The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Measuring proteins in H_2O using 2D-IR spectroscopy: pre-processing steps and applications toward a protein library

Cite as: J. Chem. Phys. 157, 205102 (2022); doi: 10.1063/5.0127680
Submitted: 23 September 2022 • Accepted: 7 November 2022 •
Published Online: 22 November 2022

Samantha H. Rutherford,[1] Gregory M. Greetham,[2] Anthony W. Parker,[2] Alison Nordon,[3] Matthew J. Baker,[1,4] and Neil T. Hunt[5,a)]

AFFILIATIONS

[1] WestCHEM, Department of Pure and Applied Chemistry, Technology and Innovation Centre, University of Strathclyde, 99 George Street, Glasgow G1 1RD, United Kingdom
[2] STFC Central Laser Facility, Research Complex at Harwell, Rutherford Appleton Laboratory, Harwell Campus, Didcot OX11 0QX, United Kingdom
[3] WestCHEM, Department of Pure and Applied Chemistry and CPACT, University of Strathclyde, 295 Cathedral Street, Glasgow G1 1XL, United Kingdom
[4] Dxcover Ltd., Suite RC534, 204 George Street, Glasgow G1 1XL, United Kingdom
[5] Department of Chemistry and York Biomedical Research Institute, University of York, Heslington, York YO10 5DD, United Kingdom

Note: This paper is part of the JCP Special Topic on Celebrating 25 Years of Two-Dimensional Infrared (2D IR) Spectroscopy.
[a)] **Author to whom correspondence should be addressed:** neil.hunt@york.ac.uk

ABSTRACT

The ability of two-dimensional infrared (2D-IR) spectroscopy to measure the amide I band of proteins in H_2O rather than D_2O-based solvents by evading the interfering water signals has enabled *in vivo* studies of proteins under physiological conditions and in biofluids. Future exploitation of 2D-IR in analytical settings, from diagnostics to protein screening, will, however, require comparisons between multiple datasets, necessitating control of data collection protocols to minimize measurement-to-measurement inconsistencies. Inspired by analytical spectroscopy applications in other disciplines, we describe a workflow for pre-processing 2D-IR data that aims to simplify spectral cross-comparisons. Our approach exploits the thermal water signal that is collected simultaneously with, but is temporally separated from the amide I response to guide custom baseline correction and spectral normalization strategies before combining them with Principal Component noise reduction tools. Case studies show that application of elements of the pre-processing workflow to previously published data enables improvements in quantification accuracy and detection limits. We subsequently apply the complete workflow in a new pilot study, testing the ability of a prototype library of 2D-IR spectra to quantify the four major protein constituents of blood serum in a single, label-free measurement. These advances show progress toward the robust data handling strategies that will be necessary for future applications of 2D-IR to pharmaceutical or biomedical problems.

INTRODUCTION

Ultrafast two-dimensional infrared (2D-IR) spectroscopy is a well-established technique that is capable of providing molecular level dynamic information on an ultrafast time scale.[1,2] In particular, the application of 2D-IR to solution phase systems and biomolecules has provided deeper understanding of conformational structure changes and molecular dynamics.[2–12]

Alongside the development of the 2D-IR spectroscopic method, advancements in laser technology[10,13,14] and pulse shaping[15] have reduced the time needed to acquire a single 2D-IR spectrum to less than a minute.[15,16] This means that 2D-IR data can now be produced

in high densities over short times, leading to proof-of-concept demonstrations of high-throughput analytical applications.[10,17–19]

Very recently, 2D-IR has been shown to have the potential for use as a tool for the analysis of proteins in H_2O-rich fluids.[20] Examples include analysis of the protein content of blood serum, including quantification of the albumin to globulin ratio, measurement of low molecular weight species, and the detection of drug binding to serum proteins.[20–22] 2D-IR has also been used to study the kinetics of fibril formation in H_2O,[23] complementing a number of prior studies in deuterated media, including a method employing protein-depleted serum.[7,24–26] Measurements of the protein amide I band in H_2O were based on two advantages of 2D-IR in comparison to IR absorption methods. First, in a reversal of the situation in absorption spectroscopy, the 2D-IR signal from the amide I band of proteins dominates that of the bending mode of water, which appears in the same part of the mid IR spectrum. Second, the ultrafast time resolution of 2D-IR spectroscopy enables the temporal separation of signals from proteins and H_2O.[20]

Taken together, these technological and methodological developments have created the potential for 2D-IR to be employed in screening applications, as might be used in drug development or to generate clinically relevant data to aid disease diagnosis.[20–22] Making the step from advanced spectroscopy to analytical tool brings new challenges, however. The most pressing of these is that analytical or screening experiments are based upon the comparisons of large numbers of different samples. This brings a need to control the experimental factors that can lead to measurement-to-measurement fluctuations, ranging from beam alignment and sample path length to pulse-to-pulse instabilities and even the type of laser used.[27,28] The extent of the challenge is compounded by the fact that the spectral changes to be detected—for example, a few percent change in protein concentration in a complex mixture or a change in shape of the amide I band upon drug binding—may be very small on the scale of the overall spectrum.

Practical approaches toward eliminating measurement-to-measurement variations have been developed, including normalization of spectra to an external molecular standard and the development of sample cells to facilitate different spectroscopic analyses.[28–30] Methods have also been devised to enhance sensitivity via data processing, including compressive data sampling and spectral reconstructions, which have been shown to improve signal-to-noise ratios.[31] Reductions in additive noise have also been demonstrated by exploiting the spectral correlation between reference and signal detectors,[32] and an edge-pixel referencing method to suppress correlated baseline noise has also been documented.[27]

In this study, we are inspired by methods that have been applied to analytical spectroscopies in other disciplines.[33,34] For example, spectral data pre-processing is an important step in the analytical workflow of IR spectral datasets, where standardization is necessary for translation to biomedical analysis[35] and to aid comparisons between different spectrometers and improve data interpretability. Pre-processing is also routinely used to minimize spectral noise and remove outliers and artifacts, which ultimately leads to improvements in data elucidation and produces increased quantification and classification accuracy.[33,35–37]

These studies suggest the importance of establishing a similar pre-processing workflow for use in 2D-IR analysis experiments. Our aim is to begin this process by demonstrating an approach specifically for use with proteinaceous samples, measured in H_2O-rich fluids. To do this, we exploit a thermal signal from H_2O, which is an integral part of capturing the 2D-IR amide I spectrum in aqueous media. The H_2O bending mode (δ_{H-O-H}; 1650 cm^{-1}) is excited simultaneously with the protein amide I band, but its spectral contribution is temporally separated from that of the protein amide I mode by its fast relaxation rate, which gives way to a long-lived thermal signature.[20] In a previous study, we reported a spectral normalization method that used the size of this thermal signal from H_2O as an internal standard to account for path length variations between samples.[38] The new pre-processing workflow further exploits this H_2O signal, to provide an integrated means of monitoring the laser bandwidth, to guide baseline subtraction and incorporates multivariate noise-reduction strategies currently used in signal processing,[34,39] to reduce the impacts of laser fluctuations and spectral noise on the outcome of any subsequent quantitative analyses.

We demonstrate our approach by first applying elements of the new workflow to two previously published datasets, allowing direct comparison of quantification accuracies and detection limits.[20,21] We then move on to apply the full pre-processing method in a proof-of-concept study, where a prototype library of protein 2D-IR spectra in H_2O is used to determine the concentrations of four major protein components of blood serum via their overlapping protein amide I signatures. In addition to providing a challenging test-case for the use of pre-processed 2D-IR data libraries, the ability to extract this information from a single 2D-IR measurement (sample volume ~10 μl) would be of significant practical benefit in biomedical applications, because current methodologies used to obtain quantitative information on the concentrations of multiple proteins in a serum sample would require as many wet chemistry assays involving time-consuming preparations.[40]

EXPERIMENT

Materials

The data for the two case studies were obtained via methods discussed elsewhere.[20,21] For the protein library study, samples were produced using pooled serum (equine), serum albumin (bovine), and γ-globulins (bovine) that were obtained from Sigma-Aldrich and used without further purification. The immunoglobulin proteins IgG, IgA, and IgM (human) were obtained from the same source, but were concentrated prior to use. For quantitative comparisons with spectroscopy data, the total protein and albumin content of the equine blood serum samples was established by standard laboratory testing at the Glasgow School of Veterinary Medicine. The total serum protein content was measured to be 71 g/l using the Biuret method.[41] The total serum albumin concentration was measured to be 30 g/l (0.45 mM) using the bromocresol green assay method.[42] The total globulin concentration, obtained from the difference between the albumin and total protein concentrations, was 41 g/l (~0.3 mM).

2D-IR spectroscopy

2D-IR method

2D-IR spectra used for the protein library study were recorded using the LIFEtime spectrometer at the STFC Central Laser Facility

using the Fourier transform 2D-IR technique, which utilizes a sequence of three mid-IR laser pulses arranged in a pseudo pump–probe geometry.[14,16] The laser pulses used in the experiment had a central frequency of 1650 cm^{-1}, a temporal duration of ~200 fs, and a bandwidth of ~80 cm^{-1}. The pulse repetition rate of the laser system was 100 kHz.

2D-IR data acquisition

The 2D-IR spectrometer employed mid-IR pulse shaping to produce and control the pump–pump delay time (τ).[15,16] During data collection, τ was incremented in 12 fs steps for a total of 4 ps using four-frame phase cycling, to minimize scattering artifacts. With cycle averaging, the total acquisition time per waiting time (T_w) was 60 s. For each sample, spectra were obtained at two waiting times—250 fs and 5 ps, using parallel pump–probe polarization. At T_w = 250 fs, strong protein signals are observed in the absence of the water response, while at T_w of 5 ps, the protein signal has decayed, leaving the thermal water signal (Fig. S1).[20] For all studies, samples were measured in triplicate, using identical sample conditions.

2D-IR sample preparation

To avoid saturation of the δ_{H-O-H} mode of water at 1650 cm^{-1}, the sample thickness was carefully controlled prior to measuring 2D-IR spectra in water. Samples were housed between two CaF_2 windows, with no spacer used. The tightness of the sample holder was adjusted, to obtain a consistent absorbance of ~0.1 for the δ_{H-O-H} + υ_{libr} combination mode of water located at 2130 cm^{-1}. Based upon the measured molar extinction coefficient of water, this corresponds to a sample thickness of ~2.75 μm.[20]

Data analysis

All data processing described below was carried out using a custom script written using the statistical analysis software program R.[43]

RESULTS AND DISCUSSION

Data pre-processing

In this section, we describe the data pre-processing workflow, which assumes that all data were collected as described in the "Experimental" section, including two waiting times (250 fs and 5 ps), which capture the protein amide I band and the thermal H_2O response, respectively. Subsequent sections describe the implementation of the workflow in three case studies.

The pre-processing workflow is shown schematically in Fig. 1. All 2D-IR data were acquired in the time-domain, by scanning τ, with a Fourier transformation as a function of τ used to generate the pump frequency axis of the 2D spectrum. The use of apodization functions (window) and zero padding to de-noise and enhance the signal response have been dealt with previously,[1] and we do not discuss them further here. In all cases, a Hamming function was applied prior to the Fourier transformation alongside zero-padding by a factor of 2.

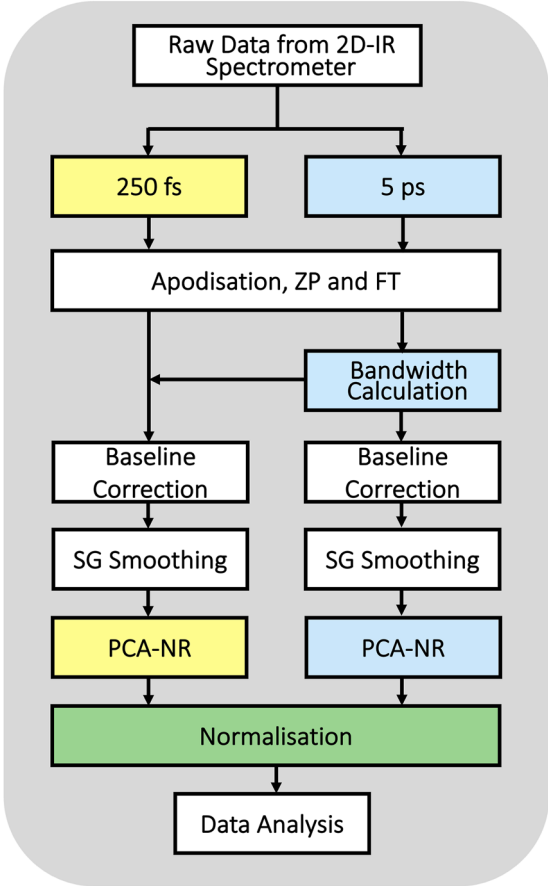

FIG. 1. Data pre-processing workflow. Colored boxes highlight when the dataset is used in its entirety for that waiting time (250 fs – yellow, 5 ps – blue, both together – green); white boxes highlight when a spectrum is processed independently on a per measurement basis. Two spectra are acquired for each sample (250 fs and 5 ps) and standard protocols [apodization, zero-padding (ZP), and Fourier transformation (FT)] are applied. The 5 ps spectrum is used to calculate the bandwidth, which forms the baseline correction for both waiting time spectra. After the baseline has been corrected, the spectra are smoothed using a Savitzky–Golay (SG) filter described in the text, and principal component analysis noise reduction (PCA-NR) is applied. The 5 ps spectrum is then used to normalize the 250 fs spectrum containing the protein information before data analysis.

Baseline correction

Following apodization and Fourier transformation, the next step in the pre-processing workflow is to apply baseline correction to the spectra. In general, 2D-IR spectra are subject to instrumental fluctuations and scattering effects, which lead to distortion of what should, in principle, be a zero spectral baseline. To correct for this, a background or reference spectrum can be measured and subtracted from the spectral data; however, fluctuations measured using these approaches may not be perfectly correlated, meaning that accurate correction of this baseline may prove difficult from one sample to another. For analytical applications, it is also preferable to use a method for baseline correction that does not add to the measurement time.

An approach that is often utilized for IR spectroscopies involves the approximation of the spectral background prior to it being removed, which leads to a more interpretable signal, allowing higher analytical classification accuracies.[33,37] This relies on adopting a model for the baseline— for example, a linear or polynomial function.[33] However, there is no standard methodology that is applied to 2D-IR spectral datasets, and one of the key challenges is to identify regions of the spectral data that constitute the baseline as opposed to the signal, to avoid erroneously subtracting the latter from the data.

Another complication that arises when acquiring 2D-IR measurements using high pulse repetition rate laser systems relates to the generally narrow bandwidth of the laser pulse produced, which can be of the order of 80–100 cm^{-1}.[14,16] Considering that the amide I band of a protein has a spectral full width at half maximum (FWHM) in excess of 40 cm^{-1}, it is important to know the bandwidth of the laser, both to guide baseline identification and avoid the potential for any bandwidth-induced distortion of the measured amide I band shapes.

Our approach uses the thermal water response as a direct measurement of the spectrometer's pump-pulse bandwidth and applies this information to inform the choice of wavenumber regions used for polynomial baseline correction.

The time-delayed thermal signal from H_2O has been described previously and is attributed to the delocalization of the bending mode of water, coupled with the fast vibrational relaxation of the fundamental vibrational mode, producing broadband responses that persist to long waiting times.[4,44] If the linewidth of the thermal signal exceeds the pump pulse bandwidth used for 2D-IR data collection, the frequency profile of the T_w = 5 ps spectrum should be defined by the pulse bandwidth. This condition applies to the LIFEtime spectrometer, where the bandwidth has been measured to be ~80 cm^{-1}.[14]

We demonstrate our approach in Fig. 2, which shows the 2D-IR spectrum of a sample of blood serum obtained with a T_w of 5 ps. The spectrum is dominated by the thermal H_2O signal.[38] The pump frequency profile of the signal is determined by taking a vertical slice through the spectrum at the probe frequency where the maximum amplitude occurs [dashed line, Fig. 2(a)]. On comparing the results of this process for several 2D-IR spectra, this was found to be well-represented by a Gaussian function [Fig. 2(b)]. The FWHM value of the fitted Gaussian functions produced an estimate of the bandwidth of the LIFEtime spectrometer of 78 cm^{-1} with a central frequency of 1648 cm^{-1} (Fig. S2). This is in excellent agreement with the experimentally determined value, giving confidence in our approach.

To apply this measurement to correct the baseline of 2D-IR spectra at other values of T_w, the fitted Gaussian function [red dashed line Fig. 2(c)] is used to inform a weighted function [black line Fig. 2(c)] that effectively allocates a ratio of "signal" to "baseline" at each frequency [Fig. 2(c) blue and red, respectively]. To exclude as much signal as possible from the baseline correction, a threshold of 12.5% of the peak amplitude was deemed appropriate, to assign regions of signal and noise [Figs. 2(c) and S2]. This weighted function is then utilized to guide a second order polynomial fit that favors the end regions of the spectrum, ensuring that the regions of signal are not influencing the baseline correction algorithm. Lower and higher order polynomials were also tested; however, upon inspection of spectra post baseline correction, it was apparent that these tended to under- and over-fit, respectively, and thus were deemed unsuitable.

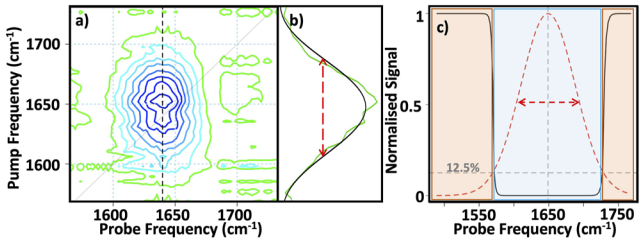

FIG. 2. Bandwidth calculation and application to the baseline subtraction process. (a) Serum 2D-IR spectrum taken at T_w = 5 ps. Black dotted vertical line indicates probe frequency 1645 cm^{-1} spanning full width of the peak, as indicated in panel (b). (b) Data (green trace) and fit (black trace) of the probe frequency identified in panel (a). Red horizontal dashed line indicates full width at half maximum (FWHM). (c) Gaussian fit to the data (red dashed trace) plotted alongside weights used for the baseline correction (solid black trace). Red dashed arrow indicates FWHM. Dashed vertical and horizontal lines mark the Gaussian peak frequency and 12.5% of peak amplitude, which are used to define noise and signal as denoted by red and blue boxes, respectively.

The bandwidth profile of a laser can change during measurements and from day-to-day, while different instruments will have unique bandwidth profiles. These factors highlight the benefits of a methodology that provides an online, customized approach to baseline correction. As the bandwidth's limited response is inherent in the water that is part of the sample being studied, this approach does not require additional measurements or instrumental adjustments.

Smoothing

The Savitzky–Golay moving filter is readily applied to remove high frequency noise from signals, while retaining the height and shape of spectral peaks.[45] The algorithm is applied identically to both the pump and probe frequency of the 2D-IR spectrum, each utilizing a third order polynomial with a 5-point window, so as not to over-fit the data.

Principal component analysis—Noise reduction

The next step in the pre-processing workflow applies multivariate noise reduction. Similar to all measurement technologies, 2D-IR is subject to noise from multiple sources. Principal component analysis (PCA) is a multivariate analysis technique, used for performing orthogonal linear transformations, that has seen success, with spectral and imaging modalities as a means of noise-reduction (NR).[34,39,46,47] PCA reduces the dimensionality of the entire dataset, by geometrically projecting the data onto fewer dimensions, while maximizing the total variance within the dataset as a whole.[48] This establishes a new coordinate system, whereby successive principal components (PCs) are defined such that they identify progressively smaller contributions to spectral variance [Fig. 3(a)]. In practice, this produces limited number of PCs that contain important spectral information, while many of the higher-numbered PCs capture little to no variance throughout the entire dataset (i.e., random noise). By rebuilding the dataset following the initial PCA using only the

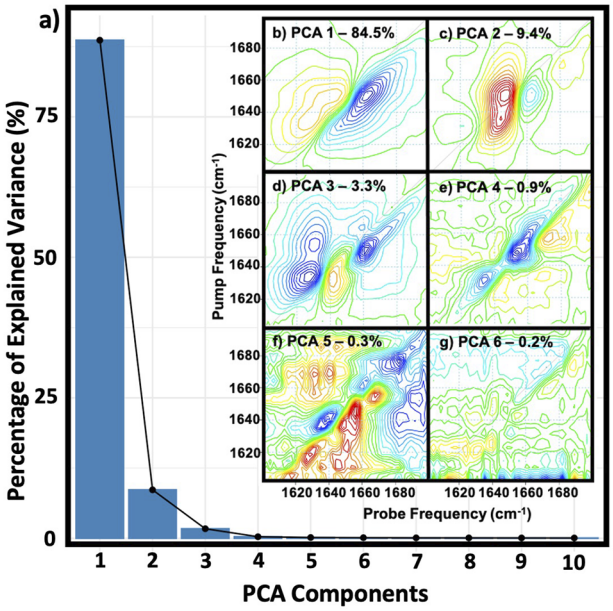

FIG. 3. Principal component analysis – noise reduction (PCA-NR) methodology. (a) Scree plot indicating total variance of each PCA component (only ten are shown here for clarity). PCA loadings are inset [panels (b)–(g)] with decreasing variance.

significant PCs that contain spectral information, it is possible to eliminate the noise that is captured by the minor PCs.[48,49]

We demonstrate this approach using the data from the final protein library study, which will be discussed in more detail below. PCA was applied to the full set of 15 spectra (four proteins and serum–each measured in triplicate) obtained with T_w values of 250 fs (Fig. 3). A PCA scree plot presenting the explained variance of each component (up to 10 for clarity) is shown (Fig. 3, main panel). The loading plots of the first six PCs are also shown [Figs. 3(b)–3(g)], where the decreasing variance, and, so, decrease in the spectral information with increasing PC number, can be seen from the percentage variance value quoted in each panel.

The next step is to identify the components which contain largely noise. A number of different methods can be applied to determine the PC that represents the cut-off point between useful signal and noise.[49] A widely accepted rule followed for inspection of the scree plot is to find the "scree elbow;" the point at which the explained variance of each subsequent component is no longer rapidly decreasing. Another technique is to apply a threshold of the explained variance, whereby only the number of components needed to account for that percentage of the total variance is included. For the data presented here (Fig. 3), the scree plot would instruct that either two or three PCs are necessary to retain spectral information, whether using a threshold of 90% of the total variance or the scree elbow method, respectively. It is important, however, to also consider the nature of the spectral dataset that is being analyzed. For example, the data in the protein library study (see below) used to generate Fig. 3 include the spectra of four individual proteins in the solution, as well as spectra of blood serum, which is a mixture of the four proteins along with other components. This experimental design dictates that the number of PCs likely to be needed to capture the data is five. This can be seen clearly in the loading plots [Figs. 3(b)–3(g)], whereby components 1 through 5 contain reasonable spectral information, but component 6 is more random in distribution and, so, contains noise. Exploratory inspection of the loadings alongside the scree plot is advised prior to implementing PCA for noise reduction.

Normalization

Following PCA-NR, the final stage in the workflow is to normalize the dataset. Normalization to a specific spectral peak is often utilized for spectroscopic data; however, this can conceal any absolute changes in magnitude associated with the peak of interest. To avoid such issues, we employ our internal normalization method, which utilizes the magnitude of the temporally separated thermal response in water to normalize the protein signals obtained at other values of T_w.[38] This approach relies on the fact that both T_w signals acquired for each sample originate from near identical laser–sample interaction processes, giving rise to magnitudes that are influenced in an equal manner and, so, correlated. This provides a route to normalize the dataset using a label-free internal standard, reducing measurement fluctuations and accounting for sample-to-sample variations in cell path length.[38]

Workflow implementation

Upon application of the full pre-processing workflow described above, the 2D-IR spectrum of a blood serum sample becomes significantly less noisy, as shown in Fig. 4. Inspection of the 2D-IR pump slices projected onto the probe axis, as shown in Figs. 4(a) and 4(b) (and Fig. S3), demonstrates how the prescribed workflow resolves the spectral bands more clearly and places them on a clean and flat baseline. Additionally, coupling to the protein amide II mode at 1550 cm^{-1} can be seen clearly after pre-processing, whereas it was previously hidden in the noise [Figs. 4(a) and 4(b), red arrows]. Furthermore, comparisons of the 2D-IR spectral profiles of the serum amide I peaks show considerably less noise after the signal processing methods have been applied [Figs. 4(c) and 4(d)], with the signal-to-noise ratio increasing from 73 to 119. A more in-depth comparison of the serum spectrum following each stage of the pre-processing workflow is shown in Fig. S4, where it can be seen that the S/N ratio improves with each successive step. After spectral normalization is complete, the dataset is ready for analysis.

It is anticipated that this methodology could be applied to other protein datasets, including protein dynamics studies. Although beyond the scope of this article, the application of this pre-processing methodology should, in principle, be applicable to studies where the waiting time is varied, as the PCA-NR and associated pre-processing methods do not appreciably alter the spectral line shapes. However, care would have to be taken in identifying the PCs associated with, for example, changes in the 2D-IR line shape due to dynamic processes, so as to not inadvertently remove them as part of the NR process.

We now demonstrate the benefits of the workflow through its application to two previously published datasets and finally extend the full pre-processing method to a new application. The workflow

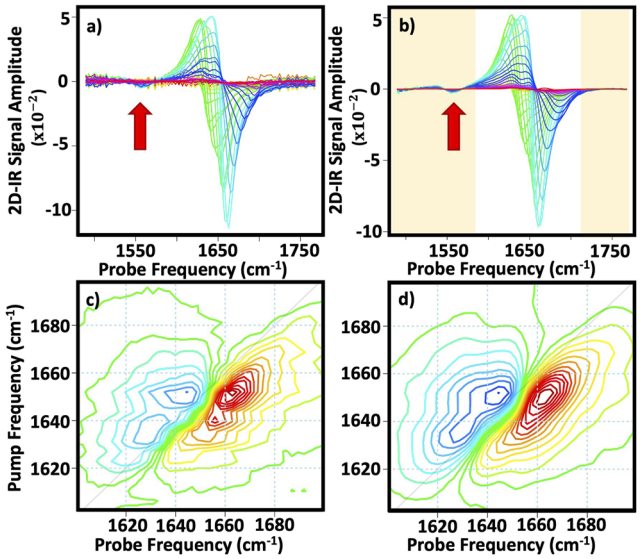

FIG. 4. Serum 2D-IR pump slices at T_w = 250 fs (a) before and (b) after application of pre-processing techniques. Yellow highlight in (b) indicates the region used to subtract the baseline. Red arrows identify the amide II region discussed in the text. Slices are displayed using a rainbow color scale (red–violet) from low to high frequencies, respectively. Enlarged versions of these panels are shown in Fig. S3. Serum 2D-IR spectra at T_w = 250 fs (c) before and (d) after application of pre-processing techniques.

is structured in such a way that it can be applied fully or partially to 2D-IR spectral datasets, and, as such, in the first case study the pre-processing is done step-wise, to highlight the benefits of each stage of the technique.

Case study 1: AGR in serum

Blood serum samples consist of a mixture of proteins (~71 g/l), along with minor components such as fatty acids, minerals, carbohydrates, and nucleic acids. The main protein constituents are serum albumin [30 g/l (0.45 mM)], which has a secondary structure dominated by α-helixes, and γ-globulins (41 g/l), which are a mixture of a number of β-sheet-rich proteins. The γ-globulins are largely comprised of the immunoglobulin proteins IgG, IgA ,and IgM, which make up to >99% of the total γ-globulin content.[50]

In a previous study, we demonstrated that the albumin: globulin concentration ratio (AGR) of a serum sample could be determined using 2D-IR spectroscopy.[20] The spectra of a number of samples were measured, in which γ-globulins were added to blood serum at a range of concentrations. The concentration ratio of albumin to globulins was determined by the magnitude of their respective amide I peaks at 1639 cm^{-1} (γ-globulins) and 1656 cm^{-1} (serum albumin), which arise from the differences in their secondary structure compositions (Fig. S5). Three methodologies were tested to quantify the AGR, and it was determined that the use of pump frequency slices isolating the γ-globulin and albumin fractions independently yielded the best results. The published data applied a second order polynomial baseline correction across the full probe range and the spectra were normalized to the albumin peak at 1656 cm^{-1} for clarity.

The linear regression is shown when plotting the AGR against additional γ-globulin concentration, following the addition of the SG-smoothing filter and PCA-NR elements of our new pre-processing pathway, in Fig. 5, alongside the results of the original study. When adding the SG-smoothing component, the linear regression analysis shows that R^2 increases from 0.963 to 0.977 and the root mean square error (RMSE) reduces from 0.017 to 0.013 g/l. Combining the SG filter with a five component PCA-NR yields a further improvement in the linear regression R^2 to a value of 0.981 (Fig. 5),[20] while the RMSE reduces from 0.017 to 0.012 g/l.

The true AGR value for each sample is shown via the solid black diagonal lines in Fig. 5. Perhaps the most notable result of adding the new pre-processing steps is that the experimental AGR values align more closely with the true AGR value.

The order in which SG smoothing and noise reduction techniques were applied was also found to be significant. The best results were obtained when applying the baseline subtraction prior to the SG filter and PCA-NR, as determined by the highest quantification accuracies. When the baseline correction was performed after the PCA-NR, the regression R^2 reduced to 0.95 and the RMSE increased to 0.021 g/l.

Case study 2: Glycine detection in serum

The second case study uses the results of an experiment in which glycine was employed as a model protein to evaluate the ability of 2D-IR to detect and quantify low-molecular weight (LMW) protein components of serum.[21] The 2D-IR spectra of blood serum samples with varying concentrations of glycine were obtained (supplementary material Fig. S6). To quantify the glycine concentration, the spectral diagonals were isolated, and the magnitude of the 2D-IR diagonal peak assigned to the amide II contribution of glycine (1515 cm^{-1}) was found to be linearly correlated with the glycine concentration (Fig. 6). Based on these measurements [Fig. 6(a)], we reported a detection limit of 3 g/l for glycine in serum. This was evaluated using the quantified noise floor level found in the 2D-IR spectra ($\pm 9.6 \times 10^{-3}$, as the 2D-IR signal has both positive and negative components), which is shown as the horizontal dashed lines in Fig. 6(b) inset.[21]

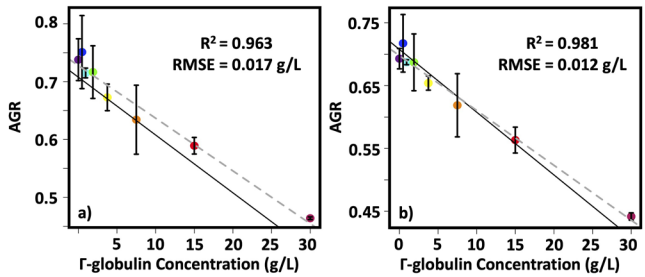

FIG. 5. Albumin to globulin ratio (AGR) of serum samples from 2D-IR spectroscopy, obtained using the pump slice method. Panels compare the AGR results when using different pre-processing methods (a) with a second order polynomial baseline subtraction and amide I normalization and (b) when combining this with an additional 5pt Savitzky–Golay smoothing filter and principal component analysis-noise reduction (PCA-NR) retaining only five components. Solid black lines indicate the "actual" AGR of the sample, and the gray dashed lines denote linear fits to the experimental result.

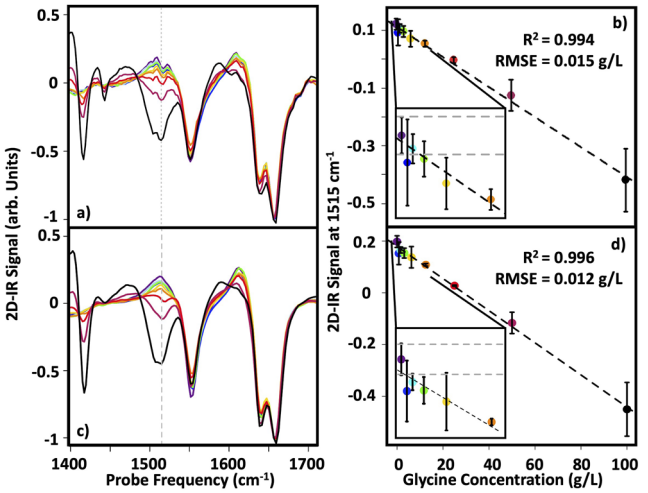

FIG. 6. Comparison of different pre-processing steps in glycine detection in serum. (a) and (c) 2D-IR diagonal slices scaled to the amide I albumin peak at 1656 cm^{-1}. (b) and (d) The average amplitude at 1515 cm^{-1} [gray dashed vertical lines in (a) and (c)] are plotted as a function of glycine concentration; error bars show 1σ variation from the measurement in triplicate. The linear regression fits (black dashed lines) and their respective R^2 and errors are shown in (b) and (d). The inset shows the data for the range 0–12.5 g/l ,expanded for clarity, and dashed horizontal gray lines show noise levels for the dataset. (a) and (b) Only a second order polynomial baseline correction and amide I normalization was used, and (c) and (d) used additional Savitzky–Golay smoothing and principal component noise reduction pre-processing steps highlighted in the text. Colored spectra in (a) and (c) align with concentrations shown in (b) and (d).

Once again, addition of the SG-smoothing filter and a seven component PCA-NR leads to a reduction in the noise levels of the spectrum diagonals when compared with those without the additional pre-processing steps [Fig. 6(c) vs Fig. 6(a)]. Improved linearity is achieved, yielding an increase in R^2 from 0.994 to 0.996 alongside a reduction in RMSE from 0.015 to 0.012 g/l [Figs. 6(b) and 6(d)]. The spectral noise floor remained unchanged, which is likely due to the PCA method being unable to remove noise from the important PCs containing spectral data; however, the detection limit is reduced from 3 to 0.8 g/l, meaning that all glycine concentrations studied are now detectable above the noise floor—[Fig. 6(d)] inset, horizontal dashed lines. As glycine is a single amino acid, this detection limit of 0.8 g/l is an upper sensitivity limit, which does not consider the effects of transition dipole coupling in secondary structures that serve to enhance the magnitude of the amide I band in 2D-IR spectra.[51] For proteins and peptides, which are composed of many amino acids, this coupling should lead to improvements in sensitivity relative to glycine. Moreover, glycine was chosen as the model protein, as it has the simplest side chain structure of all naturally occurring amino acids. In the case of other amino acids, which will present additional side chain-specific modes, we anticipate that each amino acid will yield a signature set of 2D-IR peaks, allowing similar methods to be used.

It is interesting to note that seven PCs were used to encompass the variance of this dataset, compared with the five components used for the AGR data in case study 1. This may relate to the larger wavenumber range in both pump and probe frequencies and additional number of bands being assessed in the dataset due to glycine, compared with only the amide I peak assessed with the AGR dataset. Following this, a definitive number of PCs for use in the noise-reduction aspect of the pre-processing workflow cannot be identified, but instead, each dataset should be evaluated on a case-by-case basis.

Serum protein deconvolution using protein library

The improvements to the analysis of previously published datasets now motivates an attempt to test whether 2D-IR can be used to discriminate between contributions to the serum amide I band from more than just the albumin and γ-globulin components using the full workflow. In principle, 2D-IR spectroscopy with its information-rich line shapes provides the opportunity to resolve contributions from multiple proteins. To test this, a protein library in H_2O was constructed with the aim of using this to distinguish contributions to the 2D-IR spectrum from the four major serum proteins. As described above, our equine serum sample contains ~71 g/l of total proteins [serum albumins (SA) ~30 g/l; γ-globulins ~41 g/l]. In turn, the γ globulin fraction features three major immunoglobulin (Ig) components by concentration: IgG (~13–41 g/l), IgA (~1–3 g/l), and IgM (~0.5–2.5 g/l).[52,53] Together, these account for 99% of the γ-globulins.

The average (of triplicate) 2D-IR spectra for each of the individual proteins are shown in Fig. 7, along with those of the serum [Fig. 7(a)]. Despite the fact that the Ig proteins are all dominated by β-sheet structures, each of these proteins has a unique secondary structure, which enables their differentiation using 2D-IR.[20] Two approaches were tested to quantify the components of the serum spectrum using the four-protein model library: manual deconvolution and the multivariate analysis technique—multivariate curve resolution-alternating least squares (MCR-ALS).

In the case of manual deconvolution, the spectrum of each component protein was scaled to reflect the typical concentrations found in serum. The albumin concentration was fixed to 30 g/l and the total globulin concentration was summed to 41 g/l, to align with the wet chemistry results. Specifically, the concentrations used were SA – 30 g/l, IgG – 36 g/l, IgA – 3 g/l, and IgM – 2 g/l. Building a 2D-IR spectrum from a linear combination of the four constituent components at these concentrations produced a spectrum that appears similar to that of the serum [Figs. 8(a) and 8(b)]. Upon subtraction of this from the true serum spectrum, small residual signals are obtained [Figs. 8(c) and 8(d)]. These signals lie along the spectrum diagonal, representative of α-helical and β-sheet structures still present in serum, and the coupling of the β-sheet modes between 1630 and 1670 cm^{-1} is also notable [Fig. 8(d)]. This approach indicates that it is feasible that the amide I band of serum can be deconstructed into its component parts. The fact that the residual signal that remains after subtraction of the model serum spectrum from the real one contains signals that are consistent with the presence of further proteinaceous material is reasonable. Our approach is based on accounting for a large fraction of the protein component, but there will necessarily be residual protein content, and the results of this analysis appear consistent with that. When varying each of the component concentrations, while maintaining clinical ranges, similar residuals are produced, and, so, a more sensitive methodology may be necessary to obtain results comparable with wet chemistries.

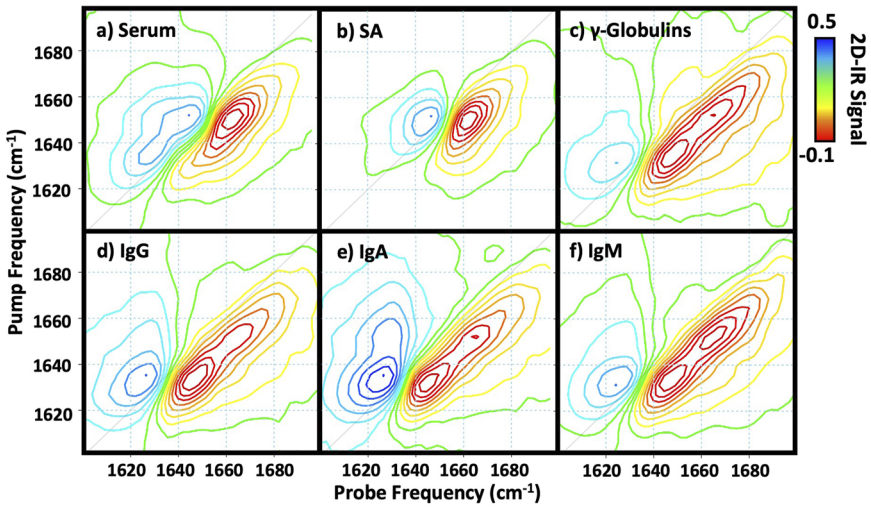

FIG. 7. Pre-processed 2D-IR spectra of (a) serum and serum proteins. (b) SA, (c) γ-globulins, (d) IgG, (e) IgA, and (f) IgM—all at concentrations of 30 g/l.

Although promising, the manual approach necessarily requires foreknowledge of the protein content of serum. Given that future analytical applications of 2D-IR will require rapid and objective analyses, we evaluated the MCR-ALS approach, as it has shown promising results and high sensitivities when decomposing chemical mixtures using other measurement techniques.[54–56] MCR-ALS is a linear model of chemically meaningful pure contributions that is used to quantify these pure contributions in mixed component samples.[54,57,58] Typically, the initial estimates of the spectra (e.g., the protein library) are input into the algorithm, along with "constraints," to help fit the library to the test spectra. Here, the "non-negativity constraint" was applied–this ensures that the concentration predictions do not produce negative values, and the initial input, the four protein spectra, were deemed "known" and, therefore, fixed–meaning that the algorithm could not change the input library.

The results of the MCR-ALS using four input components (SA, IgG, IgA, and IgM)—all at a concentration of 30 g/l (Fig. 7)—are shown in Fig. 9 and Table I. Three individual serum samples were used to test the MCR model, and the average serum spectrum is

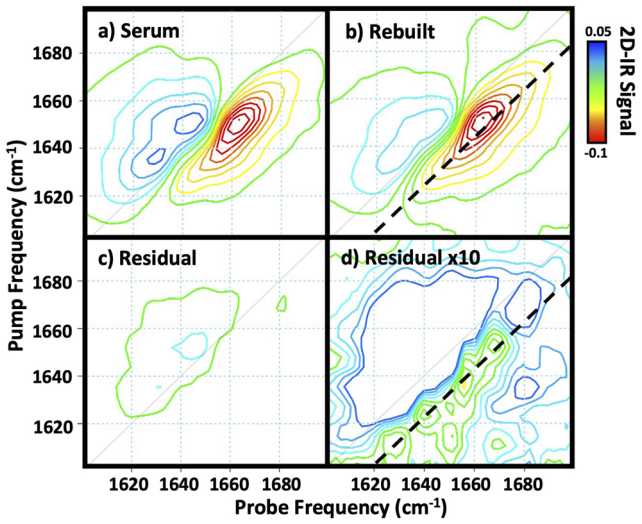

FIG. 8. Manual serum protein deconvolution. 2D-IR spectra of (a) serum containing known albumin (30 g/l) and total globulin (41 g/l) concentrations. (b) Spectrum formed from a linear combination of SA (30 g/l), IgG (36 g/l), IgA (3 g/l), and IgM (2 g/l), and (c) the residual signal after subtraction of (b) from (a), and (d) the residual signal in (c) zoomed in by a factor of 10 to observe remaining signals. Black diagonal dashed lines in panels (b) and (d) indicate spectrum diagonals.

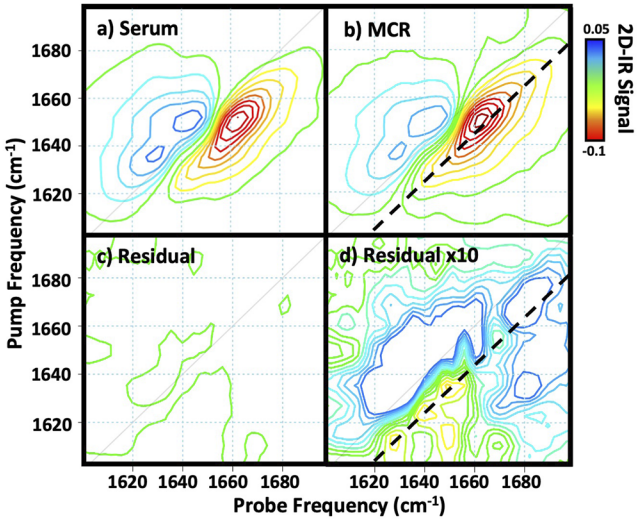

FIG. 9. Results obtained by multivariate curve resolution (MCR) using four pure component inputs—BSA, IgA, IgM, and IgG (as in Fig. 7). (a) Average 2D-IR spectrum of serum, (b) the 2D-IR spectrum obtained using MCR (c) the residual signal from subtraction of (b) from (a), and (d) the residual signal in (c) multiplied by a factor of 10 to observe remaining signals. Black diagonal dashed lines in panels (b) and (d) indicate spectrum diagonals.

TABLE I. Protein concentrations in serum samples obtained by multivariate curve resolution (MCR) with a 4-component model. The known ranges of each protein are shown in the top row for Refs. 52 and 53.

Sample	SA (g/l)	IgA (g/l)	IgG (g/l)	IgM (g/l)
⋯	30–50	1–3	13–41	0.5–2.5
1	30.1	32.4	0	0
2	28.4	47.4	0	0
3	39.0	40.5	0	0
Average	32.5	40.1	0	0

shown in Fig. 9(a). The average protein concentrations obtained using MCR (Table I) are used to calculate the original spectrum [Fig. 9(b)] and the difference between the MCR calculated and the experimental 2D-IR spectrum of serum is shown in panels (c) and (d). In this case, the algorithm was found to only identify the presence of two-components, and the contributions for IgG and IgM were returned as 0 g/l in all cases. The indication is that the technique is not able to distinguish between the immunoglobulins in the complex mixture, which may arise from the comparative similarity in their secondary structures.

This represents an ambitious challenge for 2D-IR spectral analysis, and, so, the fact that the different spectra obtained using both manual deconvolution and MCR-ALS provide similar results is encouraging. However, the quantitative results produced by MCR were not as realistic of the true serum mixture and, so, it is evident that challenges remain in developing truly unguided analysis methods. One key drawback of MCR-ALS and many least-squares fitting algorithms is that they are designed to minimize the residual fit. In this proof-of-concept trial, we attempted to use a limited library of protein spectra, to account for the spectrum of the complex mixture, which, in reality, may feature several hundred protein species.[40,59] This means that, in practice, the fitting of the spectral profile results in a minimum with a residual signal that cannot be fully removed or accounted for at this stage.

It is noteworthy that the three immunoglobulins used here have structures rich in β-sheets and are structurally very similar to one another, with subtle spectral differences between them.[20] This leads us to suggest that structurally similar α-helical proteins or intrinsically disordered proteins could also be deconvoluted in the serum matrix and would be an interesting next step in our proposed protein library method and pre-processing methods.

In summary, application of the workflow to the third dataset highlights a number of things: First, the combined effects of bandwidth-guided baseline subtraction, PCA-NR, and normalization lead to improvements in the sensitivity of the 2D-IR method in the applications described. Second, it is clear that further advancements surrounding data acquisition and sample cell design will be necessary to reduce the variability stemming from the sample path length and to enable more accurate quantification of protein concentrations. Finally, while there is promise for protein library methods, it is clear that development of larger protein libraries that are able to account for the more general aspects of the nature of serum samples that are non-specific and common will be needed alongside the application of more sophisticated analysis tools. Overall, however, these results serve to encourage further work classifying additional abundant proteins in serum, to create a larger protein library and, thus, a more inclusive study. Such a development strategy can also be expected to provide further knowledge to develop 2D-IR disease diagnosis methods through protein screening.

CONCLUSIONS

The methodology developed here demonstrates a robust and automated approach to the pre-processing of 2D-IR spectral datasets of aqueous protein samples and biofluids; for the protein library dataset, less than two minutes were required to process 46 datasets (23 samples, each with two T_w spectra), starting from time domain as-measured data to fully processed spectra, using our workflow. The pre-processing workflow has been devised and tested on previously published data and applied to a first attempt to deconvolute spectra of complex protein mixtures. Using the thermal response of water inherent in these measurements, a direct measurement of the laser pulse-bandwidth is achieved and provides a route to a guided baseline correction approach that can be customized to the laser pump–pulse and does not impact on the spectral signatures being investigated. As this information is already present in the dataset via the intrinsic water signals, the workflow does not require additional measurements or experimental alterations, offering a simple, yet powerful, approach to provide reductions in the effects of experimental noise and instrumental fluctuations.

We have verified our method in three studies, showcasing beneficial outcomes including a more accurate AGR, comparable to the current gold-standard assays, with reduced associated errors. Additionally, lower detection limits for low-molecular weight species in serum have also been demonstrated, with the upper detection limit reduced from ~3 to 0.8 g/l. It is noted that developments in experimental design may also reduce some experimental variability and lead to improvement of these works. Finally, we established the use of a small protein library to predict multiple protein concentrations in serum using 2D-IR for the first time. We demonstrated manual deconvolution of serum into its constituent components, which provided promising results; however, this approach requires prior knowledge of concentration ranges. Additionally, we found issues with the use of MCR-ALS for the purposes of a serum library, likely due to the vast range of proteins present in serum, suggesting that some of these issues with this method may be overcome with larger and more robust datasets, allowing the algorithm to recognize the 2D-IR patterns pertaining to more of the proteins present in the serum. Comparisons with gold standard assay testing demonstrate the similarity in results being obtained from both methods; however, the 2D-IR method could, in principle, achieve these results from a single label-free, non-destructive, rapid measurement of the proteins *in situ*. To our knowledge, this is the first full pre-processing workflow devised for the application toward 2D-IR spectral datasets of aqueous fluids. Combination of high-throughput systems with our devised pre-processing workflow and tighter restrictions on measurement set-up provides scope that the water content of biofluids could be the internal standard for aqueous 2D-IR datasets, and, so, the next steps to create large protein libraries become feasible. In combination with methods such as artificial intelligence and machine learning algorithms, the potential

exists for a truly unguided assessment of protein content from a single 2D-IR measurement.

SUPPLEMENTARY MATERIAL

The supplementary material contains additional figures relating to data pre-processing, methodology, and the results described in each of the three case studies.

ACKNOWLEDGMENTS

Funding from EPSRC (Grant Nos. EP/T014318/1 and EP/T014245/1) for this work is gratefully acknowledged. STFC funding for access to the ULTRA suite of spectrometers is also acknowledged. The assistance of Mr. James Harvie (University of Glasgow, School of Veterinary Medicine,) in obtaining the albumin concentration in horse serum samples is also gratefully acknowledged.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Samantha H. Rutherford: Formal analysis (equal); Investigation (equal); Software (equal); Validation (equal); Writing – original draft (equal); Writing – review & editing (equal). **Gregory M. Greetham**: Data curation (equal); Investigation (equal); Writing – original draft (equal). **Anthony W. Parker**: Investigation (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Alison Nordon**: Funding acquisition (equal); Methodology (equal); Supervision (equal); Writing – review & editing (equal). **Matthew J. Baker**: Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Methodology (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Neil T. Hunt**: Conceptualization (equal); Funding acquisition (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in University of Strathclyde KnowledgeBase at http://doi.org/10.15129/8da49741-31b6-46ad-9649-38791c6a84c7.

REFERENCES

[1] P. Hamm and M. T. Zanni, *Concepts and Methods of 2D Infrared Spectroscopy* (Cambridge University Press, 2011).
[2] P. Hamm, M. Lim, and R. M. Hochstrasser, "Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy," J. Phys. Chem. B **102**, 6123–6138 (1998).
[3] K. Kwak, S. Park, I. J. Finkelstein, and M. D. Fayer, "Frequency-frequency correlation functions and apodization in two-dimensional infrared vibrational echo spectroscopy: A new approach," J. Chem. Phys. **127**(12), 124503 (2007).
[4] L. de Marco, J. A. Fournier, M. Thämer, W. Carpenter, and A. Tokmakoff, "Anharmonic exciton dynamics and energy dissipation in liquid water from two-dimensional infrared spectroscopy," J. Chem. Phys. **145**(9), 094501 (2016).
[5] P. Hamm, "Transient 2D IR spectroscopy from micro-to milliseconds," J. Chem. Phys. **154**(10), 104201 (2021).
[6] Ł. Szyc, M. Yang, and T. Elsaesser, "Ultrafast energy exchange via water-phosphate interactions in hydrated DNA," J. Phys. Chem. B **114**(23), 7951–7957 (2010).
[7] L. Wang, C. T. Middleton, S. Singh, A. S. Reddy, A. M. Woys, D. B. Strasfeld, P. Marek, D. P. Raleigh, J. J. de Pablo, M. T. Zanni, and J. L. Skinner, "2DIR spectroscopy of human amylin fibrils reflects stable β-sheet structure," J. Am. Chem. Soc. **133**(40), 16062–16071 (2011).
[8] L. M. Kiefer and K. J. Kubarych, "Two-dimensional infrared spectroscopy of coordination complexes: From solvent dynamics to photocatalysis," Coord. Chem. Rev. **372**, 153–178 (2018).
[9] P. Pagano, Q. Guo, A. Kohen, and C. M. Cheatum, "Oscillatory enzyme dynamics revealed by two-dimensional infrared spectroscopy," J. Phys. Chem. Lett. **7**(13), 2507–2511 (2016).
[10] K. M. Tracy, M. V. Barich, C. L. Carver, B. M. Luther, and A. T. Krummel, "High-throughput two-dimensional infrared (2D IR) spectroscopy achieved by interfacing microfluidic technology with a high repetition rate 2D IR spectrometer," J. Phys. Chem. Lett. **7**(23), 4865–4870 (2016).
[11] M. C. Thielges, "Transparent window 2D IR spectroscopy of proteins," J. Chem. Phys. **155**(4), 040903 (2021).
[12] J. C. Flanagan, M. L. Valentine, and C. R. Baiz, "Ultrafast dynamics at lipid-water interfaces," Acc. Chem. Res. **53**(9), 1860–1868 (2020).
[13] G. M. Greetham, P. Burgos, Q. Cao, I. P. Clark, P. S. Codd, R. C. Farrow, M. W. George, M. Kogimtzis, P. Matousek, A. W. Parker, M. R. Pollard, D. A. Robinson, Z.-J. Xin, and M. Towrie, "Ultra: A unique instrument for time-resolved spectroscopy," Appl. Spectrosc. **64**(12), 1311–1319 (2010).
[14] P. M. Donaldson, G. M. Greetham, D. J. Shaw, A. W. Parker, and M. Towrie, "100 kHz pulse shaping 2D-IR spectrometer based on dual Yb:KGW amplifiers," J. Phys. Chem. A **122**(3), 780–787 (2018).
[15] S.-H. Shim, D. B. Strasfeld, Y. L. Ling, and M. T. Zanni, "Automated 2D IR spectroscopy using a mid-IR pulse shaper and application of this technology to the human islet amyloid polypeptide," Proc. Natl. Acad. Sci. U. S. A. **104**(36), 14197–14202 (2007).
[16] G. M. Greetham, P. M. Donaldson, C. Nation, I. V. Sazanovich, I. P. Clark, D. J. Shaw, A. W. Parker, and M. Towrie, "100 kHz time-resolved multiple-probe femtosecond to second infrared absorption spectrometer," Appl. Spectrosc. **70**(4), 645–653 (2016).
[17] R. Fritzsch, P. M. Donaldson, G. M. Greetham, M. Towrie, A. W. Parker, M. J. Baker, and N. T. Hunt, "Rapid screening of DNA–ligand complexes via 2D-IR spectroscopy and ANOVA–PCA," Anal. Chem. **90**(4), 2732–2740 (2018).
[18] J. Helbing and P. Hamm, "Compact implementation of Fourier transform two-dimensional IR spectroscopy without phase ambiguity," J. Opt. Soc. Am. B **28**(1), 171–178 (2011).
[19] J. Bredenbeck, J. Helbing, and P. Hamm, "Continuous scanning from picoseconds to microseconds in time resolved linear and nonlinear spectroscopy," Rev. Sci. Instrum. **75**(11), 4462–4466 (2004).
[20] S. Hume, G. Hithell, G. M. Greetham, P. M. Donaldson, M. Towrie, A. W. Parker, M. J. Baker, and N. T. Hunt, "Measuring proteins in H_2O with 2D-IR spectroscopy," Chem. Sci. **10**(26), 6448–6456 (2019).
[21] S. H. Rutherford, G. M. Greetham, P. M. Donaldson, M. Towrie, A. W. Parker, M. J. Baker, and N. T. Hunt, "Detection of glycine as a model protein in blood serum using 2D-IR spectroscopy," Anal. Chem. **93**(2), 920–927 (2021).
[22] S. H. Rutherford, G. M. Greetham, M. Towrie, A. W. Parker, S. Kharratian, T. F. Krauss, A. Nordon, M. J. Baker, and N. T. Hunt, "Detection of paracetamol binding to albumin in blood serum using 2D-IR spectroscopy," Analyst **147**(15), 3464–3469 (2022).
[23] S. Y. Chun, M. K. Son, C. R. Park, C. Lim, H. I. Kim, K. Kwak, and M. Cho, "Direct observation of protein structural transitions through entire amyloid aggregation processes in water using 2D-IR spectroscopy," Chem. Sci. **13**, 4482 (2022).
[24] J. P. Lomont, K. L. Rich, M. Maj, J.-J. Ho, J. S. Ostrander, and M. T. Zanni, "Spectroscopic signature for stable β-amyloid fibrils versus β-sheet-rich oligomers," J. Phys. Chem. B **122**(1), 144–153 (2018).
[25] S. D. Moran, A. M. Woys, L. E. Buchanan, E. Bixby, S. M. Decatur, and M. T. Zanni, "Two-dimensional IR spectroscopy and segmental ^{13}C labeling reveals the

domain structure of human γD-crystallin amyloid fibrils," Proc. Natl. Acad. Sci. U. S. A. **109**(9), 3329–3334 (2012).
[26]C. R. Fields, S. S. Dicke, M. K. Petti, M. T. Zanni, and J. P. Lomont, "A different hiAPP polymorph is observed in human serum than in aqueous buffer: Demonstration of a new method for studying amyloid fibril structure using infrared spectroscopy," J. Phys. Chem. Lett. **11**(15), 6382–6388 (2020).
[27]K. C. Robben and C. M. Cheatum, "Edge-pixel referencing suppresses correlated baseline noise in heterodyned spectroscopies," J. Chem. Phys. **152**(9), 094201 (2020).
[28]M. Grechko and M. T. Zanni, "Quantification of transition dipole strengths using 1D and 2D spectroscopy for the identification of molecular structures via exciton delocalization: Application to α-helices," J. Chem. Phys. **137**(18), 184202 (2012).
[29]E. B. Dunkelberger, M. Grechko, and M. T. Zanni, "Transition dipoles from 1D and 2D infrared spectroscopy help reveal the secondary structures of proteins: Application to amyloids," J. Phys. Chem. B **119**(44), 14065–14075 (2015).
[30]E. Deniz, J. G. Löffler, A. Kondratiev, A. R. Thun, Y. Shen, G. Wille, and J. Bredenbeck, "High-precision background correction and artifact suppression for ultrafast spectroscopy by quasi-simultaneous measurements in a split-sample cell," Rev. Sci. Instrum. **93**(3), 033001 (2022).
[31]J. J. Humston, I. Bhattacharya, M. Jacob, and C. M. Cheatum, "Optimized reconstructions of compressively sampled two-dimensional infrared spectra," J. Chem. Phys. **150**(23), 234202 (2019).
[32]Y. Feng, I. Vinogradov, and N.-H. Ge, "Optimized noise reduction scheme for heterodyne spectroscopy using array detectors," Opt. Express **27**(15), 20323 (2019).
[33]P. Lasch, "Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging," Chemom. Intell. Lab. Syst. **117**, 100–114 (2012).
[34]Y. Murali Mohan Babu, "PCA based image denoising," Signal Image Process. **3**(2), 236–244 (2012).
[35]M. J. Baker, H. J. Byrne, J. Chalmers, P. Gardner, R. Goodacre, A. Henderson, S. G. Kazarian, F. L. Martin, J. Moger, N. Stone, and J. Sulé-Suso, "Clinical applications of infrared and Raman spectroscopy: State of play and future challenges," Analyst **143**(8), 1735–1757 (2018).
[36]S. Magalhães, B. J. Goodfellow, and A. Nunes, "FTIR spectroscopy in biomedical research: How to get the most out of its potential," Appl. Spectrosc. Rev. **56**, 869–907 (2021).
[37]B. R. Smith, M. J. Baker, and D. S. Palmer, "PRFFECT: A versatile tool for spectroscopists," Chemom. Intell. Lab. Syst. **172**, 33–42 (2018).
[38]S. Hume, G. M. Greetham, P. M. Donaldson, M. Towrie, A. W. Parker, M. J. Baker, and N. T. Hunt, "2D-infrared spectroscopy of proteins in water: Using the solvent thermal response as an internal standard," Anal. Chem. **92**(4), 3463–3469 (2020).
[39]K. Peng, H. Guo, and X. Shang, "EEMD and multiscale PCA-based signal denoising method and its application to seismic P-phase arrival picking," Sensors **21**(16), 5271 (2021).
[40]S. Hu, J. A. Loo, and D. T. Wong, "Human body fluid proteome analysis," Proteomics **6**, 6326–6353 (2006).
[41]M. M. Lubran, "The measurement of total serum proteins by the Biuret method," Ann. Clin. Lab. Sci. **8**, 106–110 (1978).
[42]J. R. Harding and J. W. Keyser, "Bromocresol green as a reagent for serum albumin," Proc. Assoc. Clin. Biochem. **5**, 51–53 (1968).
[43]R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/.
[44]W. B. Carpenter, J. A. Fournier, R. Biswas, G. A. Voth, and A. Tokmakoff, "Delocalization and stretch-bend mixing of the HOH bend in liquid water," J. Chem. Phys. **147**(8), 084503 (2017).
[45]A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," Anal. Chem. **36**(8), 1627–1639 (1964).
[46]P. Antonelli, H. E. Revercomb, L. A. Sromovsky, W. L. Smith, R. O. Knuteson, D. C. Tobin, R. K. Garcia, H. B. Howell, H. L. Huang, and F. A. Best, "A principal component noise filter for high spectral resolution infrared measurements," J. Geophys. Res.: Atmos. **109**(D23), D23102, https://doi.org/10.1029/2004jd004862 (2004).
[47]G. Chen and S.-E. Qian, "Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage," IEEE Trans. Geosci. Remote Sens. **49**(3), 973–980 (2011).
[48]K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," London, Edinburgh Dublin Philos. Mag. J. Sci. **2**(11), 559–572 (1901).
[49]R. Bro and A. K. Smilde, "Principal component analysis," Anal. Methods **6**(9), 2812–2831 (2014).
[50]H. W. Schroeder and L. Cavacini, "Structure and function of immunoglobulins," J. Allergy Clin. Immunol. **125**(2), S41 (2010).
[51]A. Barth and C. Zscherp, "What vibrations tell us about proteins," Q. Rev. Biophys. **35**(4), 369–430 (2002).
[52]M. M. de Camargo, J. S. Kuribayashi, C. R. Bombardieri, and A. Hoge, "Normal distribution of immunoglobulin isotypes in adult horses," Vet. J. **182**(2), 359–361 (2009).
[53]G. A. Perkins, D. V. Nydam, M. J. Flaminio, and D. M. Ainsworth, "Serum IgM concentrations in normal, fit horses and horses with lymphoma or other medical conditions," J. Vet. Intern. Med. **17**, 337 (2003).
[54]A. de Juan, J. Jaumot, and R. Tauler, "Multivariate curve resolution (MCR). Solving the mixture analysis problem," Anal. Methods **6**(14), 4964–4976 (2014).
[55]A. de Juan and R. Tauler, "Multivariate curve resolution: 50 years addressing the mixture analysis problem—A review," Anal. Chim. Acta **1145**, 59–78 (2021).
[56]J. Jaumot, V. Marchán, R. Gargallo, A. Grandas, and R. Tauler, "Multivariate curve resolution applied to the analysis and resolution of two-dimensional [^{1}H, ^{15}N] NMR reaction spectra," Anal. Chem. **76**(23), 7094–7101 (2005).
[57]R. Tauler, "Multivariate curve resolution of multiway data using the multilinearity constraint," J. Chemom. **35**(2), e3279 (2021).
[58]C. Ruckebusch and L. Blanchet, "Multivariate curve resolution: A review of advanced and tailored applications and challenges," Anal. Chim. Acta **765**, 28–36 (2013).
[59]J. N. Adkins, S. M. Varnum, K. J. Auberry, R. J. Moore, N. H. Angell, R. D. Smith, D. L. Springer, and J. G. Pounds, "Toward a human blood serum proteome," Mol. Cell. Proteomics **1**(12), 947–955 (2002).