Electronic Supplementary Material (ESI) for 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]

[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

Supplemental Material:

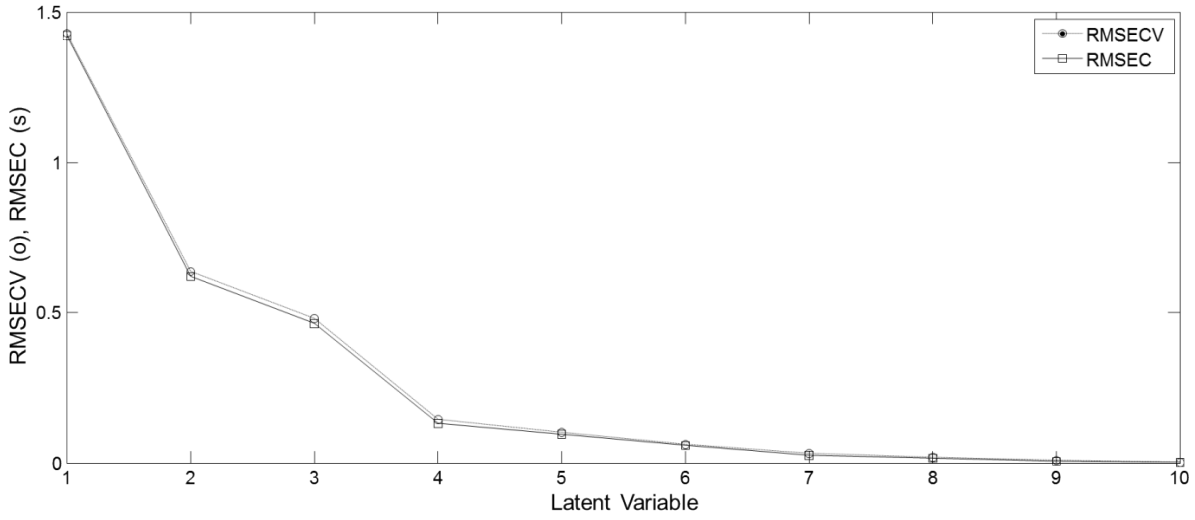

Figure S1: RMSECV and RMSEP for the first 10 LV's for the regression of Dataset 1 against Lethal Concentration 1

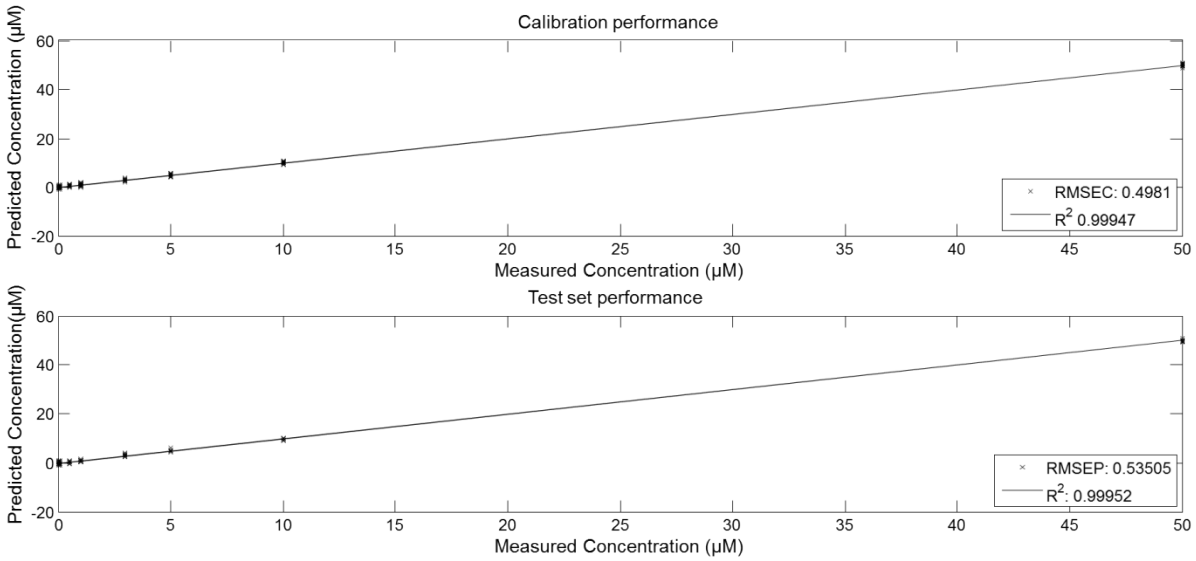

Figure S2: PLSR modelling of Dataset 2 with the Lethal Concentration range as target. Top panel shows the calibration performance and test dataset (RMSEC 0.4981, R^2 0.99947). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.53505, R^2 0.99952). Data was split in a ratio of 60:40 calibration and test respectively.

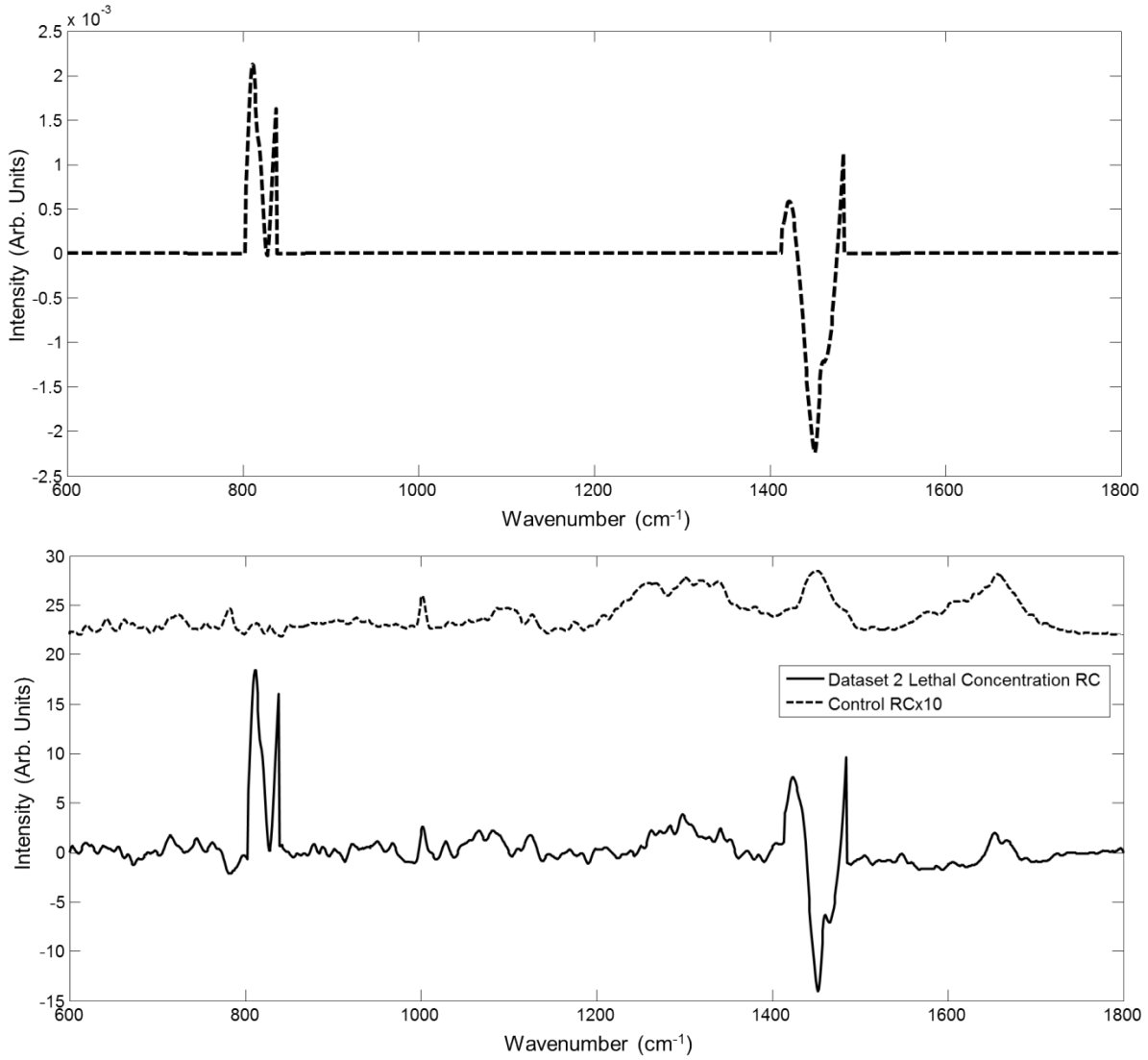

Figure S3: Plot of the regression co-efficient following PLSR modelling of Dataset 2 against Lethal Concentration. The concentration spectral construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The dashed line (bottom panel) shows the spectrum of regression co-efficients following regression of Dataset 2 against Lethal Concentration 1. The solid 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 multiplied by a factor of 10 and offset for clarity.

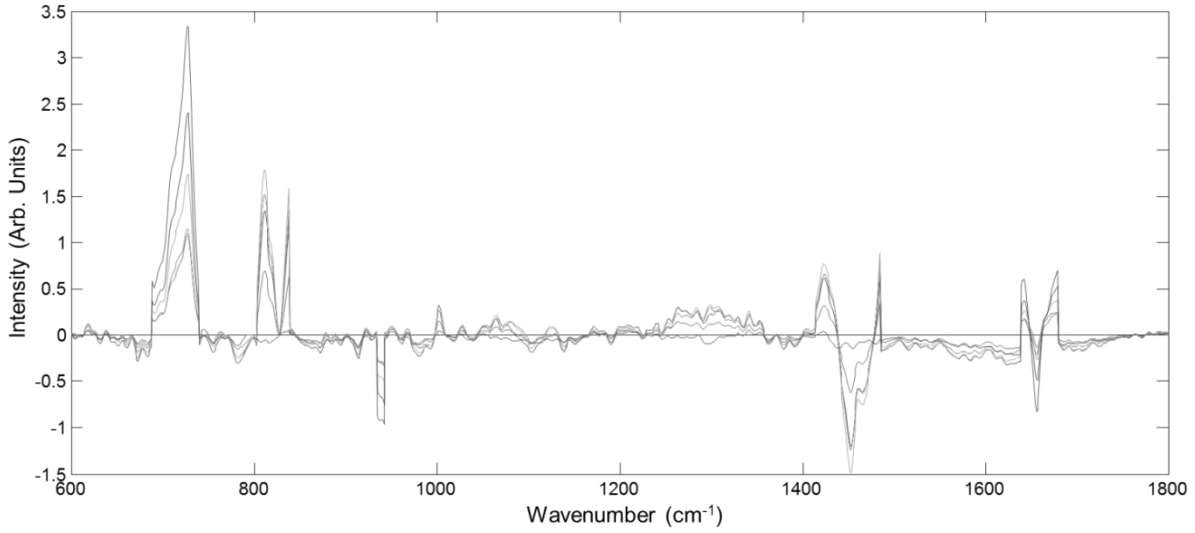

Figure S4. A plot of regression co-efficients following multiple regression of Dataset 2 against Lethal MTT with increasing data points. I.e. C+1 represents a dataset consisting of the control dataset and the data point at 0.05 µM. This then increases C+n until all data points in the dataset have been included.

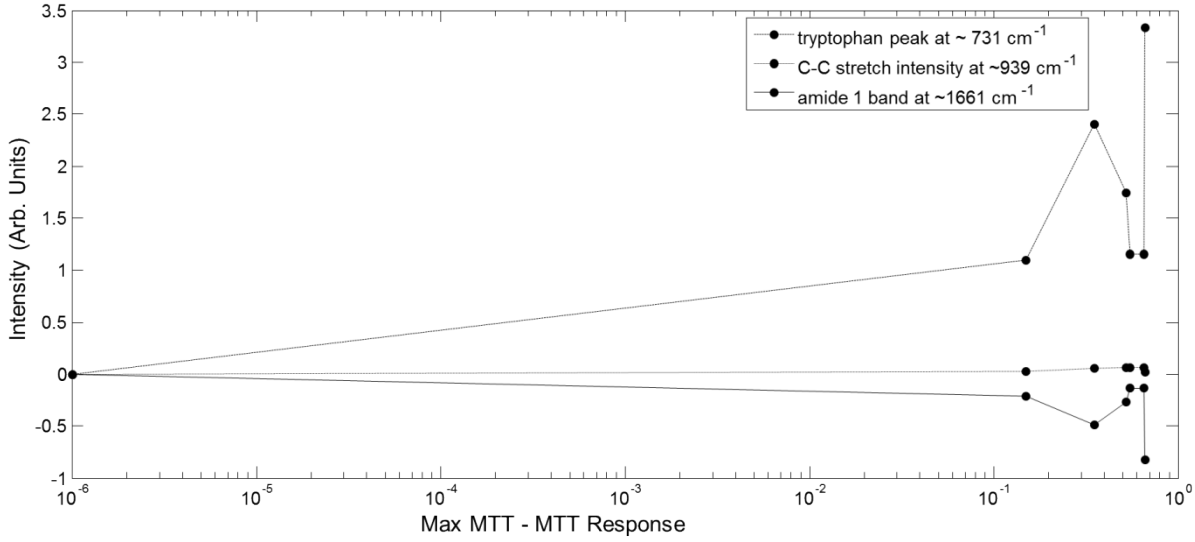

Figure S5 Plot of RC peak intensities for regression of Dataset 2 against Lethal MTT; C-C stretch intensity at ~939 cm^{-1}, the amide 1 band at ~1661 cm^{-1} and the tryptophan peak at 731 cm^{-1} of the Viability Construct (Figure 1B).

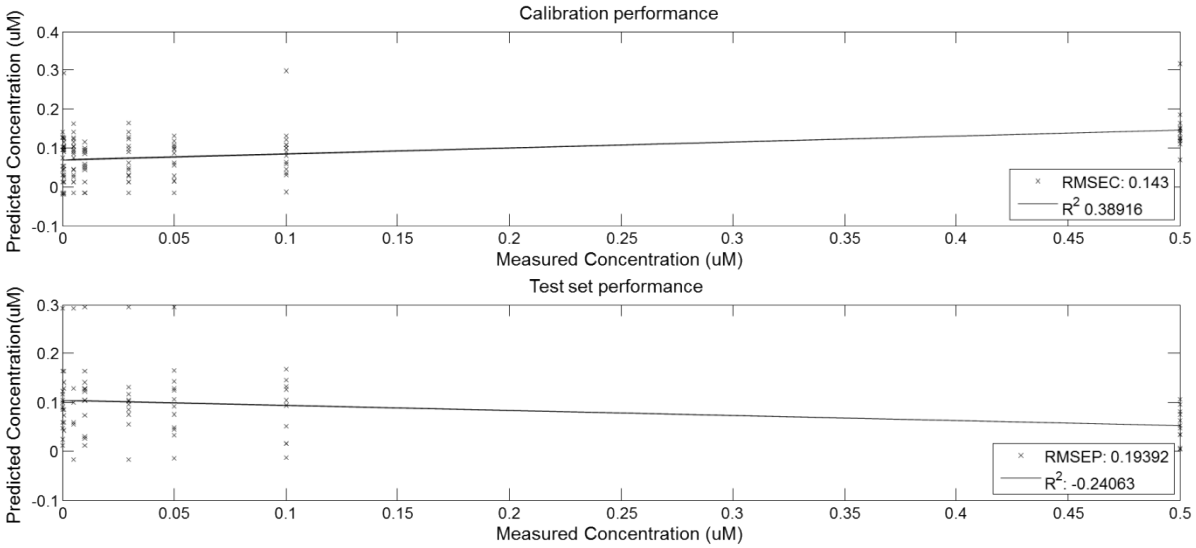

Figure S6. PLSR modelling of Dataset 3 with the Sub-lethal Concentration range as target. Top panel shows the calibration performance and test dataset (RMSEC 0.143, R2 0.38916). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.19392, R2 -0.24063). Data was split in a ratio of 60:40 calibration and test respectively.

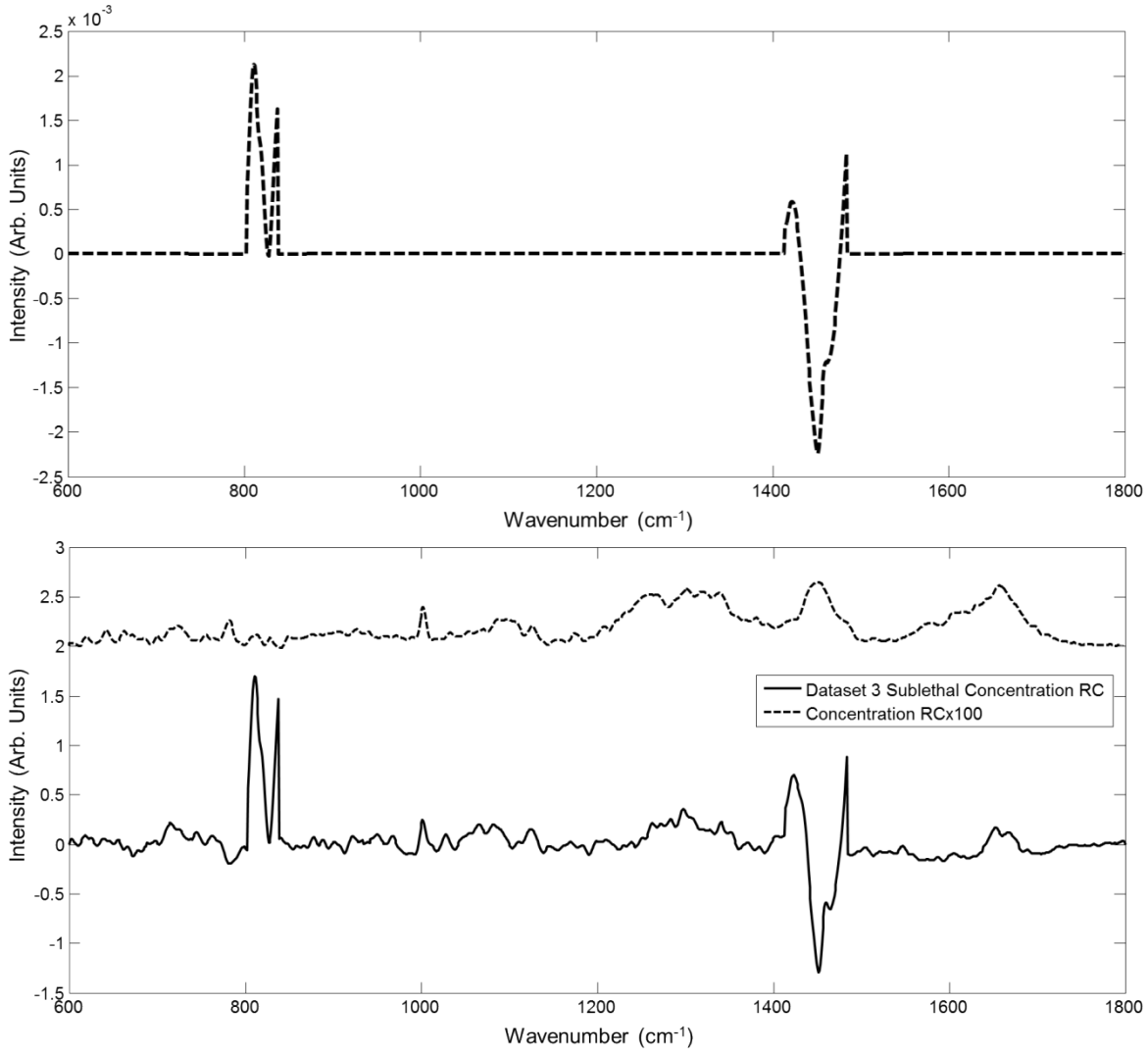

Figure S7. Plot of the regression co-efficients following PLSR of Dataset 3 against Sub-leathal Concentration . The concentration spectral construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The solid line shows the regression co-efficient following regression against sub-lethal concentration and Dataset 3 (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 sub-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 is offset and multiplied by a factor of 100 for clarity.