Biofluid Analysis and Classification using IR and 2D-IR Spectroscopy

Samantha H. Rutherford, [a] Alison Nordon,[b] Neil T. Hunt[c] and Matthew J. Baker [a,d]

a) WestCHEM, Department of Pure and Applied Chemistry, University of Strathclyde, Technology and Innovation Centre, 99 George Street, Glasgow, G1 1RD, UK

b) WestCHEM, Department of Pure and Applied Chemistry and CPACT, University of Strathclyde, 295 Cathedral Street, Glasgow, G1 1XL, UK

c) Department of Chemistry and York Biomedical Research Institute, University of York, Heslington, York, YO10 5DD, UK

d) ClinSpec Dx, 295 Cathedral Street, Glasgow, G1 1XL, UK

Abstract:
Vibrational spectroscopy has produced valuable information for biomedical research owing to its label-free and high-throughput capabilities. However, the complexity of and large number of variables of spectral datasets has seen the increasing application of multivariate analysis (MVA) and machine learning algorithms in recent years. In particular, the use of these techniques applied to the analysis of IR spectra of biological samples has been demonstrated as a powerful tool for the rapid sample analysis and diagnosis of disease. In this article, we review a variety of classification techniques employed for the analysis of infrared (IR) spectral datasets of biofluids, quoting prediction accuracies to demonstrate their effectiveness. With the advent of new technologies, two-dimensional infrared spectroscopy (2D-IR) has recently been applied to biomedical problems and shows potential future applications in biofluid analysis, however with multi-dimensional datasets there is a desire for advanced analytical techniques. As the application of 2D-IR to biofluids and physiological protein samples is in its infancy, large spectral datasets of biofluids suitable for classification are not readily available. Nevertheless, we draw on the classification techniques applied to IR datasets discussed in this review and relevant 2D-IR studies to discuss the future of machine learning algorithms in 2D-IR spectroscopy.

Keywords: IR Spectroscopy, 2D-IR, Biofluids, Machine Learning, Multivariate Analysis, Disease Diagnosis

1. Introduction

In recent years the application of infrared (IR) spectroscopy of biofluids has shown itself to be a valuable tool owing to its label-free, non-invasive and non-destructive qualities, as well as its sensitivity to changes at a molecular level.[1,2] Biofluids are those within the human body and include blood and its derivatives, saliva, urine and cerebral spinal fluid (CSF) and while some may be more advantageous than others depending on the disease being investigated, those that are easily obtained are often preferred by patients over those that require a needle biopsy or surgery.[3] Biofluids host a range of potentially diagnostic chemical markers, and the presence or absence of certain markers or changes in their concentrations can be induced via contact with the organs of which the fluid is associated with e.g. sputum – lungs, urine – kidneys, or by an associated response within the body.[4,5]

Current diagnostic methodologies within healthcare typically use antibody assays to detect the presence of specific biomarkers, however the heterogenous nature of disease means that the broad biomolecular detection that IR spectroscopy offers may be advantageous over single biomarker detection.[6,7] By providing a spectral fingerprint of the biofluid, IR spectroscopy can obtain information pertaining to a patient's health, and studies using this technique have demonstrated its usefulness in detecting disease[8–14] and various cancers[3,15–21] in a single, rapid and label-free measurement.

However, the large spectral datasets acquired and the large number of variables (wavenumbers) associated with spectral data make spectral assignments and therefore classification of disease challenging. For this reason, machine learning analytical methods have been employed in order to classify datasets demonstrating their high predictive power with spectral data of biological samples. In this review article we briefly discuss spectral pre-processing, reviewing the effects of applying these processes to linear absorption data when combined with classification analysis methods as well as highlighting some of the most common methods used for sample classification of biofluids using IR spectroscopy, quoting prediction accuracies to demonstrate the analytical power of each technique. Furthermore, two-dimensional infrared spectroscopy (2D-IR) has recently been used for biofluid analysis.[11,22,23] 2D-IR spectroscopy utilises ultrashort laser pulses to spread the vibrational spectrum over two frequency axes which provides a spectral map correlating excitation and detection frequencies.[24,25] The evolution of spectral features can also be monitored by altering the timing between pulses, providing additional spectral information as a function of time such as energy transfer pathways or solvation dynamics.[24,26] With these extra dimensions, come larger and more complex datasets and with advancing high-throughput technologies[27–29] spectra can be acquired in a matter of seconds, demonstrating the need for advanced techniques to examine large, complex spectral datasets.

While multivariate analysis (MVA) and machine learning (ML) algorithms have been shown to be successful at data classification with linear absorption techniques which will be discussed in this review, the use of these techniques have not yet been explored with 2D-IR for biofluid analysis. To this end, we assess the application of MVA to 2D-IR studies, of which there are only a few, and consider the use of MVA and ML algorithms towards the classification of biofluids using 2D-IR.

2. Spectral Pre-processing

Pre-processing spectral data is an integral part of multivariate analysis and is typically performed as the first analytic step to improve the results of subsequent analysis. Studies have shown a variety of different pre-processing steps that can be used with linear absorption spectra to detect and remove outliers, normalise datasets and reduce spectral noise allowing the production of a higher quality dataset thus improving the accuracy of classification or quantification techniques.[1,30–36] Figure 1 denotes the simplified spectral pathway with a non-exhaustive list of pre-processing techniques.

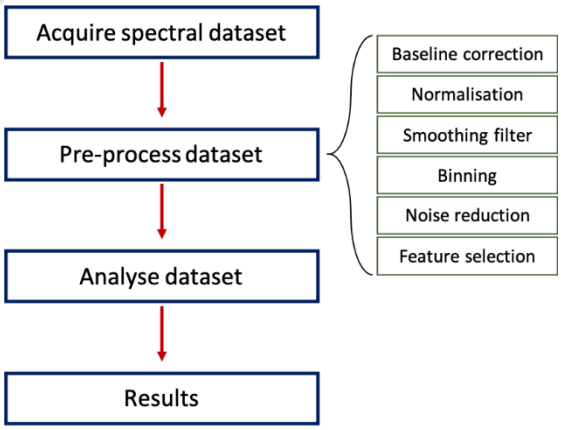

Figure 1: Schematic of the spectral dataset pathway. Initially a full dataset is acquired which is followed by the application of selected pre-processing techniques, this includes but is not limited to baseline subtraction, normalisation, binning and smoothing, noise reduction and feature selection. The dataset is then ready for analysis.

Briefly, normalisation is used to scale spectra allowing effective direct comparisons across a heterogeneous dataset, common methodologies used are min-max and vector normalisation. Baselines can often fluctuate in IR spectra owing to slight variations in sample conditions or instrumental factors and can be corrected for using baseline subtraction.[37,38] Different baseline subtraction methods exist and include subtracting a background or function from the spectrum (offset, polynomial, piecewise or rubberband) or by utilising spectral derivatives which removes broad frequency effects and in turn enhances high frequency components.[1,39] Often baseline subtraction allows spectral intensities and positions to become clearer allowing a more accurate analysis to be obtained.[36,38] Smoothing and noise filtering are common noise reduction techniques; Savitzky-Golay smoothing,[32] principal component analysis (PCA)[40] and wavelet transformations[33] have all shown success in minimising spectral noise allowing higher predictive power when later analysed using MVA.[29,30,35,41,42] For a comprehensive discussion and review of pre-processing techniques for IR spectra we refer the reader to this publication.[36]

The impact of water absorptions in biological spectra is a common topic of discussion and as such different methods have been applied to remove the water contribution from the signal.[1,11,43,44] For example, methodologies using liquid samples may require the subtraction of a background water spectrum after acquisition,[43] although this is often a subjective process and difficulties in producing optical pathlengths less than 10 um have been documented.[45] Some methodologies may involve dried droplets of biofluids whereby the water content of the fluid is decreased removing the need for water subtractions,[46–48] however drying patterns are observed which has been noted to introduce experimental difficulties. It has recently been demonstrated that no water subtractions, pre- or post-acquisition, is necessary for 2D-IR as the 2D-IR signal can circumvent the water absorption overlapping the protein amide I region. This is primarily attributed to the 4th power dependence of the vibrational transition dipole moment on the 2D-IR signal,[24] leading to the enhancement of the few protein amide I modes relative to the abundant yet weakly absorbing water molecules. This coupled with the faster vibrational lifetime of the O-H bending mode of water[11,49] allows direct measurement of the protein amide I band without the overlapping water contribution. However small optical pathlengths are necessary for this technique which can be hard to control from sample to sample.

As experimental designs and set-ups differ between IR modalities and research groups, the parameters surrounding data acquisition will also differ and so a thorough investigation of which pre-processing steps are necessary, and the order in which they should be applied, is required prior to further analysis.[30]

3. Classification Methodologies

MVA techniques have been widely applied to study linear absorption IR spectra of biofluids. In particular, the use of classification techniques with biofluids has the potential to arrange spectra within a dataset into groups, typically healthy or diseased, depending on their spectral differences and similarities.[50] The articles reviewed here have all built classification models using IR spectra of biofluids using either blind test sets or cross-validation results to test the accuracy of their models. In order to build a classification model there are 3 major steps that need to be followed.[2] Initially the dataset is split into a test and training set, and a model is built using the training data; this is known as the calibration or training phase. Based on the articles reviewed here, between 50 and 80% of the original dataset is typically selected for the training set. Next, the internal or cross validation stage utilises k-fold resampling to evaluate the model, where k is the number of groups the training data has been sub-divided into. The final step known as external validation involves the test set being classified using the model constructed from the training set, i.e., a blind testing phase.[30,51–53]

Classification methods can be split into two major groups; supervised and unsupervised. Supervised learning techniques are human guided and typically work by mapping inputs to outputs. Unsupervised

methods are given no additional information and are left to find patterns in the data. Examples of both supervised and unsupervised methods will be described in detail followed by a review of their use in biofluid analysis as well as an evaluation of how well each technique performs using sensitivity and specificity or balanced accuracy results. These parameters are commonly used when evaluating the performance of clinical tests, where the sensitivity and specificity establishes the percentage of true positive (TP) and true negative (TN) results, respectively.

$$Sensitivity = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{TN + FP}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

where FN and FP denote false negatives and false positives, respectively, and the balanced accuracy is an average of the sensitivity and specificity results. An ideal result from a classification model is one where both sensitivity, specificity and thus accuracy are 100%, where the technique does not mis-diagnose any samples from one class as the other.

3.1. Supervised Classification Techniques

Supervised classification techniques create and train a classification model with known representatives from each category involved in the dataset, e.g. healthy or diseased. This can be done in one of two ways, either by presenting example spectra of how each class is characterised or by pre-selecting information or spectral signatures pertaining to a dataset as representatives of a specific class prior to the analysis, fundamentally training the model by giving the analysis method extra information to help with the classification. Here, we focus on supervised techniques that have been commonly used for the classification of biofluid IR absorption spectra, which includes support vector machines, linear discriminant analysis, random forest and artificial neural networks.

3.1.1. Support Vector Machine

Support vector machines (SVMs) attempt to section the dataset into distinct classes, e.g. disease and healthy, by utilising the large number of variables presented by each spectrum to produce a line to separate these classes which is referred to as either the decision boundary or the hyperplane. Kernel functions are often used to map these original observations into a new dimensional space, which allows the separation between classes to become apparent. The data points from each class that lie closest to the hyperplane are known as support vectors and are the points most difficult to classify. In order to find the best hyperplane, the separation between each decision boundary and the support vectors are calculated and deemed best when this margin is maximised (Fig.2).[54,55] As well as creating separation between classes, kernel functions also help reduce the computational time taken for the calculation and can increase the classification accuracy. SVMs can be used for linear or nonlinear classification through the use of different types of kernel functions (e.g., linear, nonlinear, radial and sigmoid).

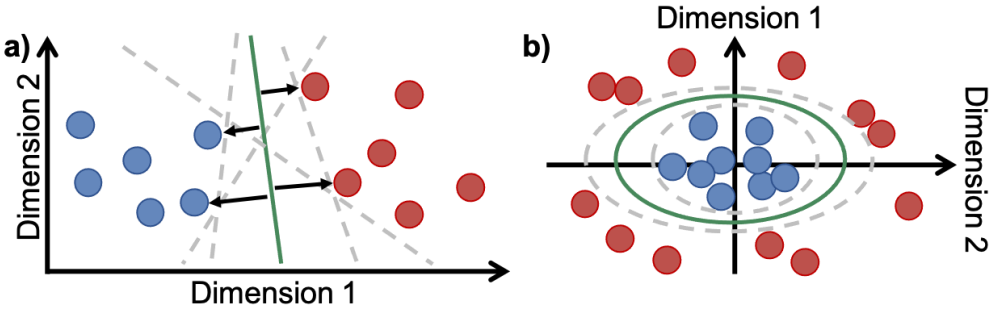

Figure 2: Schematic of the support vector machine (SVM) for both a) linear and b) radial datasets comprising two classes (red and blue circles). The optimal hyperplane is shown in green for both cases, other boundaries (grey dashed lines) are also shown however these do not maximise the margin between the two data classes. Black arrows in a) indicate the support vectors which are used to optimise the hyperplane

SVM analysis has been utilised in many studies with the aim to successfully classify biofluid samples as either diseased or healthy. The combination of SVM analysis with a large attenuated total reflection (ATR) Fourier transform IR (FTIR) spectroscopy study of serum has shown success in classifying brain cancer and non-cancer spectra.[18] A total of 724 patients were recruited and 9 spectra were acquired per serum sample. Internal validation was performed by splitting the data into test and training sets 51 times on a per patient basis and a prediction was made for each spectrum in the test set based on the training set SVM model. As there were 9 spectra for each patient, the majority result was then used as the final prediction for each patient, determining whether each sample is cancerous or healthy. Using this approach, the average sensitivity and specificity across the 51 different models was found to be 93.2% and 92.0%, respectively. Next, serum spectra from the 724 patients were used to train the SVM algorithm and 104 new patients were recruited to clinically validate the model, achieving a sensitivity and specificity of 83.3% and 87.0%, respectively. This result while notably reduced in comparison to the original study, is still above the quoted threshold of 80% required to be beneficial in healthcare settings. The authors also note that the variation of cancer types and their lack of representation in the patient training set could lead to lower sensitivity and specificity results produced from the clinical validation study, suggesting for a binary classification that each tumour type being tested must be well characterised in the training dataset.

SVM has also been utilised for classification of multi-class datasets. Hands *et al.* have shown the use of a radial basis function (RBF) kernel with SVM to predict brain tumour severity from the ATR-FTIR spectra of serum samples.[17] A total of 97 patients with high grade gliomas (50%), low grade gliomas (24%) and non-cancerous brain tumours (26%) were used and split into test and training sets containing one third and two thirds of the patients, respectively. This cross-validation process was iterated three times and an average sensitivity and specificity of 93.8% and 96.5% was achieved using this method.

Few input parameters are required to perform a SVM classification making it a favourable choice when looking to classify data compared to other techniques, in particular for users new to classification algorithms. However, the algorithm is known to underperform on noisy or large datasets,[56] where the number of variables is larger than the number of samples and so may not be the best method for classifying large patient datasets.

3.1.2. Partial-Least Squares – Discriminant Analysis

Linear discriminant analysis (LDA) is a dimensionality reduction technique that aims to find a feature subspace that maximises the separation between the mean of the two classes in a dataset whilst

minimising the spread within each class (Fig.3). To find this projected subspace, $J(w)$ the following equation is used:

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

where *m* is the mean value of each class, *s* is the spread of the data around the mean within each class and subsets 1 and 2 denote the two classes.

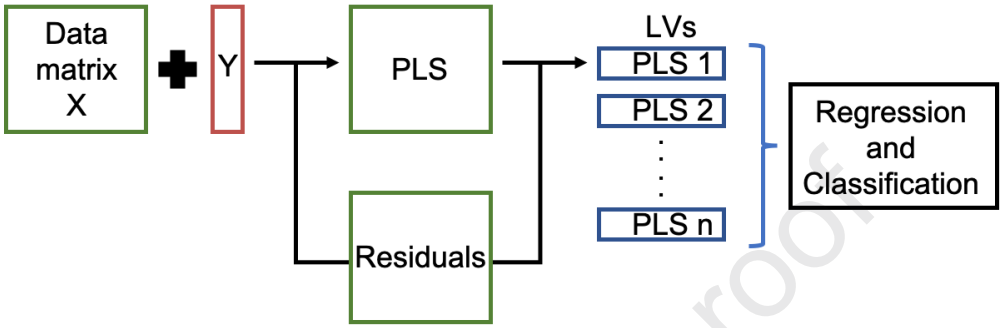

Figure 3: PLS-DA pathway showing the dataset matrix, X, and class identifier, Y, resulting in latent variables (LVs) to describe the variance between each class which can then be used for classification.

For the spectroscopic analysis of biofluids, the combination of partial least squares regression and discriminant analysis (PLS-DA) is a popular classifier.[57–62] It is sometimes also referred to as discriminant PLS (DPLS), PLS-Linear DA (PLS-LDA) or PLSDA. There are two types of PLS-DA algorithms, PLS1-DA which is commonly used for binary classification, i.e., a two-class problem and can also be used for a multi-class dataset however PLS2-DA is more suitable for multi-class problems. It is a supervised multivariate dimension reduction and classifier technique that has shown good predictabilities of classes over numerous IR biofluid studies of different classifiers.[9,19,20,63–66]

PLS-DA produces components known as latent variables (LV) which are projected into a new feature space and account for the spectral variation between the classes, where the first LV accounts for the greatest variation between the categories, the second explains the next greatest variation and so on (Fig.2(c)). PLS-DA using all the latent variables will produce the same result as a linear discriminant analysis, as no data reduction has taken place. Scores and loadings are produced by the analysis for each LV where the loadings depict where in the spectrum the variation is occurring and the scores provide coefficients for each sample spectrum. The optimum number of components needed to explain the variance between the classes and thus allow classification is found during evaluation of cross validation models across a number of test/train iterations. The ideal number of components to use will minimise the cross-validation error and provide good classification results. As PLS is a regression technique and will predict values between 0 (class 1) and 1 (class 2), decision rules need to be employed for the prediction of new samples in order to assign each spectrum to a class accurately,[58] and there are few different rules that can be employed (e.g. naïve Bayes, cut off point and boundary line).

PLS-DA has been utilised to distinguish between patients exhibiting disease and controls from analysis of high-throughput FTIR spectra of human plasma. Medipally *et al.* have shown that PLS-DA can distinguish between controls from healthy patients and samples from patients with prostate cancer with high a sensitivity (99%) and specificity (98.4%).[20] In addition, information from a scoring technique (Gleason score (GS)) used to determine how cancerous a tissue biopsy looks, was used to attempt to classify patient plasma samples further. However, this utilised less data as each GS group

contained on average 10 samples, a large reduction from the 76 patients in the whole dataset and subsequently lower sensitivities and specificities, ranging from 64.6% to 95%, were observed.

A 2014 study of ATR-FTIR spectra of human serum aimed to identify patients who were HIV positive (n=55) from those who were HIV negative (controls, n=30).[64] Of those who were HIV positive, 39 were undergoing anti-retroviral treatment (ART) while the remaining 16 were not on any treatment regimen. Using PLS-DA, 100% accuracy was achieved when categorising HIV positive patients who were not undergoing ART relative to the controls, while those undergoing ART were identified from the controls with a sensitivity and specificity of 100% and 95.2%, respectively. Furthermore, the authors classified HIV positive patients receiving no treatment and those on ART with a sensitivity and specificity of 83.3% and 100%, respectively. The authors note the drop to 83.3% may be due to an imbalance in the number of patient sera in each class.

Another study, demonstrating the high predictive power of PLS-DA, successfully categorised children and adolescents with autistic spectrum disorder (ASD, n=30) from controls (n=30) using ATR-FTIR spectra of blood serum.[63] Test and training sets used 50% of the patients from each group and achieved a prediction accuracy of 100%, which was attributed to clearly observable differences in the protein regions of the ATR-FTIR spectra.

PLS-DA is a very useful analytical tool owing to its ability to separate classes within a dataset with high predictabilities while delivering feature information, in the form of loading plots, relevant to the largest differences between the classes studied. While PLS-DA demonstrates a high predictive power, a number of parameters need to be optimised by the user in order to produce these results. Unfortunately, while performing well on two-class problems, the algorithm tends to struggle and produce poor results where a larger number of classes are involved.[58,67]

3.1.3. Random Forest

Random forest (RF) is a classifier based on a collective of decision trees and has shown promise in its analysis of data from biofluids.[43,65,68] A single tree represents a sequence of binary steps on single variables whereby in this scenario each variable is a wavenumber (Fig.4).[69,70] These decision trees are trained by iteratively making splits in the data to find the best way to group the data, this is quantified using the Gini impurity and gain calculations, which are measures that each decision tree aims to minimise and determines the probability of predicting data correctly.[70] This process is continued on subsets of data features until a Gini gain of 0 is reached, and the data is then classified. Random sampling with replacement (bootstrapping) is typically used in RF to randomly assign data into each decision tree, this means that some data may be duplicated and some will never be entered into a specific tree. Each decision tree in the RF will only test a subset of features described by the data, typically the square root of the number of features, in order to introduce randomness thus making each tree unique and ultimately improving the performance of the forest. Finally, the RF takes the majority result predicted by each tree as the final classification for each unknown spectrum.

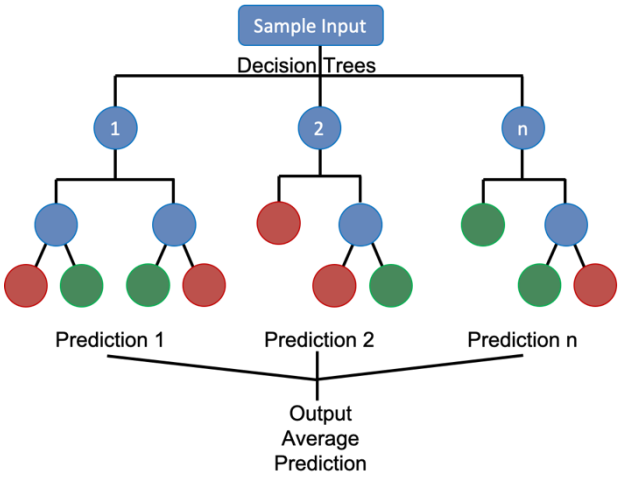

Figure 4: Visualisation of a simple Random forest (RF), with n=3 decision trees. Blue circles indicate questions posed on the data to help determine its class and red and green circles denote the class prediction.

RF with 500 different trees was used to categorise cancer and non-cancer in a study of serum samples from 433 patients; the serum samples were air-dried and then analysed using ATR-FTIR spectroscopy.[68] A 5-fold cross validation procedure was performed where 80% of the 2nd derivative spectra of patient samples were used to train the RF algorithm and the remaining 20% were used to test the model, yielding a sensitivity and specificity up to 92.8% and 91.5%, respectively. RF analysis applied to digitally dried serum spectra from 150 patients has been shown to classify brain cancer and non-cancer with a sensitivity of 93.7% and specificity of 84.0%.[43] This study demonstrates an increase in sensitivity and specificity compared to the values obtained from the analysis of the same serum samples in liquid form or air dried, providing greater accuracy in a much quicker timeframe when compared to the analysis of dried samples. Furthermore, a study by Cameron *et al.* has shown the ability of RF to classify brain tumour sub types with accuracies greater than 78%.[65] The authors note than with a more balanced dataset of tumour type, accuracies could be improved upon.

Unbalanced datasets can be a large source of error in prediction models and so 'balancing' the data can be effective in helping train a model and different sampling methods have been utilised to help eliminate bias within a model. Frequently used in prediction models are: up-sampling, which randomly re-selects data from the minority class to increase the number of samples; down-sampling, which does the opposite and randomly selects a subset of the majority class and removes the others in order to match the number of samples in the minority class; and, finally, synthetic minority over-sampling technique (SMOTE) which artificially creates 'new' samples from the minority class.[71]

RF models have the ability to handle unbalanced datasets relatively well, owing to the fact that each decision tree utilises bootstrapping and so having a class imbalance within the dataset is not as important when compared to other classification techniques, however caution should still be taken.[72] RF models have shown success in many fields and are known to handle large datasets well and arguably the biggest advantage of RF is its versatility. It can be used as a tool for both regression and classification tasks as well as being able to identify those features that are most important and responsible for the classification.[30] However RF models are computationally more complex than SVM and PLS-DA and where forests contain a large number of trees, predictions can be slow.[73]

3.1.4. Neural Networks

Advancements in artificial intelligence technologies have seen the arrival and use of neural networks (NNs) in disease diagnosis studies.[47,74–82] NNs are typically described as systems designed to operate

like the human brain and are often referred to as artificial neural networks (ANNs). Typically, NNs are mapped inputs to outputs through various 'layers' of feature extraction and calculations and there are many different types of NNs that can be used depending on the type of analysis required for a given dataset.[83]

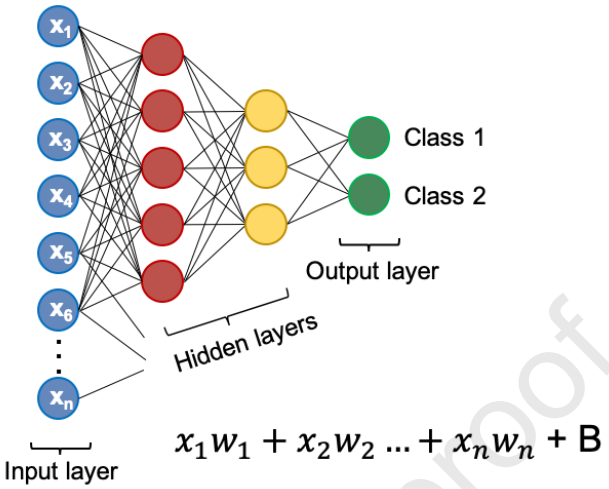

Figure 5: Representation of a neural network (NN) with three layers: input, hidden and output, along with linking channels between each layer. x_n refers to each input, w_n the weighting assigned to each neuron and B is the bias associated with each hidden layer.

Within a dataset, each wavenumber and its corresponding absorbance is used as the input layer (x_n). A mathematical function (neuron) weights (w_n) the input values (x_n) and sums all products before adding a corresponding bias (B) associated with that layer (Fig.5). This value is then sent through a threshold function to decide whether it will get passed onto a new neuron in the subsequent layer. Threshold functions known as 'activation functions' are utilised within NNs and these determine which neurons become activated and passed on to the next hidden layer, eliminating those that no longer hold useful information (Fig.5). This gradually reduces the neurons that propagate through the layers which eventually results in the production of the probability of the data being in each class. The class with the highest probability results in the final classification. These internal layers within the network are known as 'hidden layers' and this is where the calculations take place before the final classification is made in the 'output layer'. This whole process in known as 'forward propagation' and the architecture of a NN can be adjusted so that a specific number of layers and neurons exist within each layer.

In supervised learning, the NN inputs are mapped to known outputs. If the algorithm does not produce satisfactory class probabilities, the error is computed and the algorithm loops back into the network using the known classifications to adjust the weightings, producing a more accurate result. This process is known as 'backpropagation' and is iterated a number of times, improving the performance of the NN each time until it arrives at the correct classification within a given tolerance.

NNs of varying sizes have been shown to predict classes from analysis of spectra of biofluids with a high accuracy.[47,75–79,82] In a high-throughput FTIR study of human serum, NNs were used to identify breast cancer from controls from analysis of spectra acquired using two measurement modalities, transflection and transmission.[76] Samples from a total of 193 patients were analysed and the data were split into three groups for training, validation and testing. The architecture of the network was investigated and included an input layer consisting of 15 - 200 inputs, 1 - 20 different hidden layers and 2 outputs corresponding to the two classes: breast cancer or control. The optimum set-up was found to be 46 inputs and 2 hidden layers for transmission mode, which produced identical sensitivity

and specificity of 95%. In transflection mode, the highest sensitivity (100%) and specificity (92%) were recorded using a network with 43 inputs and 3 hidden layers. Furthermore, the breast cancer patient sera were compared against 11 other diseases, including sepsis, Alzheimer's disease and other cancers, with a total of 3119 patients, and using NNs an average accuracy of 91% was found across all 12 classes.

Combining ATR-FTIR spectroscopy of the saliva from type 2 diabetic patients with NNs has allowed patients with the disease 'controlled' and 'uncontrolled' to be identified.[77] Using inputs from the 2000 - 4000 cm^{-1} region, 1 hidden layer with only one neuron, a 100% accuracy was obtained in determining controlled from uncontrolled diabetic patients. Additionally, patients were divided into 3 sub groups based on their blood glucose levels and the NNs again classified each patient serum with 100% accuracy based on a leave one out cross validation.

In principle, the more data provided to a neural network the higher predictive power it will have, however often neural networks are quoted to require very large amounts of input data in order to produce high prediction accuracies.[74] The use of backpropagation in the NN framework gives the model a high degree of adaptability, allowing itself to fix any mistakes, an advantage over other classification models. However, the complexity of the network and amount of data provided can have a huge impact on the amount of computational power required and successful training of deep learning algorithms can take days or even weeks/months to compute. The major disadvantage of NNs however is the lack of interpretability. The 'black box' approach of NNs doesn't allow the user to understand how the network has made its prediction, which depending on the aim of the experiment may be problematic.[83,84]

Overall supervised learning has been shown to yield disease classification with high accuracies, and works are currently underway to compare different disease states in order to hone in on disease classification. However, supervised techniques are not always suitable, in particular during exploratory analysis where differences between classes are unknown and this is where unsupervised techniques may become useful. Often the results from an unsupervised method indicate particular features important for classification and can be used as a precursor to a supervised method, essentially as a tool for feature extraction. Unsupervised techniques may be more appropriate when the dataset is large, but not large enough for NN and so the additional work required to input each sample class into the chosen algorithm may be impractical.

3.2. Unsupervised Classification Techniques

Unsupervised methods are especially useful when the spectral differences between groups are unknown or if a dataset is particularly large as inputting class information can be time-consuming. The variability of healthy biofluids is vast let alone when considering disease states which can make classification of controls and different disease states especially difficult. With this in mind, knowing the class of a patient sample does not necessarily provide enough information for a machine learning algorithm to define classes as there may be strong similarities and overlaps between the categories being studied. As the class is not known to the unsupervised algorithm, typically they tend to underperform when compared to supervised methods. Nevertheless, for larger datasets their lack of supervision can be advantageous. Some of the most common unsupervised techniques used to analyse biofluids are clustering and principal component analysis which have seen increasing use in recent years.[85–89]

3.2.1. Clustering

Clustering is the process of finding structures or patterns within a dataset, which helps to define natural groups or clusters whose members are similar allowing identification of data classes.[87,88] Essentially clustering provides segregation of data into groups which contain similar data within a

cluster and data that is dissimilar assigned to other clusters. There are a few basic clustering methods used in data analytics, however the most commonly used in IR spectroscopy of biofluids are k-means clustering and hierarchical clustering.

K-means clustering (KMC) sorts the data into k clusters, where k is the number of classes expected to be observed within the dataset, and aims to sort each individual spectrum into 1 of k clusters.[87,88] By assigning k random datapoints as cluster centres (the number of clusters must be user defined), the Euclidean distance is then calculated between each datapoint and the k cluster centres. Each datapoint is then assigned to its nearest cluster centre (Fig.6(a)). Next, the average value of each data point within each cluster is calculated allowing re-centring of the cluster centres. This process is repeated until the spread within each cluster can no longer be reduced with each iteration. This method is then repeated using new random datapoints assigned as new cluster centres. This is done a user defined number of times and the most common result is then used for classification.

For hierarchical clustering the algorithm provides a ladder of clusters, typically displayed via a dendrogram (Fig.6(b)). One of the advantages of this method is that, unlike KMC, the number of classes does not have to be defined prior to starting the clustering algorithm.[90,91] Hierarchical clustering can work in one of two ways: top-down or bottom-up. In the bottom-up approach, all datapoints belong to their own cluster, two clusters that are similar are then merged and this process continues until all datapoints are assigned into one cluster. The opposite is done for the top-down approach, using differences in the dataset to split into clusters. Many different methods can be used to measure the cluster similarity and merge the clusters together, typically the Euclidean distance is used, however other methods have been documented for IR analysis of biofluids.[92–95] Using the dendrogram, the optimum number of clusters for the dataset can be identified.

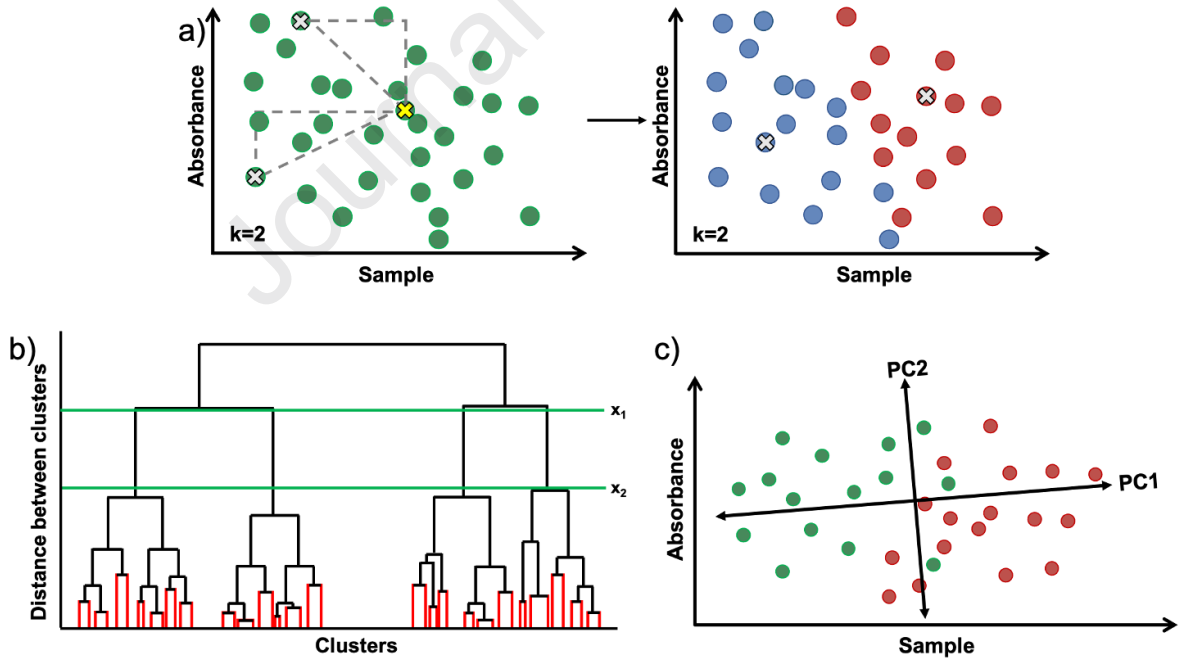

Figure 6: Illustration of unsupervised classification techniques. a) Representation of k-means clustering, where two random datapoints are assigned as cluster centres (white crosses) and the Euclidean distance is calculated between each datapoint and the cluster centres. The datapoint is then assigned to the closest cluster (shown for datapoint denoted with yellow cross). In theory, after a number of iterations, cluster centres will be in the centre of each class (red and blue datapoints). b) A hierarchical clustering dendrogram. Initial individual data clusters are shown in red. Horizontal green lines show the maximum distance between a given number of clusters, indicating the optimum number of clusters for a given dataset. c) PCA showing the first two principal components explaining the 2 largest directions of variation.

Both clustering methods have shown successful identification of biofluid classes using IR spectrosopy.[47,76,89,95–97] Using 193 patient sera from breast cancer patients and controls, Backhaus *et al.* assessed the predictability of KMC analysis, where k=2, using two different modalities of FTIR; transflection and transmission.[76] The reported sensitivity and specificity for transmission mode are 96% and 93%, and KMC analysis in transflection mode yielded slightly better results of 98% and 95%, respectively. The results quoted for KMC are comparable to those obtained from the application of artificial neural networks to the same dataset.

ATR-FTIR spectra of bladder wash collected during a cystoscopy were analysed using hierarchical cluster analysis. Identical sensitivities and specificities of 81.8% and 80.9% were obtained, irrespective of the spectral region used, for determination of bladder cancer from controls.[89] Poorer results were obtained from HCA of IR spectra acquired in transmission mode of the same samples. Utilising KMC as a classification method is relatively simple and is easy to implement owing to the single input parameter required, k. It also has several parameters that can be modified to adjust the performance of the algorithm. Due to the use of the Euclidean distance in KMC, the cluster centres are highly dependent on the selection of the initial cluster centres.[98] Both KMC and HCA are sensitive to noisy data and outliers, which can have a big impact on their predictability. Adaptations can be applied to the algorithm to help overcome this, however this requires a more in-depth knowledge of the technique. Hierarchical clustering does not require apriori information about the number of clusters and produces a visual map (dendrogram) showing how the data can be separated. One of the major drawbacks of hierarchical clustering is the time taken to compute meaning that it is rarely suitable for larger datasets,[99] however the dataset can be reduced using principal component analysis to reduce dimensionality before the application of HCA.[89]

3.2.2. Principal Component Analysis

Principal component analysis (PCA) aims to maximise the variance within the dataset as a whole, and does not have any prior knowledge of data classes (Fig.6(c)).[40,100] PCA reduces the dimensionality of the data by geometrically projecting the dataset onto fewer dimensions, known as principal components (PCs), establishing a new co-ordinate system where the largest variance is described by the first principal component (PC1), second largest variance described by the second PC (PC2) and so on.[100,101] The data matrix is broken down into two components, scores and loadings. Loadings identify which variables are most strongly associated with each PC, identifying sources of variation within the dataset, and scores denote the value of a sample when projected (orthogonally) onto a PC. PCA has seen increasing use for the analysis of IR spectra of biofluids, in particular where datasets are particularly large and spectral markers used to discriminate between classes are unknown.[13,17,20,41,89,92,93,102]

An FTIR study of plasma samples from patients with varying stages of prostate cancer and controls, displayed clear discrimination between the two classes when the data were analysed using PCA.[20] Moreover, when plotting PC1 vs. PC2 the PCA algorithm was able to separate patients with prostate cancer into clusters based on cancer stage and further proved that the age of the patients studied was not the source for the variations observed.

In the bladder wash study, which was described above in section 3.2.1, the data obtained were also analysed using PCA and showed successful results towards identifying bladder cancer patients from controls.[89] PCA was applied to data acquired using two different measurement modalities, transmission FTIR (with KBr pellets) and ATR-FTIR (with dried wash samples), and the scores of PC1 vs. PC2 displayed clear discrimination between samples from the cancer patients and the controls. The PC loadings were then used to identify the spectral features associated with largest variance in the dataset which were passed on to the cluster analysis for further evaluation. PCA is often used as

a feature extraction technique and has seen success when combined with supervised classification techniques.[41,103–105] In doing so, the dimensionality of the data is reduced allowing computation of discrimination between classes to be faster and based on real spectral differences within the dataset as opposed to pre-assigning the spectral differences between classes.[106]

One of the key advantages of PCA is visualisation of important features within a dataset provided by the scores and loadings plots, allowing interpretation of the differences within the dataset being analysed.[40] The ability of PCA to handle noisy datasets is another key advantage, by reducing the dimensionality of the dataset, the variations due to noise can be ignored more easily. PCA is strongly influenced by the scale of the data and so if other data besides spectroscopic are being investigated (e.g., protein levels) then the data must be standardised, and any categorical features must be converted into numerical values before applying PCA. However PCA alone is typically not a strong enough classifier and is often used a feature selection technique prior to classification analysis.[86,104,105,107]

4. 2D-IR Spectroscopy

2D-IR spectroscopy has been used to probe the molecular vibrations of many systems including biological molecules and solvents and has provided visualisations of conformational structural changes and dynamics of molecules.[25,29,108–116] Proof of principle studies have shown the power and sensitivity of 2D-IR as it observes molecular structures and changes within a system allowing identification of proteins, ligands, DNA and other markers, and it has the potential to aid in disease diagnosis.[11,29,109,112,116–121] However for the use of 2D-IR to study biofluids it is in its infancy and requires development which faces many challenges.

As the application of 2D-IR to the study of protein amide I modes in biofluids is in the initial exploratory stages, the application of machine learning algorithms to 2D-IR datasets is small and classification of disease in biofluids has yet to be demonstrated. In order to assess the potential for biofluid classification using 2D-IR it is important to consider MVA techniques applied to 2D-IR datasets and how they compare with other methods.

It was recently demonstrated that the 2D-IR amide I signature of proteins dominates that of water even at sub-millimolar protein concentrations, in contrast to IR absorption experiments, allowing the quantification of proteins in serum without the overlapping water absorptions. [11] The 2D-IR spectral lineshapes have been shown to be extremely sensitive to protein secondary structure, allowing the identification of individual proteins in serum, and the ratio of albumin to globulin proteins was determined in a model serum dataset with accuracies <4%. Contributions due to structurally similar individual immunoglobin sub-groups were also identified due to their unique spectral lineshapes. Comparisons of the same samples analysed with FTIR spectroscopy show that this information is not easily retrievable from FTIR analysis.[11]

PCA has been shown to be a powerful tool for use with 2D-IR spectroscopy and has enabled extraction of information from 2D-IR datasets.[23,29,118] It is important to note that prior to fit using a bilinear model such as PCA, the data must be first altered into a suitable format, ideally where each 2D matrix is reshaped into a single dimensional vector. In a 2D-IR serum study, the concentration of supplemented glycine in serum was evaluated using PCA.[23] Notably, the spectral information contained in the first two principal components of the PCA loading plots produced results similar to individual 2D-IR spectra of neat serum and glycine in H_2O, respectively. Comparisons of the PCA result with transmission mode FTIR data showed that while the PCA score data gave similar linear correlations with glycine concentrations for the first two components, it was unable to split the two major contributions (serum and glycine) into individual loadings. PCA was also used to help establish a detection limit using the amino acid glycine which allowed projection to potential sensitivities for larger peptides/proteins of

~200 µM. It is suggested that with further investigation the use of PCA with 2D-IR spectroscopy has the potential to produce different principal components corresponding to the peptides or proteins causing changes from that of a healthy serum sample, whereby in an ideal scenario producing a library of 2D-IR spectra. With PCA being an unsupervised analytical method, the differences within each sample do not have to be highlighted to the algorithm. In this scenario, this is highly advantageous over supervised analytical methods as the composition of biofluids infamously fluctuates depending on many factors including age, diet, stress, and disease.[97,122–125]

Biofluids primarily contain proteins, not including water, however there are only a few 2D-IR protein studies that have been conducted in H_2O. There are many evaluations of proteins in D_2O solutions which have allowed identification of protein secondary structures producing results similar to those achieved from more conventional techniques,[116,118,120,121] and so it is insightful to evaluate the use of MVA techniques for 2D-IR disease diagnostics on non-physiological protein studies.

The first application of PCA with 2D-IR spectroscopy shows the ability to quantify changes in the protein secondary structure of the messenger protein calmodulin associated with temperature and calcium binding.[118] Two principal components were revealed, PC1 and PC2, and found to be independent and dependent on temperature, respectively. Furthermore, the loading information of PC2 was found to match the features found in difference spectra from the highest and lowest temperature spectra. The calmodulin 2D-IR spectra in its calcium free or bound form differed enough to allow identification of each system. This binding is reported to affect only 4.7% of the protein's residues, showcasing the sensitivity of 2D-IR spectroscopy.[118] This result also compared positively to secondary structure assignments achieved via circular dichroism spectroscopy.

Additionally, the use of PCA combined with analysis of variance (ANOVA-PCA) has shown success in categorising ligand binding to DNA sequences using 2016 different 2D-IR spectra.[29] Due to the high information content of the 2D-IR spectral dataset, Fritzsch *et al.* have shown the ability of this technique to retrieve the base composition of the 12 different DNA sequences used and detect the presence of a bound ligand, as each DNA base pair produces a distinctive 2D-IR peak pattern. As the understanding of base couplings improve, determination of DNA sequences may be possible via large libraries of data and MVA analysis. Among the analytical advancements making 2D-IR suitable for biomedical research, this study in particular showcases the current state of the art in laser technologies, demonstrating rapid collection of a 2D-IR spectrum in seconds showing that a high-throughput screening application of 2D-IR coupled with the categorisation of a large number of samples is possible.

Utilising singular value decomposition (SVD), Baiz *et al.* constructed a library of 2D-IR spectra from 16 different proteins with known crystal structures.[116] SVD is a matrix diagonalisation technique used to describe a dataset in terms of principal component spectra. By utilising the specific signatures produced by each secondary structure component, e.g. α-helix and β-sheet, the SVD algorithm predicted the fraction of each structure contained in a single protein sample, negating the need for complex and subjective curve fitting and deconvolutions. A leave-one-out process was performed and using the remaining 15 protein spectra, the percentage of α-helix, β-sheet and unassigned protein conformations were predicted. The authors compare their results and errors using circular dichroism spectroscopy, the current standard for protein secondary structure, and note similar errors for each structural element between the two methods. The authors also comment that a larger library of proteins would likely result in a decrease in the associated error. The power and sensitivity of MVA combined with 2D-IR protein spectral datasets is encouraging, alluding to its potential high value for their use in 2D-IR disease diagnostics.

For current investigations of biofluids using IR spectroscopy, measurement, pre-analytical and analytical protocols are often varied between different research groups and no single protocol exists despite efforts to standardise these. While the application of 2D-IR to biofluids is still in its initial stages, it would be prudent to learn from this community and take this opportunity to establish a protocol for the use of spectral pre-processing and multivariate analysis methods to help eliminate the interoperability of these processes. Initially thorough investigation of pre-analytical techniques on a 'basic' 2D-IR biofluid dataset should be performed. The methods used will likely be similar to those already utilised in the IR spectroscopy community however the effects on a multi-dimensional dataset may require additional study, in particular the order in which they are applied. Multi-dimensional datasets often require manipulation into a suitable form prior to application of analysis techniques, for example a single 2D-IR spectrum exists as a 2D matrix requiring reshaping into a single dimensional array (similar to that of a FTIR spectrum). However, techniques have been developed for application with data cubes (for example with 2D nuclear magnetic resonance and mass spectrometry) which could be useful for 2D-IR datasets. Once a pre-analytical protocol has been established, the application of different multivariate analysis techniques can be tested. In order to standardise these procedures, we need to be aware of both the clinical needs and instrumental capabilities; what information is clinically relevant and can reproducible results be obtained quickly. This will require a rigorous study on the limitations of each process and determination of standards that should be adhered to in order to use these advanced analytical techniques to produce fair and comprehensive results. Realistically, standardised procedures will only be accepted when they are shown to be successful time and time again, establishing the need for many studies and their careful evaluation, a slow yet necessary development.

As with many detection technologies, the shortcoming of 2D-IR lies with sensitivity and detection limits, and generally speaking 2D-IR is not considered a high-sensitivity technique. However in recent years, developments in laser technologies have seen the acquisition of a single 2D-IR spectrum in seconds, increased signal-to-noise levels resulting in lower detection sensitives and new methodologies have been shown to minimise measurement to measurement variability.[22,27,28,126–130] Further developments in technology, sample handling and microfluidics or even surface enhancements could lead to the potential increased sensitivity of biomolecule concentrations. This coupled with the ability of 2D-IR to circumvent water absorptions and the minimal sample preparation required, presents 2D-IR as a complementary method to IR absorption spectroscopy in the biomedical arena.

5. Conclusion

The studies discussed in this review highlight the application and power of machine learning classification algorithms towards the analysis of IR spectra of biofluids. However, a common remark posed by authors in the studies mentioned express the need for larger datasets to encompass a more balanced dataset, i.e., a similar number of patients in each tested class, to help increase test accuracies. Additionally, numerous large-scale studies would help steer vibrational spectroscopy towards mainstream healthcare. With the arrival of 2D-IR spectroscopy to biofluid analysis, the need to establish analytical protocols while in its infancy is vital, and the need for strong analytical tools is essential to deal with the large and complex datasets that 2D-IR possesses. Proof-of-principle 2D-IR protein studies combined with MVA have shown great potential for the analysis of biofluids with its sensitivity to small molecular changes in large systems, and the potential to build a large spectral library which could be used to evaluate unknown samples has been recognized. The multidimensionality of a 2D-IR spectrum and thus the wealth of information gives 2D-IR a strong advantage over IR absorption methods and with this in mind the use of classification algorithms with 2D-IR biofluid spectral datasets has an encouraging potential.

Declaration of competing interest

Matthew J. Baker is a director of ClinSpec Diagnostics Ltd.

References

1. Baker, M. J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **9**, 1771–1791 (2014).
2. Baker, M. J., Sockalingum, G. D., Hughes, C. & Lukaszewski, R. A. Developing and Understanding Biofluid Vibrational Spectroscopy: a Critical Review. *Chem. Soc. Rev.* **45**, 1803–1818 (2016).
3. Hands, J. R. *et al.* Brain tumour differentiation: rapid stratified serum diagnostics via attenuated total reflection Fourier-transform infrared spectroscopy. *J. Neurooncol.* **127**, 463–472 (2016).
4. Hu, S., Loo, J. & Wong, D. Human body fluid proteome analysis. *Proteomics* vol. 6 6326–6353 (2006).
5. Li, Y. & Bahassi, E. M. Biofluid-Based Circulating Tumor Molecules as Diagnostic Tools for Use in Personalized Medicine. *J. Mol. Biomarkers Diagnosis* **5**, 157 (2013).
6. Wallstrom, G., Anderson, K. S. & Labaer, J. Biomarker discovery for heterogeneous diseases. *Cancer Epidemiol. Biomarkers Prev.* **22**, 747–755 (2013).
7. Considine, E. C. The search for clinically useful biomarkers of complex disease: A data analysis perspective. *Metabolites* **9**, (2019).
8. Petrich, W. *et al.* Potential of mid-infrared spectroscopy to aid the triage of patients with acute chest pain. *Analyst* **134**, 1092–1098 (2009).
9. Roy, S. *et al.* Simultaneous ATR-FTIR Based Determination of Malaria Parasitemia, Glucose and Urea in Whole Blood Dried onto a Glass Slide. *Anal. Chem.* **89**, 5238–5245 (2017).
10. Rohleder, D. *et al.* Comparison of mid-infrared and Raman spectroscopy in the quantitative analysis of serum. *J. Biomed. Opt.* **10**, 031108 (2005).
11. Hume, S. *et al.* Measuring Proteins in H2O with 2D-IR Spectroscopy. *Chem. Sci.* **10**, 6448–6456 (2019).
12. Spalding, K. *et al.* Enabling quantification of protein concentration in human serum biopsies using attenuated total reflectance – Fourier transform infrared (ATR-FTIR) spectroscopy. *Vib. Spectrosc.* **99**, 50–58 (2018).
13. Bonnier, F. *et al.* Screening the low molecular weight fraction of human serum using ATR-IR spectroscopy. *J. Biophotonics* **9**, 1085–1097 (2016).
14. Bonnier, F. *et al.* Ultra-filtration of human serum for improved quantitative analysis of low molecular weight biomarkers using ATR-IR spectroscopy. *Analyst* **142**, 1285–1298 (2017).
15. Gajjar, K. *et al.* Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. *Analyst* **138**, 3917–3926 (2013).
16. Cameron, J. M. *et al.* Developing infrared spectroscopic detection for stratifying brain tumour patients: glioblastoma multiforme vs . lymphoma. *Analyst* **144**, 6736–6750 (2019).
17. Hands, J. R. *et al.* Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectral discrimination of brain tumour severity from serum samples. *J. Biophotonics* **7**, 189–199 (2014).
18. Butler, H. J. *et al.* Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. *Nat. Commun.* **10**, 4501 (2019).
19. Medipally, D. K. R. *et al.* Monitoring Radiotherapeutic response in prostate cancer patients using high throughput FTIR Spectroscopy of Liquid Biopsies. *Cancers (Basel).* **11**, 1–18 (2019).
20. Medipally, D. K. R. *et al.* Vibrational spectroscopy of liquid biopsies for prostate cancer diagnosis. *Ther. Adv. Med. Oncol.* **12**, 1–23 (2020).
21. Sheng, D. *et al.* Comparison of serum from gastric cancer patients and from healthy persons using FTIR spectroscopy. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **116**, 365–369 (2013).
22. Hume, S. *et al.* 2D-Infrared Spectroscopy of Proteins in Water: Using the Solvent Thermal Response as an Internal Standard. *Anal. Chem.* **92**, 3463–3469 (2020).
23. Rutherford, S. H. *et al.* Detection of Glycine as a Model Protein in Blood Serum Using 2D-IR Spectroscopy. *Anal. Chem.* (2020) doi:10.1021/acs.analchem.0c03567.
24. Hamm, P. & Zanni, M. T. *Concepts and Methods of 2D Infrared Spectroscopy*. *Journal of Experimental Psychology: General* (Cambridge University Press, 2011).
25. Hamm, P., Lim, M. & Hochstrasser, R. M. Structure of the Amide I Band of Peptides Measured by Femtosecond Nonlinear-Infrared Spectroscopy. *J. Phys. Chem. B* **102**, 6123–6138 (1998).

26. Zanni, M. T. & Hochstrasser, R. M. Two-dimensional infrared spectroscopy: A promising new method for the time resolution of structures. *Curr. Opin. Struct. Biol.* **11**, 516–522 (2001).
27. Donaldson, P. M., Greetham, G. M., Shaw, D. J., Parker, A. W. & Towrie, M. A 100 kHz Pulse Shaping 2D-IR Spectrometer Based on Dual Yb:KGW Amplifiers. *J. Phys. Chem. A* **122**, 780–787 (2018).
28. Greetham, G. M. *et al.* A 100 kHz Time-Resolved Multiple-Probe Femtosecond to Second Infrared Absorption Spectrometer. *Appl. Spectrosc.* **70**, 645–653 (2016).
29. Fritzsch, R. *et al.* Rapid Screening of DNA-Ligand Complexes via 2D-IR Spectroscopy and ANOVA-PCA. *Anal. Chem.* **90**, 2732–2740 (2018).
30. Smith, B. R., Baker, M. J. & Palmer, D. S. PRFFECT: A versatile tool for spectroscopists. *Chemom. Intell. Lab. Syst.* **172**, 33–42 (2018).
31. Rinnan, Å., Berg, F. van den & Engelsen, S. B. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC - Trends Anal. Chem.* **28**, 1201–1222 (2009).
32. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
33. Alsberg, B. K., Woodward, A. M. & Kell, D. B. An introduction to wavelet transforms for chemometricians: A time- frequency approach. *Chemom. Intell. Lab. Syst.* **37**, 215–239 (1997).
34. Surewicz, W. K., Mantsch, H. H. & Chapman, D. Determination of Protein Secondary Structure by Fourier Transform Infrared Spectroscopy: A Critical Assessment. *Biochemistry* **32**, 389–394 (1993).
35. Mallet, Y., Coomans, D. & De Vel, O. Recent developments in discriminant analysis on high dimensional spectral data. *Chemom. Intell. Lab. Syst.* **35**, 157–173 (1996).
36. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.* **117**, 100–114 (2012).
37. Schulze, G. *et al.* Investigation of selected baseline removal techniques as candidates for automated implementation. *Appl. Spectrosc.* **59**, 545–574 (2005).
38. Mazet, V., Carteret, C., Brie, D., Idier, J. & Humbert, B. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemom. Intell. Lab. Syst.* **76**, 121–133 (2005).
39. Holler, F., Burns, D. H. & Callis, J. B. Direct use of second derivatives in curve-fitting procedures, Applied Spectroscopy. *Appl. Spectrosc.* **43**, 977–882 (1989).
40. Bro, R. & Smilde, A. K. Principal component analysis. *Anal. Methods* **6**, 2812–2831 (2014).
41. Lovergne, L. *et al.* Investigating optimum sample preparation for infrared spectroscopic serum diagnostics. *Anal. Methods* **7**, 7140–7149 (2015).
42. Jernelv, I. L., Hjelme, D. R., Matsuura, Y. & Aksnes, A. Convolutional neural networks for classification and regression analysis of one-dimensional spectral data. (2020).
43. Sala, A. *et al.* Rapid analysis of disease state in liquid human serum combining infrared spectroscopy and "digital drying". *J. Biophotonics* 1–10 (2020) doi:10.1002/jbio.202000118.
44. Cameron, J. M., Butler, H. J., Palmer, D. S. & Baker, M. J. Biofluid spectroscopic disease diagnostics : A review on the processes and spectral impact of drying. *J. Biophotonics* (2018) doi:10.1002/jbio.201700299.
45. Lasch, P. & Kneipp, J. *Biomedical Vibrational Spectroscopy*. (Wiley-Interscience, 2010).
46. Ollesch, J. *et al.* It's in your blood: Spectral biomarker candidates for urinary bladder cancer from automated FTIR spectroscopy. *J. Biophotonics* **7**, 210–221 (2014).
47. Haas, S. L. *et al.* Spectroscopic diagnosis of myocardial infarction and heart failure by fourier transform infrared spectroscopy in serum samples. *Appl. Spectrosc.* **64**, 262–267 (2010).
48. Cameron, J. M., Butler, H. J., Palmer, D. S. & Baker, M. J. Biofluid spectroscopic disease diagnostics : A review on the processes and spectral impact of drying. *J. Biophotonics* **11**, 1–12 (2018).
49. Huse, N., Ashihara, S., Nibbering, E. T. J. & Elsaesser, T. Ultrafast vibrational relaxation of O – H bending and librational excitations in liquid H 2 O. vol. 404 389–393 (2005).
50. Wilks, D. S. Discrimination and Classification. in *Statistical Methods in the Atmospheric Sciences* (ed. Wilks, D. S.) vol. 100 583–602 (Elsevier, 2011).
51. Huang, Y., Li, W., Macheret, F., Gabriel, R. A. & Ohno-Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Informatics Assoc.* **27**, 621–633 (2020).
52. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).
53. Baker, M. J. *et al.* Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem. Soc. Rev.* **45**, 1803–1818 (2016).
54. Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*

2, 121–167 (1998).
55. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).
56. Brereton, R. G. & Lloyd, G. R. Support Vector Machines for classification and regression. *Analyst* **135**, 230–267 (2010).
57. Brereton, R. G. & Lloyd, G. R. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* **28**, 213–225 (2014).
58. Lee, L. C., Liong, C. Y. & Jemain, A. A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategies and knowledge gaps. *Analyst* **143**, 3526–3539 (2018).
59. Mehmood, T. & Ahmed, B. The diversity in the applications of partial least squares: An overview. *J. Chemom.* **30**, 4–17 (2016).
60. Ballabio, D. & Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Anal. Methods* **5**, 3790–3798 (2013).
61. Barker, M. & Rayens, W. Partial least squares for discrimination. *J. Chemom.* **17**, 166–173 (2003).
62. Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* **879**, 10–23 (2015).
63. Ildiz, G. O., Bayari, S., Karadag, A. & Kaygisiz, E. Complementary Diagnosis Tool for Autism Spectrum Disorder in Children and Adolescents. *Molecules* **25**, 2079–2091 (2020).
64. Sitole, L., Steffens, F., Krüger, T. P. J. & Meyer, D. Mid-ATR-FTIR spectroscopic profiling of HIV/AIDS sera for novel systems diagnostics in global health. *Omi. A J. Integr. Biol.* **18**, 513–523 (2014).
65. Cameron, J. M. *et al.* Stratifying brain tumour histological sub-types: The application of ATR-FTIR serum spectroscopy in secondary care. *Cancers (Basel).* **12**, 1–16 (2020).
66. Paraskevaidi, M. *et al.* Potential of mid-infrared spectroscopy as a non-invasive diagnostic test in urine for endometrial or ovarian cancer. *Analyst* **143**, 3156–3163 (2018).
67. Ruiz-Perez, D. & Narasimhan, G. So you think you can PLS-DA? *bioRxiv* (2017) doi:10.1101/207225.
68. Smith, B. R. *et al.* Combining random forest and 2D correlation analysis to identify serum spectral signatures for neuro-oncology. *Analyst* **141**, 3668–3678 (2016).
69. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
70. Cutler, A., Cutler, R. D. & Stevens, J. R. Machine Learning. in *Ensemble Machine Learning: Methods and Applications* (eds. Zhang, C. & Ma, Y.) (Springer, 2012). doi:10.1007/978-1-4419-9326-7.
71. Chawla, N. V, Bowyer, K. W., Hall, L. O. & Kegelmeyer, P. W. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
72. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **26**, 217–222 (2005).
73. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23**, 2507–2517 (2007).
74. Shahid, N., Rappon, T. & Berta, W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS One* **14**, e0212356 (2019).
75. Thomas, N., Goodacre, R., Timmins, É. M., Gaudoin, M. & Fleming, R. Fourier transform infrared spectroscopy of follicular fluids from large and small antral follicles. *Hum. Reprod.* **15**, 1667–1671 (2000).
76. Backhaus, J. *et al.* Vibrational Spectroscopy Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vib. Spectrosc.* **52**, 173–177 (2010).
77. Sánchez-Brito, M. *et al.* A Machine-Learning Strategy to Evaluate the Use of FTIR Spectra of Saliva for a Good Control of Type 2 Diabetes. *Talanta* **221**, 121650 (2021).
78. Lux, A. *et al.* HHT diagnosis by Mid-infrared spectroscopy and artificial neural network analysis. *Orphanet J. Rare Dis.* **8**, 1–15 (2013).
79. Peters, A. S. *et al.* Serum-infrared spectroscopy is suitable for diagnosis of atherosclerosis and its clinical manifestations. *Vib. Spectrosc.* **92**, 20–26 (2017).
80. Rzaei-tavirani, M., Hassan, S. & Ranjbar, B. The Effects of Acetaminophen on Human Serum Albumin (HSA). *Iran. J. Pharm. Res.* **4**, 239–244 (2005).
81. Bacsik, Z., Mink, J. & Keresztury, G. FTIR spectroscopy of the atmosphere Part 2. Applications. *Appl. Spectrosc. Rev.* **40**, 327–390 (2005).
82. Ahmed, S. S. S. J., Santosh, W., Kumar, S. & Thanka Christlet, T. H. Neural network algorithm for the early detection of Parkinson's disease from blood plasma by FTIR micro-spectroscopy. *Vib. Spectrosc.* **53**, 181–188 (2010).
83. Almeida, J. S. Predictive non-linear modeling of complex data by artificial neural networks. *Curr. Opin.*

Biotechnol. **13**, 72–76 (2002).
84. Tu, J. V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **49**, 1225–1231 (1996).
85. Custers, D. *et al.* ATR-FTIR spectroscopy and chemometrics: An interesting tool to discriminate and characterize counterfeit medicines. *J. Pharm. Biomed. Anal.* **112**, 181–189 (2015).
86. Zlotogorski-Hurvitz, A., Dekel, B. Z., Malonek, D., Yahalom, R. & Vered, M. FTIR-based spectrum of salivary exosomes coupled with computational-aided discriminating analysis in the diagnosis of oral cancer. *J. Cancer Res. Clin. Oncol.* **145**, 685–694 (2019).
87. Romesburg, C. *Cluster Analysis for Researchers*. (Lulu Press, 2004).
88. Kaufman, L. & Rousseuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis.* (John Wiley & Sons, 2005). doi:10.2307/2532178.
89. Gok, S. *et al.* Bladder cancer diagnosis from bladder wash by Fourier transform infrared spectroscopy as a novel test for tumor recurrence. *J. Biophotonics* **9**, 967–975 (2016).
90. Revelle, W. Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behav. Res.* **14**, 57–74 (1979).
91. Bridges, C. C. Hierarchical cluster analysis. *Physiol. Rep.* **18**, 851–854 (1966).
92. Lovergne, L. *et al.* Biofluid infrared spectro-diagnostics: pre-analytical considerations for clinical applications. *Faraday Discuss.* **187**, 521–537 (2016).
93. Lewis, P. D. *et al.* Evaluation of FTIR spectroscopy as a diagnostic tool for lung cancer using sputum. *BMC Cancer* **10**, 640–650 (2010).
94. Caixeta, D. C. *et al.* Salivary molecular spectroscopy : A sustainable , rapid and non-invasive monitoring tool for diabetes mellitus during insulin treatment. *PLoS One* **15**, e0223461 (2020).
95. Takamura, A., Watanabe, K., Akutsu, T. & Ozawa, T. Soft and Robust Identification of Body Fluid Using Fourier Transform Infrared Spectroscopy and Chemometric Strategies for Forensic Analysis. *Sci. Rep.* **8**, 1–10 (2018).
96. Caetano Júnior, P. C., Lemes, L. C., Aguiar, J. C., Strixino, J. F. & Raniero, L. Application of FT-IR spectroscopy to assess physiological stress in rugby players during fatigue test. *Res. Biomed. Eng.* **32**, 123–128 (2016).
97. Caetano Júnior, P. C., Strixino, J. F. & Raniero, L. Analysis of saliva by Fourier Transform Infrared Spectroscopy for diagnosis of physiological stress in athletes. *Res. Biomed. Eng.* **31**, 293–300 (2015).
98. Celebi, M. E., Kingravi, H. A. & Vela, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst. Appl.* **40**, 200–210 (2013).
99. Murtagh, F. A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* **26**, 354–359 (1983).
100. Lever, J., Krzywinski, M. & Altman, N. Points of Significance: Principal component analysis. *Nat. Methods* **14**, 641–642 (2017).
101. Bro, R. & Smilde, A. K. Principal component analysis. *Anal. Methods* **6**, 2812–2831 (2014).
102. Travo, A. *et al.* Potential of FTIR spectroscopy for analysis of tears for diagnosis purposes. *Anal. Bioanal. Chem.* **406**, 2367–2376 (2014).
103. Lima, K. M. G., Gajjar, K. B., Martin-Hirsch, P. L. & Martin, F. L. Segregation of ovarian cancer stage exploiting spectral biomarkers derived from blood plasma or serum analysis: ATR-FTIR spectroscopy coupled with variable selection methods. *Biotechnol. Prog.* **31**, 832–839 (2015).
104. Rashid, N. A., Hussain, W. S. E. C., Ahmad, A. R. & Abdullah, F. N. Performance of classification analysis: A comparative study between PLS-DA and integrating PCA+LDA. *Math. Stat.* **7**, 24–28 (2019).
105. Makki, A. A. *et al.* Qualitative and quantitative analysis of therapeutic solutions using Raman and infrared spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **218**, 97–108 (2019).
106. Diem, M. *Modern Vibrational Spectroscopy and Micro-Spectroscopy*. *Modern Vibrational Spectroscopy and Micro-Spectroscopy* (2015). doi:10.1002/9781118824924.
107. Haenlein, M. & Kaplan, A. M. A Beginner's Guide to Partial Least Squares Analysis. *Underst. Stat.* **3**, 283–297 (2004).
108. Woutersen, S. & Hamm, P. Nonlinear two-dimensional vibrational spectroscopy of peptides. *J. Phys. Condens. Matter* **14**, 1035–1062 (2002).
109. Alperstein, A. M., Ostrander, J. S., Zhang, T. O. & Zanni, M. T. Amyloid found in human cataracts with two-dimensional infrared spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 6602–6607 (2019).
110. Strasfeld, D. B., Ling, Y. L., Gupta, R., Raleigh, D. P. & Zanni, M. T. Strategies for Extracting Structural information from 2D IR Spectroscopy of Amyloid: Application to Islet Amyloid Polypeptide. *J. Phys.* **113**,

15679–15691 (2009).
111. Krummel, A. T., Mukherjee, P. & Zanni, M. T. Inter and Intrastrand Vibrational Coupling in DNA Studied with Heterodyned 2D-IR Spectroscopy. *J. Phys. Chem. B* **107**, 9165–9169 (2003).
112. Shaw, D. J. *et al.* Disruption of key NADH-binding pocket residues of the Mycobacterium tuberculosis InhA affects DD-CoA binding ability. *Sci. Rep.* **7**, 1–7 (2017).
113. Shaw, D. J. *et al.* Multidimensional infrared spectroscopy reveals the vibrational and solvation dynamics of isoniazid. *J. Chem. Phys.* **142**, 212401 (2015).
114. Simpson, N. & Hunt, N. T. Ultrafast 2D-IR spectroscopy of haemoproteins. *Int. Rev. Phys. Chem.* **34**, 361–383 (2015).
115. Demirdöven, N. *et al.* Two-dimensional infrared spectroscopy of antiparallel beta-sheet secondary structure. *J. Am. Chem. Soc.* **126**, 7981–7990 (2004).
116. Baiz, C. R., Peng, C. S., Reppert, M. E., Jones, K. C. & Tokmakoff, A. Coherent two-dimensional infrared spectroscopy: Quantitative analysis of protein secondary structure in solution. *Analyst* **137**, 1793–1799 (2012).
117. Zhang, T. O., Alperstein, A. M. & Zanni, M. T. Amyloid β-Sheet Secondary Structure Identified in UV-Induced Cataracts of Porcine Lenses using 2D IR Spectroscopy. *J. Mol. Biol.* **429**, 1705–1721 (2017).
118. Minnes, L. *et al.* Quantifying Secondary Structure Changes in Calmodulin Using 2D-IR Spectroscopy. *Anal. Chem.* **89**, 10898–10906 (2017).
119. Bredenbeck, J., Helbing, J., Nienhaus, K., Nienhaus, G. U. & Hamm, P. Protein ligand migration mapped by nonequilibrium 2D-IR exchange spectroscopy. *Proc. Natl. Acad. Sci.* **104**, 14243–14248 (2007).
120. Dunkelberger, E. B., Grechko, M. & Zanni, M. T. Transition Dipoles from 1D and 2D Infrared Spectroscopy Help Reveal the Secondary Structures of Proteins: Application to Amyloids. *J. Phys. Chem. B* **119**, 14065–14075 (2015).
121. Grechko, M. & Zanni, M. T. Quantification of transition dipole strengths using 1D and 2D spectroscopy for the identification of molecular structures via exciton delocalization: Application to alpha-helices. *J. Chem. Phys.* **137**, 184202 (2012).
122. Scott, D. A. *et al.* Diabetes-related molecular signatures in infrared spectra of human saliva. *Diabetol. Metab. Syndr.* **2**, 1–9 (2010).
123. Lemes, L. C., Caetano Júnior, P. C., Strixino, J. F., Aguiar, J. & Raniero, L. Analysis of serum cortisol levels by Fourier Transform Infrared Spectroscopy for diagnosis of stress in athletes. *Res. Biomed. Eng.* **32**, 293–300 (2016).
124. Llewellyn, D., Langa, K., Friedland, R. & Lang, I. Serum Albumin Concentration and Cognitive Impairment. *Curr. Alzheimer Res.* **7**, 91–96 (2010).
125. Feig, M. A., Jehmlich, N. & Völker, U. Chapter 6: Personalized proteomics of human biofluids for clinical applications. in *Advanced LC-MS Applications for Proteomics* (ed. Pennington, S. R.) (Future Science Group, 2013).
126. Greetham, G. M. *et al.* ULTRA: A Unique Instrument for Time-Resolved Spectroscopy. *Appl. Spectrosc.* **64**, 1311–1319 (2010).
127. Luther, B. M., Tracy, K. M., Gerrity, M., Brown, S. & Krummel, A. T. 2D IR spectroscopy at 100 kHz utilizing a Mid-IR OPCPA laser source. *Opt. Express* **24**, 4117 (2016).
128. Shim, S. H., Strasfeld, D. B., Ling, Y. L. & Zanni, M. T. 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.* **104**, 14197–14202 (2007).
129. Bloem, R., Garrett-Roe, S., Strzalka, H., Hamm, P. & Donaldson, P. Enhancing signal detection and completely eliminating scattering using quasi-phase-cycling in 2D IR experiments. *Opt. Express* **18**, 27067 (2010).
130. Réhault, J. & Helbing, J. Angle determination and scattering suppression in polarization-enhanced two-dimensional infrared spectroscopy in the pump-probe geometry. *Opt. Express* **20**, 21665 (2012).

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Matthew J. Baker is a director of Dxcover Ltd.