg/mL and 50 mg/mL of glycine (respectively b-j) - offset for clarity; **B:** Scatter plot from the PCA performed on ATR-IR spectra collected from the different solutions. Colour coding has been used to better identify the different groups of spectra. The top arrow indicates the distribution of the data across the scatter plot according to concentrations of glycine added; **C:** Loadings 1 and 2 corresponding to the PCA analysis performed on serum-based solutions (respectively blue and red); **D:** Polynomial regression fit of ATR-IR absorption versus concentration of glycine in whole human serum in solution. The plot represents the distance between each groups and the un-supplemented serum (0 mg/mL) calculated using the PC₁ score from the PCA.

Although removing a dominant molecule such as H₂O from the signal collected could be expected to enhance the relative contribution of other weaker bands, resulting in a better separation of the different data groups, only the standard deviation has been improved but the accuracy remains unacceptable. Although those results are encouraging, in the context of human serum analysis, the difficulty comes from unbalanced concentrations existing between the different constituents. For example, Albumin itself represents about 50% of the overall composition which, combined with other abundant proteins such as immunoglobulin, fibrinogen or transferrin, brings this number up to 90%. As a result, the LMWF of the serum represents only 1% of the total composition [21]. Therefore, similar to other approaches to identifying serum biomarkers, such as chromatography and mass spectroscopy [22], the potential of vibrational

spectroscopy techniques could also be greatly improved when coupled with optimised sample preparation methods aiming to isolate the LMWF for more specific spectral characterisation.

3.4 Centrifugal filtration and HMW protein depletion

Mass spectroscopy is the most commonly reported technique for the recent investigations screening the LMWF for disease-related information, probing and profiling proteomic information of hundreds of candidate disease biomarkers without the need for *a priori* knowledge of their existence or relevance to disease states [38-40]. Facing the complexity of human serum, most approaches aiming for the identification and validation of LMWF diagnostic biomarkers have introduced HMWF depletion in their protocols [41]. Although ATR-IR spectroscopy is an attractive, rapid and cost effective alternative for the screening of patient samples in order to deliver a specific diagnostic, similarly to mass spectroscopy, the depletion of abundant proteins (HMWF) is a non-optional step in the investigation of the relevant fraction of the serum, in order to highlight and potentially target small circulating biomolecules. When applying vibrational spectroscopy techniques for rapid screening of human body fluids, numerous constraints towards the integrity of the samples need to be considered in order to provide an adapted and optimised methodology. While a wide range of chemical extraction/precipitation protocols of the HMWF can be found in the literature [42, 43], the impact of the solvents involved in the different steps is unclear, considering that most of them would exhibit strong IR features, contributing to the spectra collected, adding complexity to the identification of specific spectral variations in serum biomolecules. Commercially available centrifugal filtration devices such as the Amicon ultra-0.5 (Merk Millipore) provide, as stated by the manufacturer, “a fast ultrafiltration, with the capability for high concentration factors and easy concentrate recovery from the dilute and complex matrices”. This aspect has already been explored in previous studies which demonstrated that the concentrating factor of 10, when centrifuging human serum at 14000g for 30 mins, offers possibilities for analysing samples in liquid form with both IR and Raman spectroscopy [33]. The main advantage of this approach resides in the centrifugal separation of the molecules based on their sizes without the need for additional solvents, resulting in the preservation of the biomolecules from the serum as close as possible to their native forms. Another attractive aspect of the centrifugal centrifugation for biomedical applications is the possibility to collect both the concentrated HMW and the filtered LMW

fractions. For instance, using a 10kDa cut-off point, the depletion of abundant proteins with high molecular weight can be achieved, as only small biomolecules are able to freely pass through the membrane, providing a perfectly suited *in vitro* tool for quick separation of molecular species for subsequent spectroscopic characterisation [44].

Evaluation of the efficiency of the HMW protein depletion has been performed based on size exclusion chromatography (Figure 6). The whole unprocessed human serum, presented in Figure 6A, has been used as reference in order to visualise the different features present and defined windows on the chromatogram corresponding to relevant elution times. The regions corresponding to the HMWF and LMWF have been highlighted and are respectively found between elution times of 9-15 mins and 19-22 mins. The dominant peak, positioned at 14 mins, corresponds to the contribution of albumin, the most abundant protein in human serum (Figure 6B). Comparison of the data collected from unprocessed (Figure 6A) and 10kDa filtered human serum samples (Figure 6C) clearly highlights a strong decrease in the intensities of the bands located in the HMWF window of the chromatogram. In order to enable a better estimation of the degree of depletion achieved, the Area Under the Curve (AUC) for the albumin peak was calculated. Considering the dilution factors used for the sample preparation, an average of 0.16% remaining albumin can be found in the samples following centrifugal filtration (Table 1). *Although, only done in triplicates, the reproducibility of the HMW proteins depletion has been found quite acceptable with remaining percentages of albumin comprised between 0% and 0.37%.* According to the chromatogram, most of the HMWF proteins are characterised by higher molecular masses, with elution times between 9-14 mins (Figure 6A). Therefore, an efficient depletion of the albumin would also impact on the other abundant proteins. Finally, the LMWF has been found to be totally preserved during the centrifugal filtration process. A similar AUC analysis for the peak located around 20 mins, estimated values of 0.188 +/- 0.004 and 0.190 +/- 0.012 for the whole unprocessed and the 10kDa filtered human serum respectively (Figure 6A and 6C).

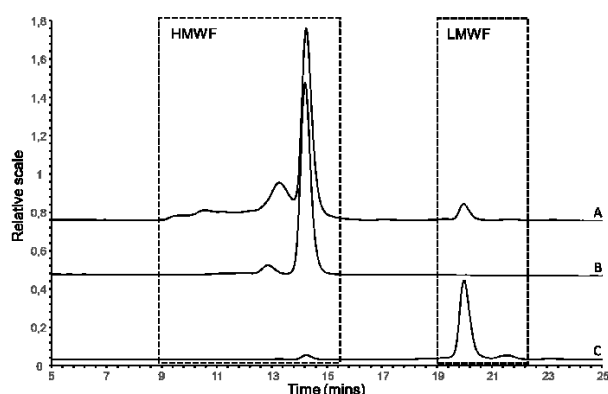


Figure 6: Size Exclusion Chromatography performed on **A:** Whole serum (1:50 dilution); **B:** Human serum albumin (1 mg/mL); and **C:** Filtrate obtained from human serum after centrifugal filtration using a 10kDa device (serum diluted 1:10 prior to centrifugation) and Chromatograms off set for clarity. The regions corresponding to the High Molecular Weight Fraction (HMWF) and the Low Molecular Weight Fraction (LMWF) have been highlighted.

Table 1: Area Under the Curve (AUC) calculated for the albumin peak observed in the chromatograms displayed in figure 6. The percentage of Albumin (Alb) takes into account the dilution factors used for the sample preparation.

	Whole serum	Filtrate (10kDa)
AUC (min-max)	0.4725 - 0.5048	0 – 0.0115
AUC (average / STD)	0.489 +/- 0.023	0.004 +/- 0.0053
Vol injected (µl)	50	50
Dilution factor	50	10
% HSA (average)	100% (reference)	0.16%

3.5 Centrifugal filtration for human serum

LMWF spectroscopic analysis

Following centrifugal filtration using 10kDa filters, the depletion of abundant proteins can be efficiently achieved as highlighted using exclusion chromatography. The removal of the HMWF will naturally have huge consequences on the IR spectra collected from the filtered part of the serum. The glycine, selected as a model, being the smallest amino acid, will freely pass through the centrifugal device membrane and be collected with the LMWF. Samples of whole serum and filtrates obtained from centrifugal filtration with 10kDa filters have been prepared by freeze drying and weighing of lyophilisates was performed to quantify the amounts. In the present case, the serum stock solution used in this study has an initial concentration of 98.5 +/- 0.2 mg/mL, taking into account all molecular species naturally present in the serum, while, after processing with centrifugal filtration, the LMWF was found to be 13.3 +/- 0.5 mg/mL. Therefore, the range of concentrations of glycine analysed for filtered human serum has been considerably reduced, extending from 0.01 mg/mL to 2.5 mg/mL.

The mean spectra collected from the different LMWF dry samples are presented in figure 7. Based on a similar approach to the one used previously on whole serum, the position for each group of data has been estimated according to the scores corresponding to PC 1 (see equation 1) and a correlation curve has been drawn. Once again, the curve follows a polynomial pattern and a trend line can be fitted with $R^2 = 0.9981$. Following depletion of the abundant proteins, the sensitivity for the detection of glycine has been considerably improved and the regression model built allows to deliver a predicted value of 0.011 mg/mL +/- 0.006 mg/mL, for the 0.01mg/mL sample. Although the standard deviation remains quite high compared to the average predicted value, the depletion of HMW proteins greatly enhanced the sensitivity of the method. The main limitations encountered remain purely technical, as both the LMWF fraction and the added glycine are particularly low in concentrations which make the data collected strongly influenced by the reproducibility of the drops deposited. A much lower standard deviation could easily be achieved using an automated system for deposition rather than micropipettes which could lead to uncertainties.

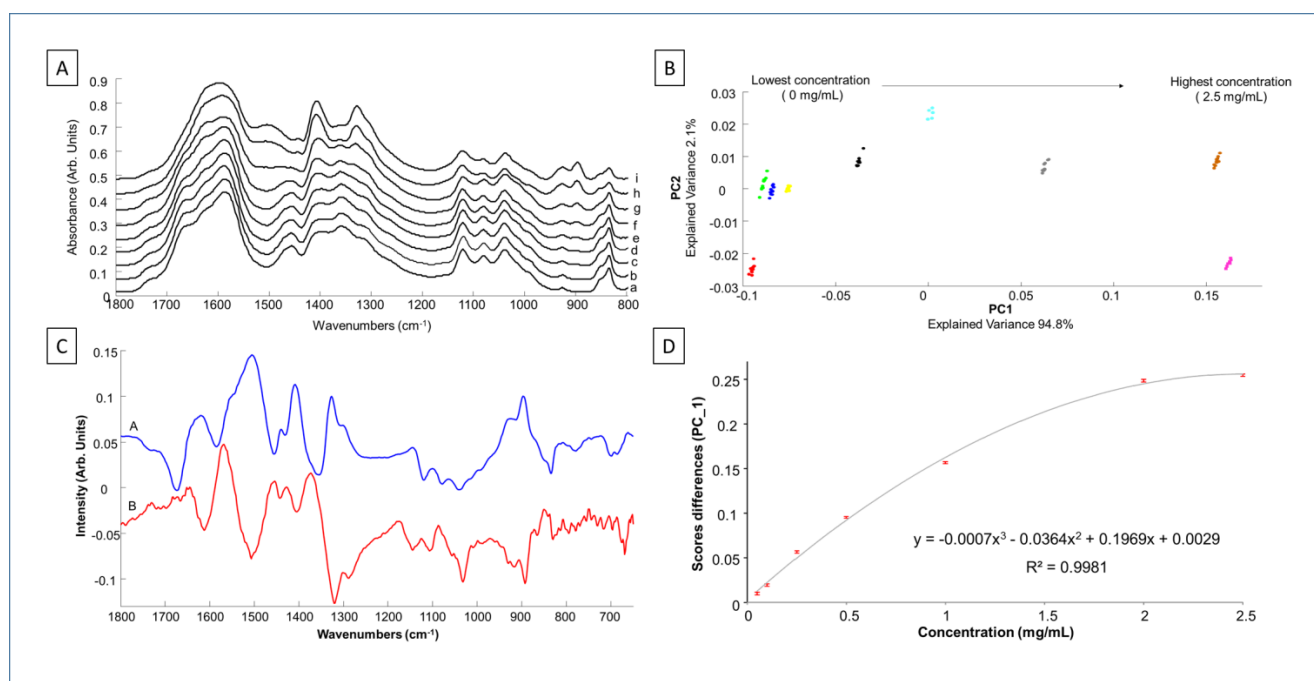


Figure 7: **A:** Mean ATR IR spectra collected from the ultrafiltered human serum (a) and supplemented with 0.01 mg/mL, 0.05 mg/mL, 0.1 mg/mL, 0.25 mg/mL, 0.5 mg/mL, 1 mg/mL, 2 mg/mL and 2.5 mg/mL of glycine (respectively b-i)-offset for clarity; **B:** Scatter plot from the PCA performed on ATR-IR spectra collected from the different solutions. Colour coding has been used to better identify the different groups of spectra. The top arrow indicates the distribution of the data across the scatter plot according to concentrations of glycine added; **C:** Loadings 1 and 2 corresponding to the PCA analysis performed on serum solutions (respectively blue and red); **D:** Polynomial regression fit of ATR-IR absorption versus concentration of glycine in whole human serum in solution. The plot represents the distance between each groups and un-supplemented serum (0 mg/mL) calculated using the PC1 score from the PCA.

3.6 Discussion

The focus in the field of clinical serology has recently been shifted to an unexplored reservoir of potentially relevant biomolecules gathered in the LMWF of the proteome, incorporating their metabolic products (peptidome). While the LMWF serum peptidome has been dismissed as too small to be biologically relevant, it has been also been reported that this fraction may contain a rich, untapped source of disease-specific diagnostic information [44, 45]. The low molecular mass range (<10 kDa) remains largely uncharacterized but promises to contain a precious source of previously undiscovered biomarkers, mostly because it has now been recognized to contain shed proteins and protein fragments (clipped or cleaved) originating from physiologic and pathologic events taking place in all perfused tissues [43, 45, 46]. While the size of HMW proteins prevents them from entering the circulation, fragments from those could be found in the serum proteome, mixed with the normal constituents, the remaining challenge being to detect them among the broad molecular population.

The metabolites are physiological constituents of the human serum LMWF that can potentially reflect a patient's health status. Variations in their concentrations can be a good indicator to detect undergoing pathological processes, and therefore as physiological markers for rapid diagnostics. Glucose remains the most common example, but also one of most targeted molecules for the development of *ex vivo* quantitative models, notably IR spectroscopy [47, 48]. While the normal average

concentration for glucose is roughly about 0.8 mg/mL [49], making its detection achievable in whole serum, monitoring and estimation of its blood levels based on spectral features will be significantly limited in the presence of highly abundant proteins (HMWF). With a molecular mass of 180 Da, glucose can freely pass through the 10kDa centrifugal devices used in this work and be collected entirely in the filtered fraction. That means blood glucose levels could be determined more accurately but also that the spectroscopic analysis could be used for patients with extremely low glucose levels located in the range of a few $\mu\text{g/mL}$, which is not achievable before HMWF depletion. Moreover, glucose remains one of the most abundant metabolites found in the human serum, which probably explains its recurrent apparition in the literature. The methodology proposed here opens numerous perspectives toward the broadening of applications as a direct consequence of the isolation of the LMWF to potentially reveal the contribution of numerous other metabolites which will fall within the lowered limit of detection range estimated [50]. For instance, most amino acids can be found naturally in the human serum, levels ranging from 0.262 mg/mL (proline), to 0.0132 mg/mL (alanine) to 0.024 mg/mL (glycine), 0.023 mg/mL (phenylalanine). Moreover, a large number of other metabolites with clinical potential for health monitoring can be found in human serum [49, 50], including, as examples, lactate (acid base homeostasis), creatinine (renal health) or carnitine (heart dysfunction), with respective normal concentrations documented at 0.29 mg/mL, 0.009 mg/mL and 0.005 mg/mL [50]. Centrifugation can process many vials at the same time, and thus is much faster than for example chromatographic separation techniques. The study demonstrates the potential advantages of spectroscopic methods over the subsequent analysis techniques. Simplifying the serum composition by depletion could lead to the development of more suitable methods for *in vitro* diagnostics which, coupled to the considerable improvements made in recent years in the field of data preprocessing and analysis, will allow accurate and reproducible models for quantitative screening of key molecules present in the LMWF to be built.

4. Conclusion

The rapid increase of scientific works referenced in the literature demonstrates the growing interest in translating and adapting the capabilities of vibrational spectroscopic techniques to the development of serological diagnostic methods. However, little or no consideration has been given to the protocols employed for sample preparation, leading to a poor relevancy of the outcomes due to the

high degree of complexity of the serum. Although vibrational spectroscopy delivers the chemical fingerprint of the analyzed samples, with, in principle, contributions from all molecular species present, the limit of detection found experimentally would suggest that performing analysis on whole human serum will imply negligible contribution from numerous constituents present at low concentrations. Human serum screening, and by extension all body fluids, is still an emerging field of application and in order to strengthen the position of Infrared and Raman spectroscopy as clinical tools, numerous questions need to be considered and addressed by the community, firstly, to be in accordance with other approaches currently used such as Mass spectroscopy and secondly, be able to deliver clinically relevant information for diagnosis. Adapting and optimizing sample preparation methods have a considerable impact on the sensitivity of the spectroscopic analysis of the LMWF of the human serum. The model built on glycine monitored by ATR-IR clearly demonstrates that removal of water, by drying, and the HMWF, by centrifugal filtration, leads to an improvement of the sensitivity of the technique which allowed to identify blinded concentrations of glycine up to 50 times lower compared to whole human serum. Furthermore, as the sample processing is also applicable to a wider range of body fluids, the methodology presented in the present work will certainly be beneficial for the field and lead to drastic improvements in the strategies oriented toward their implementation as the next generation of diagnostics techniques clinical tools.

Acknowledgements

This work was funded in part by the Science Foundation Ireland, Principle Investigator Award 11/PI/1108.

Plateforme « PC Biomed », CHRU Tours et Université François Rabelais.

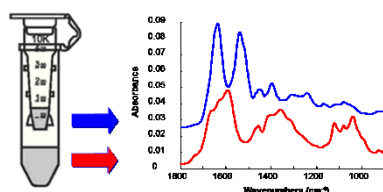
References

- [1] Kohler, K. and H. Seitz, *Sensors (Basel)* **12**, 12710-12728 (2012).
- [2] Bellisola, G. and C. Sorio, *Am J Cancer Res* **2**, 1-21 (2012).
- [3] Diem, M., A. Mazur, K. Lenau, J. Schubert, B. Bird, M. Miljkovic, C. Krafft, and J. Popp, *J Biophotonics* **6**, 855-886 (2013).
- [4] Downes, A. and A. Elfick, *Sensors (Basel)* **10**, 1871-1889 (2010).
- [5] Kendall, C., M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker, and N. Stone, *Analyst* **134**, 1029-1045 (2009).

- [6] Byrne, H.J., M. Baranska, G.J. Puppels, N. Stone, B. Wood, K.M. Gough, P. Lasch, P. Heraud, J. Sule-Suso, and G.D. Sockalingum, *Analyst* **140**, 2066-2073 (2015).
- [7] Draux, F., C. Gobinet, J. Sule-Suso, M. Manfait, P. Jeannesson, and G.D. Sockalingum, *Analyst* **136**, 2718-2725 (2011).
- [8] Nawaz, H., F. Bonnier, A.D. Meade, F.M. Lyng, and H.J. Byrne, *Analyst* **136**, 2450-2463 (2011).
- [9] Dorney, J., F. Bonnier, A. Garcia, A. Casey, G. Chambers, and H.J. Byrne, *Analyst* **137**, 1111-1119 (2012).
- [10] Farhane, Z., F. Bonnier, A. Casey, and H.J. Byrne, *Analyst* **140**, 4212-4223 (2015).
- [11] Efeoglu, E., M. Keating, J. McIntyre, A. Casey, and H.J. Byrne, *Analytical Methods* **7**, 10000-10017 (2015).
- [12] Clemens, G., J.R. Hands, K.M. Dorling, and M.J. Baker, *Analyst* **139**, 4411-4444 (2014).
- [13] Lyng, F.M., D. Traynor, I.R. Ramos, F. Bonnier, and H.J. Byrne, *Anal Bioanal Chem* **407**, 8279-8289 (2015).
- [14] Farhane, Z., F. Bonnier, A. Casey, A. Maguire, L. O'Neill, and H.J. Byrne, *Analyst* **140**, 5908-5919 (2015).
- [15] Baker, M.J., S.R. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R.A. Lukaszewski, G. Thieffn, and G.D. Sockalingum, *Chem Soc Rev* (2015).
- [16] Thomas, C.E., W. Sexton, K. Benson, R. Sutphen, and J. Koomen, *Cancer Epidemiol Biomarkers Prev* **19**, 953-959 (2010).
- [17] Arellano, M., J. Jiang, X. Zhou, L. Zhang, H. Ye, D.T. Wong, and S. Hu, *Front Biosci (Schol Ed)* **1**, 296-303 (2009).
- [18] Hanash, S.M., C.S. Baik, and O. Kallioniemi, *Nat Rev Clin Oncol* **8**, 142-150 (2011).
- [19] Pieper, R., C.L. Gatlin, A.J. Makusky, P.S. Russo, C.R. Schatz, S.S. Miller, Q. Su, A.M. McGrath, M.A. Estock, P.P. Parmar, M. Zhao, S.T. Huang, J. Zhou, F. Wang, R. Esquer-Blasco, N.L. Anderson, J. Taylor, and S. Steiner, *Proteomics* **3**, 1345-1364 (2003).
- [20] Greening, D.W. and R.J. Simpson, *J Proteomics* **73**, 637-648 (2010).
- [21] Tirumalai, R.S., K.C. Chan, D.A. Prieto, H.J. Issaq, T.P. Conrads, and T.D. Veenstra, *Mol Cell Proteomics* **2**, 1096-1103 (2003).
- [22] Tessitore, A., A. Gaggiano, G. Ciciarelli, D. Verzella, D. Capece, M. Fischietti, F. Zazzeroni, and E. Alesse, *Int J Proteomics* **2013**, 125858 (2013).
- [23] Martorella, A. and R. Robbins, *Acta Biomed* **78 Suppl 1**, 123-128 (2007).
- [24] Chertov, O., J.T. Simpson, A. Biragyn, T.P. Conrads, T.D. Veenstra, and R.J. Fisher, *Expert Rev Proteomics* **2**, 139-145 (2005).
- [25] Basso, D., A. Valerio, R. Seraglia, S. Mazza, M.G. Piva, E. Greco, P. Fogar, N. Gallo, S. Pedrazzoli, A. Tiengo, and M. Plebani, *Pancreas* **24**, 8-14 (2002).
- [26] Rubin, R. B., and M. Merchant, *Am. Clin. Lab.* **19**, 28-29 (2000).
- [27] Kong, K., C. Kendall, N. Stone, and I. Notingher, *Adv Drug Deliv Rev* **89**, 121-134 (2015).
- [28] Lacombe, C., V. Untereiner, C. Gobinet, M. Zater, G.D. Sockalingum, and R. Garnotel, *Analyst* **140**, 2280-2286 (2015).
- [29] Di Girolamo, F., J. Alessandrini, P. Somma, and F. Guadagni, *J Proteomics* **73**, 667-677 (2010).
- [30] Bonnier, F., M.J. Baker, and H.J. Byrne, *Anal. Methods* **6**, (2014).
- [31] Dorling, K.M. and M.J. Baker, *Trends Biotechnol* **31**, 327-328 (2013).
- [32] Hughes, C., M. Brown, G. Clemens, A. Henderson, G. Monjardez, N.W. Clarke, and P. Gardner, *J Biophotonics* **7**, 180-188 (2014).
- [33] Bonnier, F., F. Petitjean, M.J. Baker, and H.J. Byrne, *J Biophotonics* (2013).
- [34] Hands, J.R., P. Abel, K. Ashton, T. Dawson, C. Davis, R.W. Lea, A.J. McIntosh, and M.J. Baker, *Anal Bioanal Chem* **405**, 7347-7355 (2013).
- [35] Hands, J.R., K.M. Dorling, P. Abel, K.M. Ashton, A. Brodbelt, C. Davis, T. Dawson, M.D. Jenkinson, R.W. Lea, C. Walker, and M.J. Baker, *J Biophotonics* **7**, 189-199 (2014).
- [36] Bonnier, F. and H.J. Byrne, *Analyst* **137**, 322-332 (2012).
- [37] Penzkofer, A. and W. Blau, *Optical and Quantum Electronics* **15**(4), 325-347 (1983).
- [38] Hu, L., M. Ye, and H. Zou, *Expert Rev Proteomics* **6**, 433-447 (2009).
- [39] Klupeczynska, A., A. Swiatly, J. Hajduk, J. Matysiak, W. Dyszkiewicz, K. Pawlak, and Z.J. Kokot, *Int J Mol Sci* **17**, 410 (2016).
- [40] Labots, M., L. M. Schutte, J. C. Van Dermijn, T V. Pham, C. R. Jimenez, H. M.W. Verheul, *The Oncologist* **19**, 1028 (2014).
- [41] Luque-Garcia, J.L. and T.A. Neubert, *J Chromatogr A* **1153**, 259-276 (2007).
- [42] Chertov, O., A. Biragyn, L.W. Kwak, J.T. Simpson, T. Boronina, V.M. Hoang, D.A. Prieto, T.P. Conrads, T.D. Veenstra, and R.J. Fisher, *Proteomics* **4**, 1195-1203 (2004).
- [43] Finoulst, I., M. Pinkse, W. Van Dongen, and P. Verhaert, *J Biomed Biotechnol* **2011**, 245291 (2011).
- [44] Simpson, R.J. and D.W. Greening, *Serum/Plasma Proteomics. Methods and Protocols*. 2011: Humana Press.
- [45] Liotta, L.A. and E.F. Petricoin, *J Clin Invest* **116**, 26-30 (2006).
- [46] Anderson, N.L. and N.G. Anderson, *Mol Cell Proteomics* **1**, 845-867 (2002).
- [47] Kajiwar, K., H. Fukushima, H. Kishikawa, K. Nishida, Y. Hashiguchi, M. Sakakida, M. Uehara, and M. Shichiri, *Med Prog Technol* **18**, 181-189 (1992).
- [48] Petibois, C., V. Rigalleau, A.M. Melin, A. Perromat, G. Cazorla, H. Gin, and G. Deleris, *Clin Chem* **45**, 1530-1535 (1999).

-
- [49] Nagana Gowda, G.A., Y.N. Gowda, and D. Raftery, *Anal Chem* **87**, 706-715 (2015).
- [50] Psychogios, N., D.D. Hau, J. Peng, A.C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T.L. Pedersen, S.R. Smith, F. Bamforth, R. Greiner, B. McManus, J.W. Newman, T. Goodfriend, and D.S. Wishart, *PLoS One* **6**, e16957 (2011).

Graphical Abstract for Table of Contents



Text:

The Low Molecular Weight Fraction (LMWF) of serum holds great potential for diagnostics and health monitoring. The low concentrations of peptides and metabolites compared to highly abundant proteins remain a challenging obstacle to overcome. In order to collect specific chemical fingerprint from small serum biomolecules using vibrational spectroscopy techniques, High Molecular Weight (HMW) proteins depletion is required. Physical separation by mean of ultrafiltration offers a new dimension in serological diagnostic and further support the potential of Infrared and Raman spectroscopy as clinical tools.