Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Multivariate statistical methodologies applied in biomedical Raman spectroscopy: Assessing the validity of partial least squares regression using simulated model datasets.

Mark E. Keating[1,2]*, Haq Nawaz[3] , Franck Bonnier[1,4] and Hugh J. Byrne[1]

Abstract

Raman spectroscopy is fast becoming a valuable analytical tool in a number of biomedical scenarios, most notably disease diagnostics. Importantly, the technique has also shown increasing promise in the assessment of drug interactions on a cellular and subcellular level, particularly when coupled with multivariate statistical analysis. However, with respect to both Raman spectroscopy and the associated statistical methodologies, an important consideration is the accuracy of these techniques and more specifically the sensitivities which can be achieved, and ultimately the limits of detection of the various methods. The purpose of this study is thus the construction of a model simulated data set with the aim of testing the accuracy and sensitivity of the partial least squares regression (PLSR) approach to spectral analysis. The basis of the dataset is the experimental spectral profiles of a previously reported Raman spectroscopic analysis of the interaction of the cancer chemotherapeutic agent cisplatin in an adenocarcinomic human alveolar basal epithelial cell- line, *in vitro*, and is thus reflective of actual experimental data. The simulated spectroscopic data is constructed by adding known perturbations which are independently linear in drug dose, as well as cytological response, experimentally determined by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) cytotoxicity assay. It is demonstrated that, through appropriate choice of dose range, PLSR against the respective targets can differentiate between the spectroscopic signatures of the direct chemical effect of the drug dose and the indirect cytological effect it produces.

Keywords: Raman Spectroscopy, Drug interaction studies, Partial Least Squares Regression, Simulated dataset.

[1]FOCAS Research Institute, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland.

[2]School of Physics, Dublin Institute of Technology, Kevin Street, Dublin 8 Ireland.

[3]National Institute for Biotechnology and Genetic Engineering (NIBGE), P.O.Box 577, Jhang Road Faisalabad, Pakistan.

[4]Faculty of Pharmacy, EA 6295 – NM/NP, Université François-Rabelais de Tours, 60 rue du Plat D'Etain, 37020 Tours Cedex 1, France

*E-mail: Mark.Keating@mydit.ie

Introduction

Over the past couple of decades, vibrational spectroscopy (in particular Raman and infrared absorption) has emerged as a powerful tool for biomedical applications. The numerous studies explore applications such as disease diagnostics[1–4], cellular imaging[5–8], the study of drug[9–11]and nanoparticle interactions[12–14] on a cellular and sub-cellular level, to name but a few. In both modalities, the spectrum of tissue or cells contains a wealth of information and represents the combined molecular fingerprints of the ensemble of biomolecules contained in the sample. As a result, only in the simplest of cases can a valid interpretation be made by visual inspection of the spectrum. Multivariate statistical methods are thus critical in the analysis, interpretation and representation of the complex information contained within. However, given the critical nature of the outcomes of the application, whether in terms of medical diagnostics or in preliminary screening of drug efficacy and action mechanisms, it is imperative that the combination of spectroscopic techniques and multivariate analysis are rigorously and quantifiably validated. Such validation can also establish realistic limits to what is often purported as a high content screening methodology. To this aim, the use of simulated datasets based on experimental studies can play a crucial role[14,15].

A multitude of multivariate analytical methods exists, each of which aims to simplify complex biospectroscopic information and provide a tool with which to draw conclusions about the state of the sample. These include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Vertex Component Analysis (VCA), Spectral Cross Correlation Analysis (SCCA), K-means Clustering Analysis (KMCA), Hierarchal Cluster Analysis (HCA) to name but a few. Importantly, there also exists a number of variants of these methods which differ slightly and can give, in some instances, different answers[14,16,17].

Recently, regression modelling (e.g. Partial Least Squares Regression, PLSR) has seen a number of biomedical uses in both Raman and IR spectroscopies . The core idea of using this method is to investigate the spectral variability as a function of a systematic conditional change such as

radiation dose[18] or viral infection[19]. PLSR can be employed to construct predictive models for spectral response as a function of the target variable. Therefore, an unknown dose or degree of infection can be determined from its spectrum, having obvious potential clinical applications. Furthermore, feature selection techniques such as PLSR o-efficients, Jack-Knifing (JK) and genetic algorithms, amongst others[20], can be employed to identify the most statistically relevant spectral changes, such that the biological mechanisms underlying the spectral changes can explored and understood. Importantly, there are many variants of the PLSR algorithm and, in some instances, hybrid methods which use a combination of two statistical tools in order to extract relevant chemical information have been employed. Although these methods have been applied to a wide range of studies, the details are beyond the scope of this paper although good examples can be found in literature[10,11,18,21–24]

The potential of Raman spectroscopic microscopy for initial screening of chemotherapeutic efficacy and mechanism of action has been demonstrated by Nawaz *et al*.[10,11,23]. Taking the interaction of cisplatin with the human lung adenocarcinoma cell line, A549, *in vitro*, as an example, PLSR of Raman spectroscopic datasets was reported to identify and differentiate the direct effects of cisplatin on the cellular biochemistry as a function of drug concentration (dose) and the resultant toxicological response as measured by the MTT cytotoxicity assay. This simultaneously provides a parallel gold standard technique to compare to the spectroscopic endpoint as well as range finding for the initial dose response curve i.e. establishing values of Inhibitory Concentrations (IC) etc. In an operational model of pharmacological agonism, the former is a linear process, where as the latter results in the more complex sigmoidal response of cell populations to drug exposure[25]. PLSR against the drug concentration returned changes in the Raman peaks associated with both conformational and chemical changes in DNA, while changes to the lipid and protein distributions were dominant when the data was regressed against the cytotoxicological end point, indicating the biochemical changes associated with the resultant cytological response to the interaction with cisplatin. The statistic relevance of the results were confirmed using the JK approach.

The potential to differentiate the direct chemical effects from the subsequent cytological responses opens the way to the use of the techniques to visualise and interpret the mode of action of chemotherapeutic agents intracellularly and to quantify the efficacy to produce the desired cellular response in a single truly label free measurement. The emergence of ever higher throughput spectrometers would enable realtime and time resolved visualisation of the respective processes as they evolve. Notably, however, while the studies of Nawaz *et al*. show great promise towards this end, the technique is as yet unvalidated. The expected changes in the spectra with concentration and toxicological endpoint are inferred, based on prior knowledge about the biological action of cisplatin in the model *in vitro* system. This leads to a difficulty when trying to confirm the validity of the method or compare two different methods to quantitatively assess the sensitivity, accuracy and specificity of the technique.

Here, we aim to validate the application of these methodologies using simulated datasets based on the previously published experimental results of Nawaz *et al.*. In particular, we aim to test the ability of PLSR to model and thus extract spectroscopic variations (based on the regression co-efficients) which vary systematically as a function of different targets. Thus, the study will confirm whether the method is capable of extracting and differentiating spectroscopic features which differ based on linear or non-linear changes of the targets. Additionally, the accuracy or fidelity of the method in extracting systematically varied features will be explored as the spectral perturbations introduced decrease in magnitude, exploring the sensitivity of the method. Thus, the overarching aim is to establish the validity of the algorithms applied to Raman spectral datasets containing changes pertaining to the direct and indirect effects of the anti-cancer drug cisplatin *in vitro*. For the purposes of this study, we propose the use of a modelled simulated dataset. The dataset is constructed based on experimental observations, but the systematic spectral variation that is introduced is known precisely and thus an exact and complete assessment of the method can be carried out.

Experimental

Experimental results were obtained as described in previous publications by Nawaz *et al*[10,11] which investigated Raman spectroscopy as a tool to study cisplatin-cellular interactions *in vitro*. The experimental methods are described in detail in the publications, but are summarised in brief as follows.

Human lung adenocarcinoma (A549) cells were routinely cultured at 37 °C, 5 % CO_2 in DMEM F12 supplimented with 10% FBS, 1% pen/strep and 2mM l-glutamine. Cells were cultured until 70-80% confluency and plated on quartz substrates for Raman spectroscopy . A stardard MTT assay, using a concentration range of 0.05μM – 50 μM, was used to assess the toxicity of cisplatin to provide a comparison to Raman spectroscopy. This was carried out in standard 96 well plates and experiments were all completed in triplicate. This range resulted in a sigmoidal variation in cell culture viability over the range ~90% to ~20%, from which the Inhibitory Concentration (IC_{50}) of cisplatin in A549 cells *in vitro* was determined to be 1.2 ± 0.2 μM.

Cisplatin, at varying concentrations in the range 0.05 μM - 50μM, was added to cells and Raman microscopic measurements of cells exposed to each dose, including unexposed control, were acquired at a source wavelength of 785nm for both nuclear[10]and cytoplasmic regions[11] . In both cases, multiple spectra were recorded from a total of 60 cells at each exposure level. The PLSR approach was used to model the spectroscopic data as well as to select and distinguish the relevant features indicative of the chemical effects of cisplatin and the cellular response to cisplatin via a regression against dose and the MTT cytotoxicity endpoint respectively. By examination of the regression co-efficients, it was possible to discern the major features responsible for model construction.

In this work, these experimental spectral datasets are employed to construct semi-realistic simulated data to probe the reliability, sensitivity and quantitative nature of these methods when

applied to drug-interaction studies. More details of the experimental set up can be found in Nawaz *et al.* [10,11]

Partial Least Squares Regression

PLSR is a multivariate statistical method which aims to establish a model that relates the variations of the spectral data to a series of relevant targets. The spectral data (X matrix) is thus related to the targets (Y matrix) according to the linear equation Y = XB +E, where B is a matrix of regression coefficients and E is a matrix of residuals. The PLSR algorithms used in this study have been previously published elsewhere[10,11,18,22] and are based on scripts written in house using Matlab 7.2 (The Mathworks Inc.). The algorthim allows for the construction of a regression model which can be used to predict the outcome in a number of different situations. In this case, the examples used are concentration and MTT response, and therefore the algorthim can be used to predict for example the toxicological response of a particular drug dose.

Latent variables (LV's) in PLSR modelling are a seris of underlying variables which aim to describe the behaviour of the modelled system. The exact number of latent variables which are necessary to build an entirely accurate model is not known *a priori*. However, it is one of the goals of PLSR models to accurately predict the number necessary to build a robust and accurate model[26]. Predicting the number of LVs which will bulid an accurate model is usually achieved during the cross validation step, typically using the root mean squared error of cross validation (RMSECV) as a metric for latent variable selection.

Spectral Constructs

Spectral constructs were generated for the purpose of imparting a known perturbation to the dataset which could be systematically varied to evaluate the capability of the PLSR modeling to accurately predict and extract spectral variations correlated to a known external variable, in this case, drug dose and the resultant cytological changes. Using the original datasets of Nawaz *et al.*,

derived from the nuclear and cytoplasmic regions, specific spectral changes were identified in the mean difference spectra of a 3μM exposed cell population versus the unexposed control (Figure 3, of reference 10, Figure 4 of reference 11). In this way, spectral constructs were generated from the changes in the spectra of the nuclear region, including increases in the characteristic A form of DNA peak at 807 cm^{-1} and the B form peak at 833 cm^{-1} and a change in the C-H deformation at 1449 cm^{-1} (Figure 1A) and in the cytoplasmic region, containing the following peak changes or shifts; a change in the amide 1 band at ~1661 cm^{-1}, a decrease in the C-C stretch intensity at ~939 cm^{-1} and an increase in the tryptophan peak at 731 cm^{-1} (Figure 1B). The relative intensities of the peaks in each construct were derived from the experimental difference spectra at a cisplatin exposure dose of 3μM[10] and were normalised for concentration (Figure 1A) and a loss of viability at that concentration of 0.52 [10] (Figure 1B). Different weightings of these spectral constructs (termed hereafter the Concentration and Viability construct respectively) were then added to a control dataset as described in the following section.

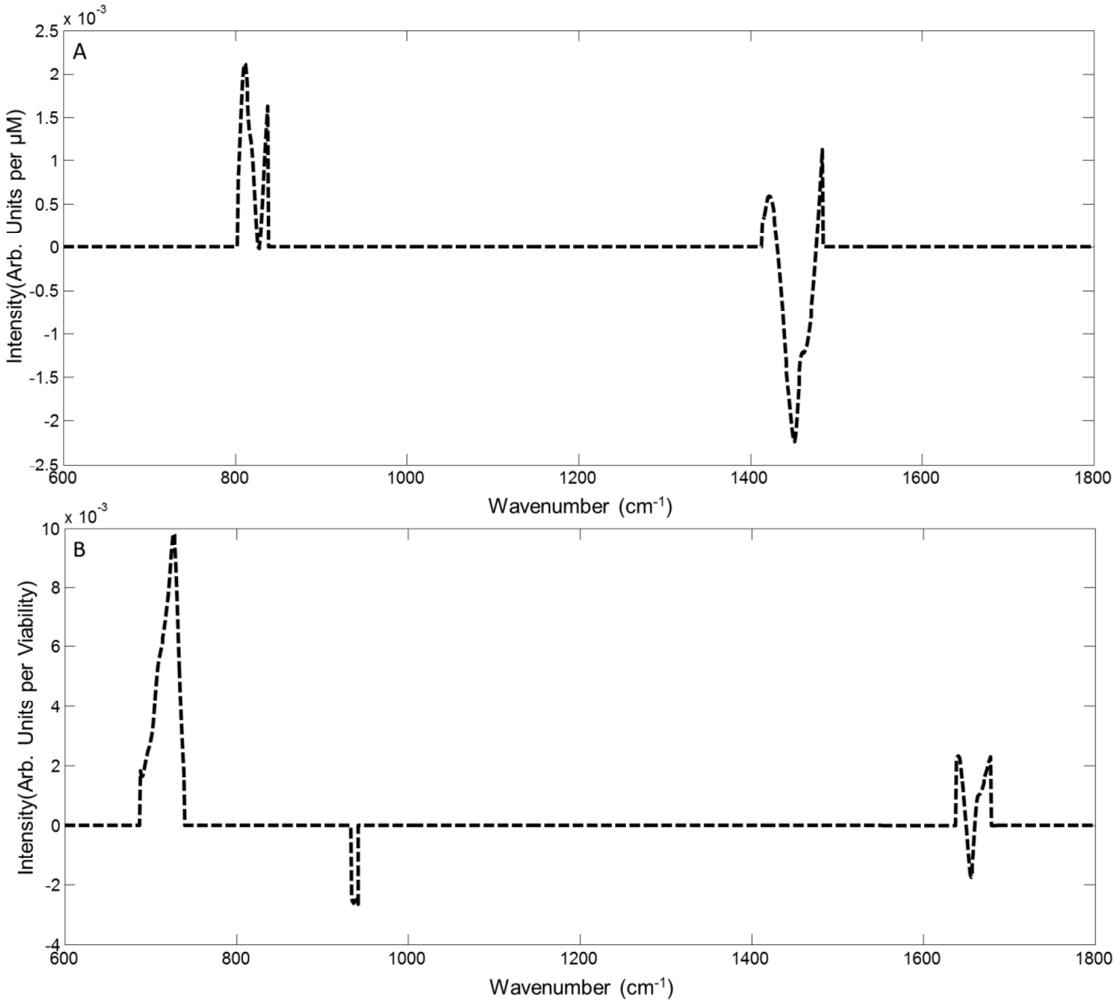

Figure 1: Spectral Constructs based on the normalised difference spectra between control and exposed nucleus (A) [10,] *and cytoplasm* [11] *(B). Selected Raman peaks were used to avoid over complexity in the simulated data; (A) the A form peak of DNA at 807 cm^{-1} and the B form peak at 833 cm^{-1} and the C-H deformation at 1449 cm^{-1} (B) the amide 1 band at ~1661 cm^{-1}, the C-C stretch intensity at ~939 cm^{-1} and the tryptophan peak at 731 cm^{-1}.*

Simulated data

Simulated datasets were generated in the following manner. A control dataset containing 25 spectra acquired from the nucleus of non-cisplatin exposed (control) cells was selected from Nawaz *et al* [10] (Figure 2). Notably, this real experimental dataset contains instrumental noise and sample variability. To this dataset, weighted contributions of the Concentration construct shown in Figure 1A, based on the experimentally observed difference spectra of the nuclear region, were added,

over the Lethal Concentration range 0.05 µM - 50µM used in the original study, which includes the IC_{50}, based on a direct weighting of the spectral construct by the range of concentrations (Table 1). Initially, only the concentration dependent weighted constructs were added to the control, to produce Dataset 1.

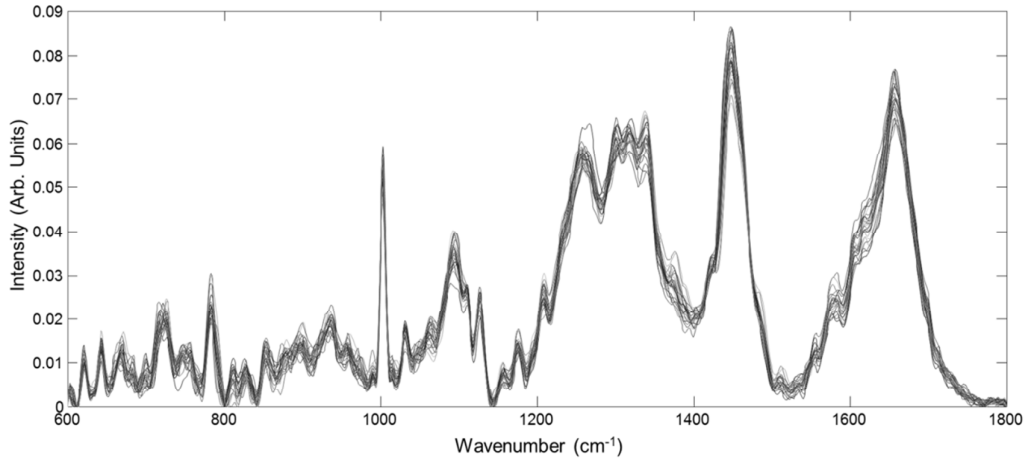

Figure 2: Control dataset taken from Nawaz et al.[10]; 25 control spectra taken from the nucleus of cells not exposed to cisplatin. Spectra have been baseline corrected and vector normalised. The inherentspectral variability in the data is representative of real experimental conditions. These spectra were then used in the construction of 3 simulated datasets, each containing 8 different dose/viability points with systematically introduced variation of the spectral constructs shown in figure 1.

As the MTT assay is expressed in viability compared to control (0.845 being maximum (V_{max}) and 0.135 being minimum values of fit to the experimentally observed viability over the concentration range[10]), the spectral construct of Figure 1B, derived from the experimentally observed differences in the cytoplasmic region, was similarly weighted by the (V_{max} – MTT) endpoints in Table 1 and also added to Dataset 1. Each spectral construct was therefore added following a linear trend based on concentration (Figure 1A) plus a linear trend based on MTT response (Figure 1B). The MTT endpoint data are, however, nonlinearly related to the concentration, in a sigmoidal fashion typical of cytotoxic responses, as shown in Nawaz *et al*[10,11]. The resultant dataset therefore contains 25 spectra for each of 8 dose points (including control) which incorporate

spectral variations, systematically dependent on both the exposure dose and the measured cytological response. For simplicity, this is referred to as Dataset 2.

It is noted that the spectral construct of Figure 1B is derived from exposure dose dependent, experimentally observed, spectral changes in the cytoplasmic region. No direct biological significance is inferred by the weighted addition of this spectral construct to the dataset derived from the nuclear regions. However, the addition serves to provide an independently variable perturbation to the dataset, which may serve to mimic a cytological effect of the direct action of the drug in the nucleus.

To probe the sensitivity of the methodology, the experimental range for cisplatin (Lethal Concentration, in table 1) has been extended (Sub lethal Concentration in table 1) to represent non-lethal doses of the drug. The MTT values have also been extrapolated according to the original fit of the Hill equation[10] to reflect these changes in concentration (Sub-lethal MTT in table 1). The corresponding simulated dataset will be referred to as Dataset 3. A dataset was also constructed which consisted solely of control spectra. This Control dataset did not contain any systematically introduced spectral variations and was used to establish a baseline regression endpoint for both Lethal Concentration and Lethal MTT.

Lethal Concentration	***Sub-lethal Concentration***	***Lethal MTT***	***Sub-lethal MTT***
0.05	*0.0005*	*0*	0.000001
0.5	*0.005*	*0.15*	0.000001
1	*0.01*	*0.35*	0.000001
3	*0.03*	*0.52*	0.00001
5	*0.05*	*0.55*	0.0001
10	*0.1*	*0.65*	0.001
50	*0.5*	*0.66*	0.01

Table 1: The weightings of the spectral constructs added to the control data. The Lethal Concentration and Lethal MTT ranges are derived from the actual experiment data of references [10,11]*. Lethal MTT represents the values obtained when the experimental MTT value is subtracted from* V_{max}*. The Subletha Concentrations extend the concentration range and are representative of sub-lethal doses of cisplatin, for which sub-lethal MTT values are derived from the extrapolated fit of the Hill equation in Reference 10.*

Results

Concentration Simulated data

The PLSR method aims to establish a model that relates the variations of the spectral data to a series of relevant targets. In this case, the spectral data is a series of simulated datasets which are based on known introduced pertubations based on cisplatin-cellular interactions as described in the previous sections.

Regression of Dataset 1 against the Lethal Concentration range (table 1) yielded the model shown in figure 3. The data were split, 60:40, to create calibration and test sets to build the model. 60% of the data was used to calibrate the model and 40% of the data was then used to assess the performance of the model in predicting the expected target with unseen data. Leave-one out cross validation with the calibration set was used to determine the optimal model complexity for use in testing (Meade et al., 2010)[27]. This process was performed with randomization of the data matrix and splitting of the data to prevent data bias (Varmuza and Filzmoser, 2009)[28]. Control of over fitting was achieved using a procedure previously described by Martens and Naes[29]. The procedure involves selection of the optimal number of latent variables (LV) to retain within the PLSR model via cross-validation with the calibration data set. The optimal number of LV's was then selected on the basis of the number which provided the lowest root mean squared error after cross validation. This is illustrated in Supplementary Material figure S1A and B, which show plots of the RMSECV and RMSEP for the first 10 LV's for the regression of Dataset 1 against Lethal Concentration 1. The values for RMSECV and RMSEP approach zero in an asymptotic fashion, and as there is no significant further

decrease after 10 LVs, 10 was chosen as the optimum number. and thus the optimum number of LV's was selected as 10. The calibration and test set had RMSEC=0.49673, RMSEP=0.52389 and R^2 values of 0.99948 and 0.99953 respectively, indicating a good linear fit of the model.

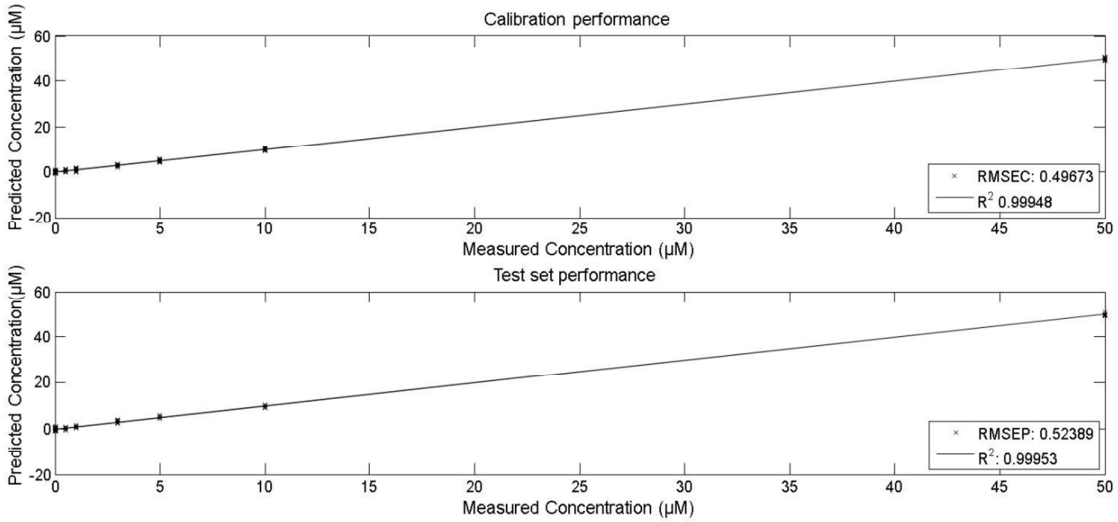

Figure 3. PLSR modelling against Lethal Concentration for Dataset 1. Top panel shows the calibration performance and test dataset (RMSEC 0.49673, R^2 0.99948). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.52389, R^2 0.99953). Data was split in a ratio of 60:40 calibration and test respectively.

As the regression co-efficients (RC) are descriptors of the spectral features which are used to build the model, we also aimed to assess the accuracy with which the alogorithm can faithfuly extract the known spectral perturbations introduced in the dataset. For regression of Dataset 1 against Lethal Concentration, we expect that the spectrum of the RC will be comprised of the Concentration construct which has been added based on the Lethal Concentration range (Figure 1A).

In figure 4, a direct comparison between the RC of regression of Dataset 1 against the Lethal Concentration range and the concentration spectral construct is shown. The spectrum of the RC is dominated by the peaks of the systematically added spectral construct, at 807cm^{-1}, 833cm^{-1} , which

correspond to A and B form DNA[10] and the C-H deformation at 1449cm^{-1} (solid line figure 4 bottom panel). This verifies that the simulated changes are the major contributors to the PLSR model construction.

However, it should be noted that the RC spectrum in figure 4 also contains other peaks which are not present in the spectral construct and so should not show a systematic variation with concentration. By regression of just the control data (with no spectral perturbations) against the Y target (Lethal Concentration) it was possible to establish a Control RC, as shown by the dotted line (bottom panel) in figure 4 (offset and multiplied by a factor of 10 for clarity). The control RC spectrum shows a high degree of similarity with the original cellular spectra (Figure 2) and thus derives from the inherent variability in the experimental measurement. Close examination of the RC for the Dataset 1 regression reveals that some of the peaks in the Control RC are also present.

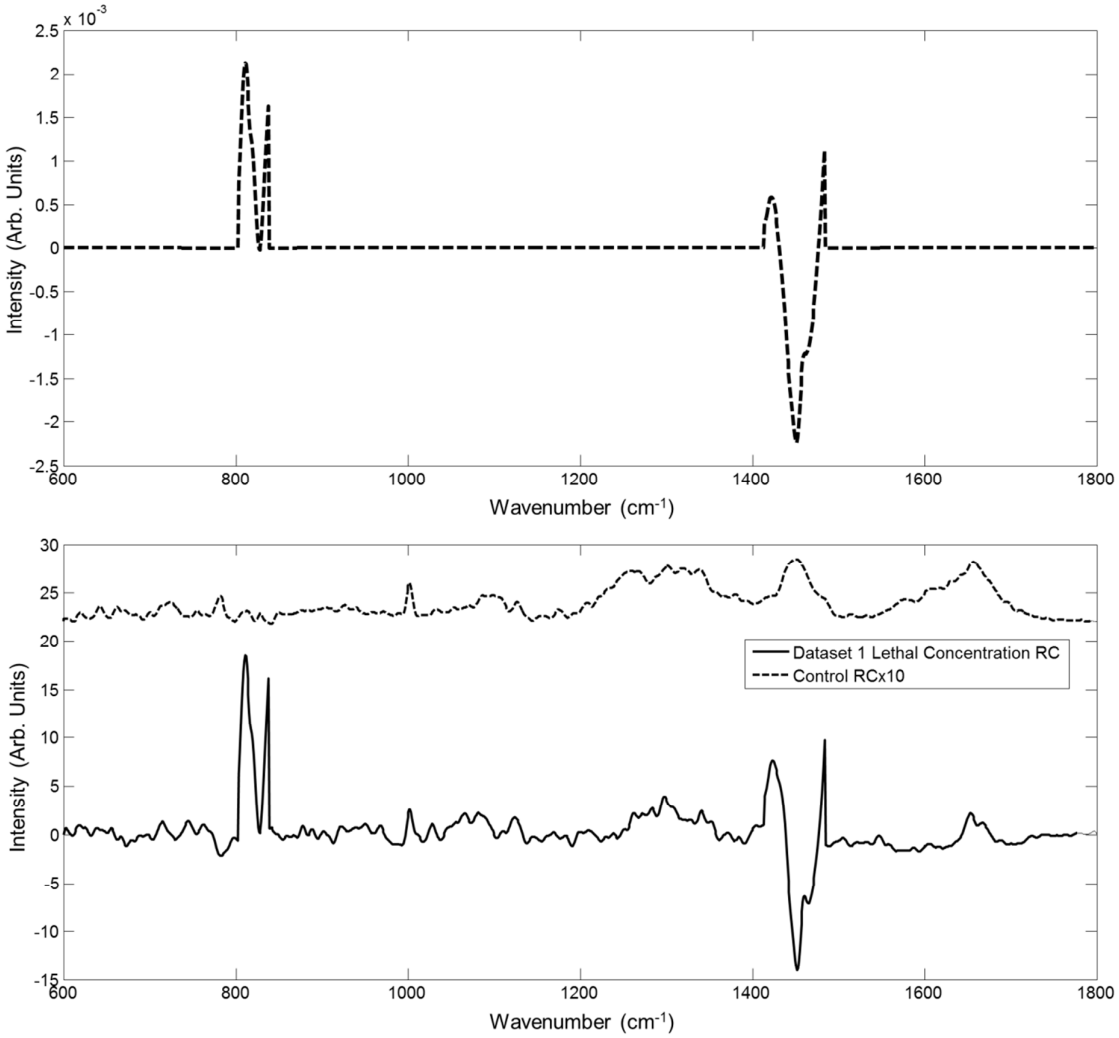

Figure 4: Plot of the regression co-efficients following PLSR of Dataset 1 against Lethal Concentration. The Concentration construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The solid line (bottom panel) shows the regression co-efficient following regression of Dataset 1 against Lethal Concentration. The dotted line shows a plot of the regression co-efficient following regression of a dataset consisting of just control spectra against Lethal Concentration, in effect showing the baseline regression co-efficient when no introduced spectral perturbation (not including sample/instrumental variations) is present. The Control RC has been offset and multiplied by a factor of 10 for clarity.

The PLSR modelling process was repeated for Dataset 2, which included the combined perturbations of the Concentration construct of Figure 1A, linearly weighted according to Lethal Concentration of Table 1, and the MTT Construct of Figure 1B, linearly weighted according to Lethal MTT of Table 1. A similar performance of model calibration and test were achieved, with RMSEC=0.4981, RMSEP=0.53505 and R^2 values of 0.99947 and 0.99952 respectively, again indicating a good linear fit of the model (Figure S2). The spectrum of RC again faithfully reproduced the Concentration Construct of Figure 1A, on a background which matches well the Control RC spectrum (Figure S3).

MTT Simulated Data

Dataset 2 also contains systematic perturbations which have been weighted according to the viability as meastured using the MTT assay, and it is of critical interest whether these spectral variations can be independently extracted using PLSR, as suggested by Nawaz et al.[10]. Regression of Dataset 2 against Lethal MTT (table 1) yielded the model shown in figure 5. As for the concentration dependent model, the data are split according to 60% calibration and 40% test data. The calibration and test set had RMSEC=0.10158, RMSEP=0.12087 and R^2 values of 0.91928 and 0.89793 respectively. Based on these values, it can be seen that, while the model has fitted the data, it does not provide as good prediction as shown for concentration (figure 3). This is also reflected by the lower R^2 values, considering that the accuracy of the linear fit is measured by how close the value is to 1. A possible explaination for this is the lower magnitude and range of weightings of spectral construct added corresponding to the MTT response (Table 1, Lethal MTT).

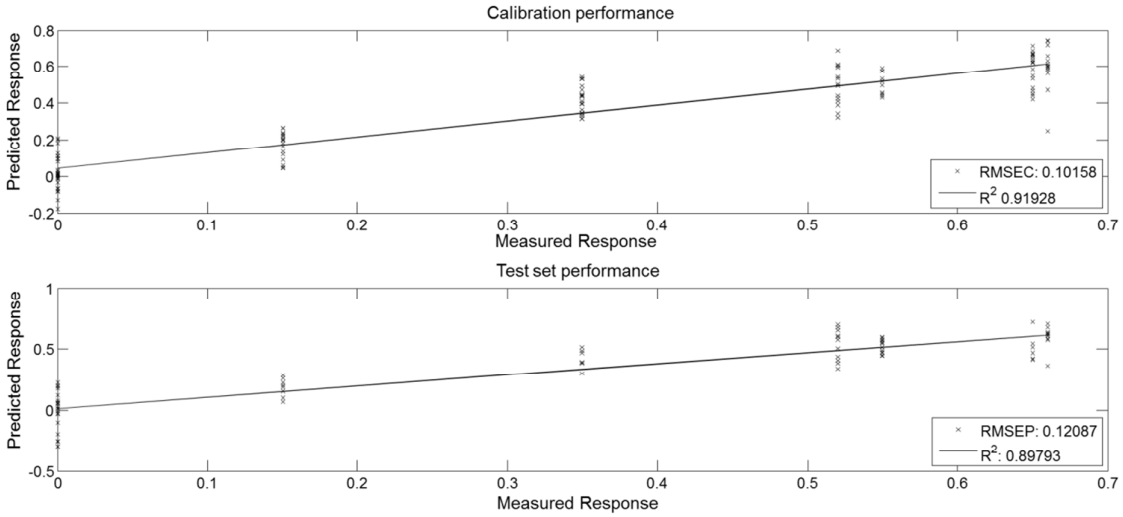

Figure 5: PLSR modelling of Dataset 2 against the Lethal MTT target. Top panel shows the calibration performance and test dataset (RMSEC 0.10158, R^2 0.91928). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.12087, R^2 0.89793). Data has been split in a ratio of 60:40 calibration and test respectively.

Inspection of the MTT RC in Figure 6 shows that the peaks of the systematically added Viability construct (Figure 6, dashed line, top panel), the amide 1 band at ~1661 cm^{-1}, the C-C stretch intensity at ~939 cm^{-1} and the tryptophan peak at 731 cm^{-1}, are faithfully reproduced and dominate the MTT RC (Figure 6, solid line, bottom panel).

The baseline sensitivity is evaluated by regressing the control dataset against the Lethal MTT target, yielding the Contol RC of Figure 6 (bottom panel, dotted line). The resultant RC spectrum has been offset and multiplied by a factor of 10, for clarity. As in the case for regression against Lethal Concentration targets, the Control RC resembles the cellular spectra of figure 2, indicating that the baseline variation is limited by the variations in the original spectral measurement.

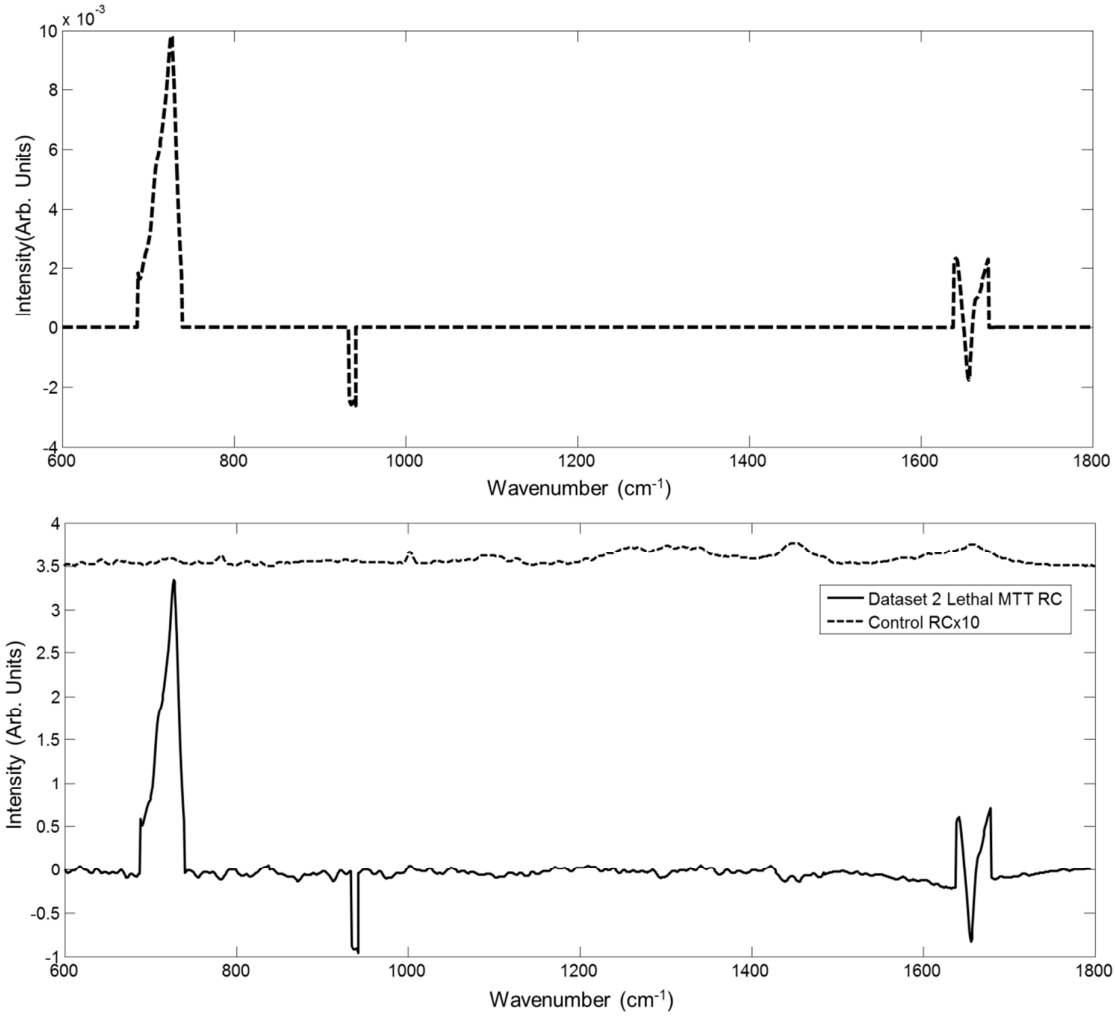

Figure 6: Plot of the regression co-efficients following PLSR modelling against MTT response. The Viability construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The solid line shows the regression co-efficient following regression against Lethal MTT and Dataset 2 (bottom panel). The dotted line (bottom panel) shows a plot of the regression co-efficient following regression of a dataset consisting of just control spectra against Lethal MTT, in effect showing the baseline regression co-efficient when no introduced spectral perturbation (not including sample/instrumental variations) is present. The Control RC is offset and multiplied by a factor of 10 for clarity.

Quantative evaluation of regression co-efficients

In an attempt to evaluate the quantitative nature of the regression co-efficients, a method was devised which looked at varying the number of datapoints used to build the PLSR model. For the analysis of the spectral variations of Dataset 1, based on variations of the Concentration construct of figure 1A weighted according to Lethal Concentration (Table 1), multiple regressions were conducted (models not shown). Each model was constructed by increasing the number of data points, C+1 being the first data set used, consisting of the control dataset (Fig 2) and the 0.05 μM datapoint of the Lethal Concentration range (Table 1). The data set was then successively extended by 1 datapoint, such that C+2 consists of control, 0.05 μM and 0.5 μM, and so on, until all data points in the Lethal Concentration were included.

For all models, the spectrum of the RC displayed a combination of the Concentration construct of Figure 1A and the Control RC of Figure 4, and, as expected, regression over the full range reproduced the RC spectrum of Figure 4. Notably, as shown in Figure 7, the peaks of the Concentration construct increase linearly as the range of the regression is increased and reach a saturation value above ~ C+4. Extension of the mocel to 1000μM results in no further significant increase of these maximum peak intensities (data not shown). The A-form DNA peak at 807cm^{-1} reaches a maximum value of 18.46. Although this does not quantitatively equate to the corresponding peak value of the Control construct of Figure 1A, the relative magnitudes of the respective peaks is consistent with those of the original Concentration construct, and notably the relative contribution of the Control RC is reduced with increasing range.

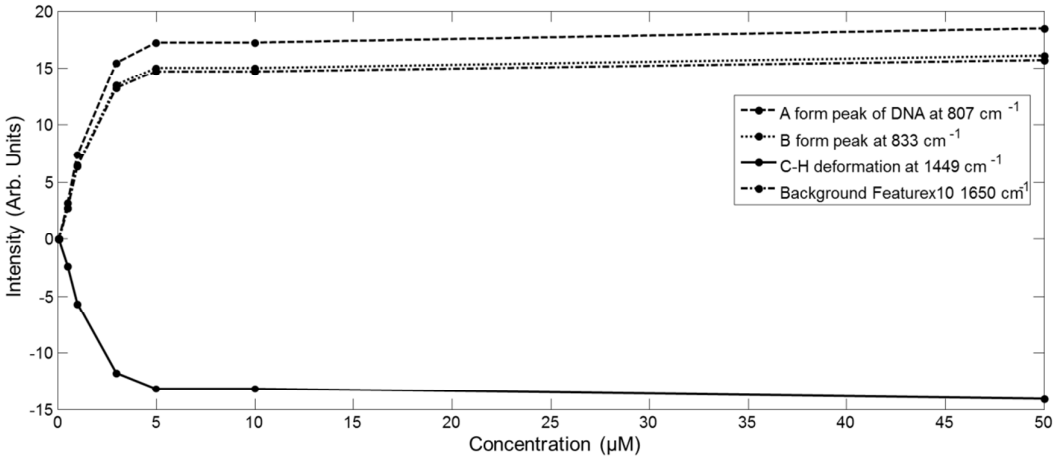

Figure 7: Evolution of the peaks of Construction construct in PLSR models of increasing range for Dataset 1.

A similar analysis was conducted for the PLSR of Dadtset 2 against the Lethal Concentration range. Figure 8 shows a plot of the extracted RCs for all successive regressions. As expected, C+7 reproduces the Lethal Concentration RC of Figure 4, and extracts the expected introduced spectral construct (Figure 1 A). However, notably for all other regressions, C+1 to C+6, the presence of peaks which are not explicitly dependant on Lethal Concentration are observed. In addition to those of the Control RC, peaks of the MTT construct (Figure 1B) are evident in the RCs of the regressions over the incomplete concentratuion range. A similar phenomenon can be seen in the equivalent sequential modelling of the MTT data of Dataset 2 (Figure S4 and S5).

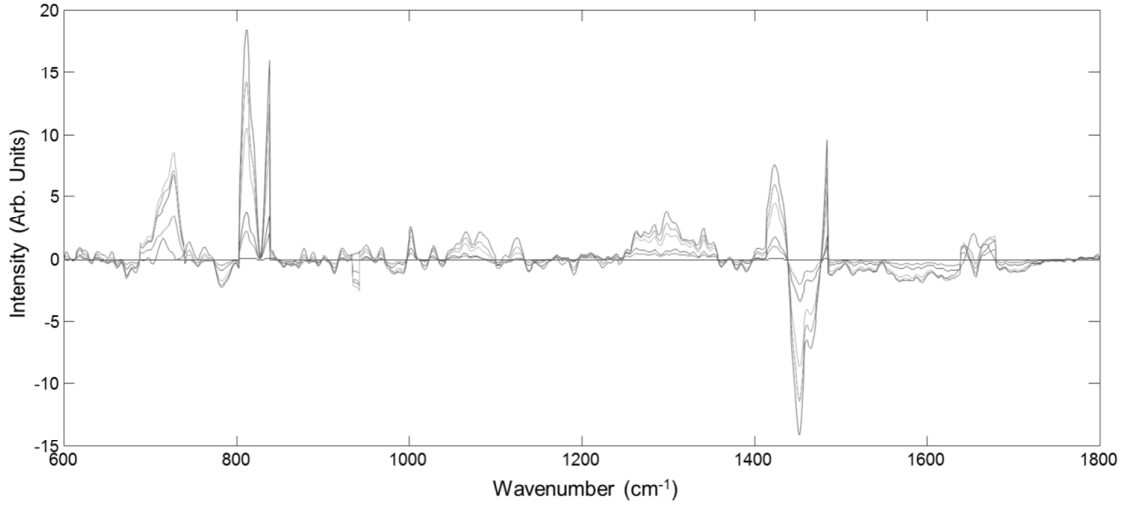

Figure 8. A plot of regression co-efficients following multiple regression against concentration with increasing data points. I.e. C+1 represents a dataset consisting of the control dataset and the data point at 0.05 μM. This then increases C+n until all data points in the dataset have been evaluated.

Figure 9 shows a plot of selected RC peak intensities associated with the spectral construct relating to concentration following successive rounds of regression as described above, namely the A form peak of DNA at 807 cm^{-1} and the B form peak at 833 cm^{-1}, which are associated with the physical changes assiociated with cisplatin-cellular interaction[11]. In fact the evolution of the peaks is observed to be identical to that observed for Dataset 1, shown in Figure 7, and although the plot of Figure 9 is in a linear/logarithmic format, it can be seen that the predicted relative intensities again increase linearly initially, before reaching a point of saturation at, or above, the dataset C+4, and further addition of datapoints makes no difference (data not shown) to the quantative prediction of the features.

Also shown in Figure 9 is the dependence of the peak of the Viability construct at 731cm^{-1}, (for example) which "bleeds through" in the regression of Dataset 2 against the incomplete concentration range. This bleed through occurs for all peaks of the MTT Construct. The contribution

of the peaks of the Viability Construct follows a trend of the derivative of the viability curve, indicating that it is the rate of change of the contributed spectral variations which governs the contribution to the RC. Notably, when the full Lethal Concentration range is included in the model, at the extremes of which the change in viability has reduced to the minimum value, the bleed through of the MTT construct is mimimal, and the Concentration Contruct of Figure 1A is faithfully extracted, albeit with an underlying background as a result of the inherent spectral variability.

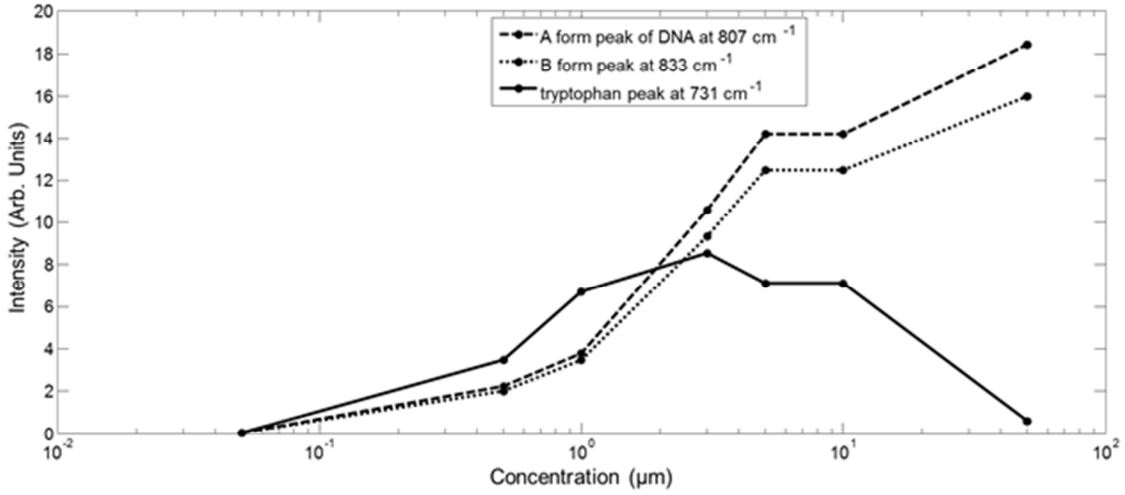

Figure 9. Plot of peak intensities vs. concentration of regression co-efficients for the A form peak of DNA at 807 cm^{-1} and the B form peak at 833 cm^{-1} of the Concentration Construct (Figure 1A). Also plotted is thecontribution of the tryptophan peak at 731cm^{-1}, a key feature of the Viability Construct (Figure 1B)

A similar PLSRA of the cntributions of the Viability construct to Dataset 2 reveals similar bleed through and more complex evolution of the features contributing to the spectrum of the RC (Supplementary Material Fugures S4 and S5). The bleed through of the features of the spectral constructs shown in Figures 8 and 9 is a clear demonstration that it is not trivial to independentlty extract the contributions of the two constructs over the lethal concentration range, as speculated by Nawaz *et al.*[10]. However, over concentration ranges in which the viability does not change

significantly, the bleed through is minimal, and the concentration dependent spectral changes can be independently extracted. Thus, it should be possible to determine the direct chemical interactions of an external agent in the sublethal range.

Figure S6 shows the calibration and test performance of the PLSR of Dataset 3 versus the Sublethal concentration range of Table 1. The model yields RMSEC and RMSEP values of 0.143 and 0.19392, respectively, with R^2 values of 0.38916 and -0.24063, accuracies considerably less that those of the equivalent model in the Lethal Concentration range. Notably, the RC spectrum is a faithful extraction of the pure Concentration construct of Figure 1 A, as shown in Figure S7. Little or no bleed through of features associacated with the Viability construct is apparent (although still present in minimal quantities) although this is not surprising as, with little or no change in viability, the contributions of the Viability construct to Dataset 3 are minimal.

Discussion

Given the drive for a reduction in the use of animal models for evaluating toxicity, screening of drugs and even cosmetics, due to regulatory developments in both the EU and US (EU Directive-2010/63/EU and US Public Law 106-545, 2010, 106th Congress)[30–32] generally based on the 3 R's of Russell and Burch[30] to replace, reduce and refine the use of animals used for scientific purposes, there is increased emphasis on the development of reliable and rapid *in vitro* sceening methodologies. This includes more representative culture models which better mimic the *in vivo* environment as well as more rapid, cost efficient, high content, and ideally label free screening technologies. It is crucial, however, that these models and technologies are well validated against established gold standards ref[33,34].

Raman spectra, in principle, contain high content information about the biochemical make up of the sample, and changes to it, related to pathology or an external agent. Raman spectra contain numerous peaks which vary dependently and independently of each other. Crucially, for real applications and particularly in the instance of drug interactions, it is difficult to tell whether these differences are inherently based on cell to cell variability or whether they are dependent on the primary action of the drug (i.e. the direct chemical effects) or the secondary effects the drug has on the cell (i.e. the response of the cell to said drug).

In this study, simulated datasets were used to evaluate the capability of PLSR to extract known and systematic spectral variation from a control dataset, which contained intrinsic experimental variability. The spectral variations introduced varied linearly with the applied drug dose and also with the measured cell population response, as measured by a standard cytotoxicity assay. Notably, however, the two spectral variations are not completely independent, as the viability response is sigmoidally dependent on the applied dose.

In the case where only a concentration dependent systematic variation in the spectra is introduced, the PLSR model provides an accurate predictive response tool, the regression co-efficients of which are based on the systematic variation which has been introduced to the dataset, linearly dependent on the targets. The model shows high sensitivity, and the limits of detection are determined only by the intrinsic variability of the experimental method, as determined by the PLSR of the Control spectral dataset. This limit can be improved by optimising sample preparation and measurement protocols. In principle, such a PLSR model can predict the response of a drug dose in a cell population, or determine an unknown drug dose from a measured spectral response.

However, the spectral changes which result from the interaction and action of a drug within a cell are manifold, and it is of interest to differentiate the spectral signatures of the direct interaction from the subsequent cellular response. Notably, this study demonstrates that, although PLSR predictive models based on regression of the combined dataset, including all spectral

responses, against the target of concentration range produce a similarly accurate, linear predictive model, the contributing RCs are only derived exclusively from the introduced concentration dependent variations in ranges where all other spectral variations are limited. For example, as shown in Figures 8 and 9, regression over the limited range of C+4 produces a model which is based on RCs which includes contributions derived from the direct effect of the interaction of the drug within the cell (Concentration construct), as well as the resultant cytological response (Viability construct). Thus, care should be taken in interpreting the spectral features which contribute to such regressions to elucidate the underlying mechanisms.

Nevertheless, in sublethal regions, the direct effects of the drug interaction can confidently be investigated employing such a PLSR analysis of Raman spectral data, independent of the cytological responses, and these are easily discernible above the intrinsic variability of the control. Although this seems a trivial conclusion, such rapid, label free analysis could prove invaluable in screening of, for example, the mechanisms and efficacy of drug interactions, evaluating drug uptake and receptor binding[25] or nanoparticle uptake and trafficking in regions where cytoxicity assays are insensitive.

The use of a parallel cytotoxic assay such as MTT serves as a range finding test to establish the IC_{50}, but also provides vital information about the sublethal doses and maximum responses. It also provides a target for regression of the data in the regions of toxicity. Thus, the subsequent cytological effects can be differentiated from the direct chemical effects of the agent and extracted from the overall spectral response in the dose range where the viability is impacted, and the cellular response can be independently mapped spectroscopically, as a function of dose and time. Notably, the model described here, which includes a single spectral construct to represent the cellular response is very simplistic, as the response is a cascade of many responses, depending on the mechanism of interaction[35]. Alternative cytological gold standard assays for cancer, such as the sulphorhodamine B assay, and human tumour cell lines such as NCI60 human tumour cell line,

should also be considered to broaden the model[36,37]. Nevertheless, the analysis presented here demonstrates that the spectral fingerprints of the direct mechanisms of interaction and the subsequent cellular responses can be independently extracted from the dose dependent spectral data, and thus, ultimately with improved screening sensitivities and speeds, Raman spectroscopy could be employed to monitor in quasi realtime, in a lable free manner, the efficacy and mode of action of, for example chemotherapeutic agents and other exogenrous agents, layng the basis for improved quantitative structure activity relationships to guide drug development or chemical regulation strategies.

Conclusions

This study demonstrates the reliability and also limitations of PLSR as a method for predictive modelling and analysis of spectroscopic signatures of cellular responses to exogeneous agents such as radiation, chemotherapeutic agents or toxins. The spectroscopic profiles at any dose/time point can derive from a complex mixture of direct interactions within the cell and a cascade of subsequent cellular response. The analysis demonstrates that care should be taken in chosing the response range and also highlights the importance of parallel cytological assays in guiding the modelling and analysis. Correct choice of range can help differentiate between the signatures of direct interactions, which are dominant at sub-lethal doses and those of the subsequent cellular response which evolve with increasing dose.

The study also demonstrates the importance of simulated datasets in exploring the potential as well as the limits of the analytical techniques. Notably, the use of real experimental data which contains sample variability and instrumental response factors as a basis of the simulated dataset helps to visualise the lower limits of sensitivity.

The results indicate that Raman spectroscopic screening combined with such regression models and feature selection techyniques, in parallel with conventional cytotoxicity assays, can be

used to screen for the efficacy of drug interactions and can contribute to understanding the mechanisms of interaction.

Acknowledgement: This research was supported by the Integrated NanoScience Platform, Ireland (INSPIRE), funded under the Higher Education Authority PRTLI (Programme for Research in Third Level Institutions) Cycle 5, co-funded by the Irish Government and the European Union Structural fund, and Science Foundation Ireland (08/PI/11).

References

1. F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan, and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

2. I. Taleb, G. Thiéfin, C. Gobinet, V. Untereiner, B. Bernard-Chabert, A. Heurgué, C. Truntzer, P. Hillon, M. Manfait, P. Ducoroy, and G. D. Sockalingum, *Analyst*, 2013, **138**, 4006–14.

3. P. Crow, B. Barrass, C. Kendall, M. Hart-Prieto, M. Wright, R. Persad, and N. Stone, *Br. J. Cancer*, 2005, **92**, 2166–70.

4. T. J. Harvey, E. Gazi, A. Henderson, R. D. Snook, N. W. Clarke, M. Brown, and P. Gardner, *Analyst*, 2009, **134**, 1083–91.

5. F. Bonnier, P. Knief, B. Lim, a D. Meade, J. Dorney, K. Bhattacharya, F. M. Lyng, and H. J. Byrne, *Analyst*, 2010, **135**, 3169–77.

6. K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark, and J. Schlegel, *Biophys. J.*, 2012, **102**, 360–8.

7. M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus, and M. Diem, *Analyst*, 2010, **135**, 2002–13.

8. C. Matthäus, T. Chernenko, J. a Newmark, C. M. Warner, and M. Diem, *Biophys. J.*, 2007, **93**, 668–73.

9. P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke, and P. Gardner, *Analyst*, 2012, **137**, 1370–7.

10. H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade, and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

11. H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng, and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

12. T. Chernenko, R. R. Sawant, M. Miljkovic, L. Quintero, M. Diem, and V. Torchilin, *Mol. Pharm.*, 2012, **9**, 930–6.

13. J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers, and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

14. M. E. Keating, F. Bonnier, and H. J. Byrne, *Analyst*, 2012, **137**, 5792–802.

15. P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, and P. Gardner, *Analyst*, 2010, **135**, 268–77.

16. H. Byrne, K. Ostrowska, and H. Nawaz, *Opt. Spectrosc. Comput. Methods Biol. Med.*, 2014, **14**, 355–399.

17. M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus, and M. Diem, *Analyst*, 2010, **135**, 2002–13.

18. A. D. Meade, H. J. Byrne, and F. M. Lyng, *Mutat. Res.*, 2010, **704**, 108–14.

19. K. M. Ostrowska, A. Malkin, A. Meade, J. O'Leary, C. Martin, C. Spillane, H. J. Byrne, and F. M. Lyng, *Analyst*, 2010, **135**, 3087–93.

20. R. M. Balabin and S. V Smirnov, *Anal. Chim. Acta*, 2011, **692**, 63–72.

21. M. Jimenez-Hernandez, C. Hughes, P. Bassan, F. Ball, M. D. Brown, N. W. Clarke, and P. Gardner, *Analyst*, 2013, **138**, 3957–66.

22. K. W. C. Poon, F. M. Lyng, P. Knief, O. Howe, A. D. Meade, J. F. Curtin, H. J. Byrne, and J. Vaughan, *Analyst*, 2012, **137**, 1807–14.

23. H. Nawaz, A. Garcia, A. D. Meade, F. M. Lyng, and H. J. Byrne, *Analyst*, 2013, **138**, 6177–84.

24. D. Rohleder, W. Kiefer, and W. Petrich, *Analyst*, 2004, **129**, 906–11.

25. J. Black and P. Leff, *Proc R Soc L. B Biol Sci.*, 1983, **220**, 141–162.

26. S. Wold, M. Sjöström, and L. Eriksson, *Chemom. Intell. Lab. …*, 2001, 109–130.

27. A. Meade, C. Clarke, H. Byrne, and F. Lyng, *Radiat. Res.*, 2010, **2**, 225–37.

28. K. Vermuza and P. Flizmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, 2009.

29. H. Martens and T. Næs, *Multivariate Calibration*, John Wiley & Sons, 1994.

30. THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, *Off. J. Eur. Union*, 2010, 33–79.

31. U. S. Congress, 2001, 2721–2725.

32. W. Russell, R. Burch, and C. Hume, *The principles of humane experimental technique*, Methuen, London, 1959.

33. A. Tfayli, F. Bonnier, Z. Farhane, D. Libong, H. J. Byrne, and A. Baillet-Guffroy, *Exp. Dermatol.*, 2014, **23**, 441–3.

34. F. Bonnier, M. Keating, T. Wróbel, K. Majzner, M. Baranska, A. Garcia, A. Blanco, and H. J. Byrne, *Toxicol. Vitr.*, 2014, **29**, 124–131.

35. M. Maher, P. C. Naha, S. P. Mukherjee, and H. J. Byrne, *Toxicol. Vitr.*, 2014, **28**, 1449–60.

36. V. Vichai and K. Kirtikara, *Nat. Protoc.*, 2006, **1**, 1112–1116.

37. R. Shoemaker, *Nat. Rev. Cancer*, 2006, **6**, 813–823.