<u>**Supplementary Information**</u>

Rapid, high-throughput, and quantitative determination of orange juice adulteration by Fourier-transform infrared spectroscopy

David I. Ellis[a], Joanne Ellis[a], Howbeer Muhamadali[a], Yun Xu[a], Andrew B. Horn[b] and

Royston Goodacre[a]*

[a]School of Chemistry and Manchester Interdisciplinary Biocentre, University of Manchester,

131 Princess Street, Manchester, M1 7ND, UK

[b]School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL

Gas chromatography-mass spectrometry (GC-MS)

GC-MS followed our standard operating procedures that have been developed for MS-based metabolomics[1]. Briefly, for derivatisation samples were dried and redissolved in 50 µL of 20 mg.mL[-1] O-methoxylamine hydrochloride in pyridine, vortexed, and incubated at 80 ºC for 15 min in a dri-block heater. 50 µL N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) was then added and the extracts incubated at 80 ºC for a further 15 min. The derivatised products were then centrifuged at 15,800 g for 15 min and 90 µL of the supernatant transferred to GC-MS vials for analysis.

GC-ToF-MS analyses were carried out using a Leco Pegasus III (4D) GCxGC-MS in GC-MS mode (Leco Corp., St. Joseph, MO), with a Gerstel MPS-2 autosampler (Gerstel, Baltimore, USA) and an Agilent 6890N Gas Chromatograph with a split/splitless injector and Agilent LPD split-mode inlet liner (Agilent Technologies, Stockport, UK). A 30 m x 0.25 mm x 0.25 µm VF17-MS bonded phase capillary column (Varian, Oxford, UK). The conditions that we used for GC-MS are as reported in Begley et al.[2]

For quantitative analysis a series of sugar standards were prepared for glucose, fructose and sucrose. These were diluted, derivatised and analysed by GC-MS. The peak areas for the three sugars were used to construct calibration curves for the quantification of these sugars in freshly squeezed orange juice.

We found that 100 mL of our freshly squeezed orange juice prepared from South African variety Valencia oranges contained 2.45 g glucose, 5.90 g sucrose and 1.90 g fructose.

1   Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J.D., Halsall, A., Haselden, J.N., Nicholls, A.W., Wilson, I.D., The Husermet consortium, Kell, D.B. & Goodacre, R. Nat. Protoc., 2011, **6**, 1060-1083.

2 Begley, P., Francis-McIntyre, S., Dunn, W.B., Broadhurst, D.I., Halsall, A., Tseng, A., Knowles, J., Husermet consortium, Goodacre, R. & Kell, D.B. *Anal. Chem.*, 2009, **81**, 7038-7046.
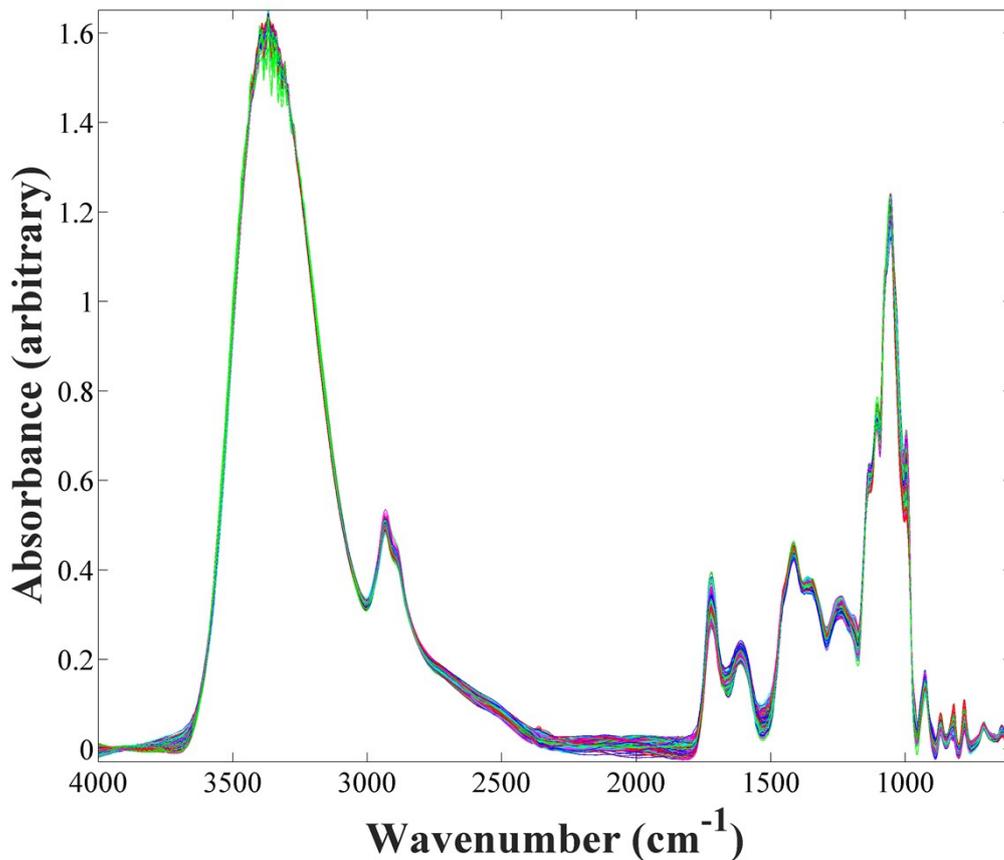
Figure S1:   Raw FT-IR spectra from the whole experiment. These spectra include pure orange juice plus 0.5 − 20.0 % adulteration with water disguised with either glucose, fructose or sucrose individually. The reproducibility is evident in that all 363 spectra are highly consistent. Note that the O-H vibration (broad peak centred at 3370 cm⁻¹) has some noise associate with it. This is due to water rotations because it was not possible to remove all the water in these sugar containing samples by drying in an oven at 50°C.
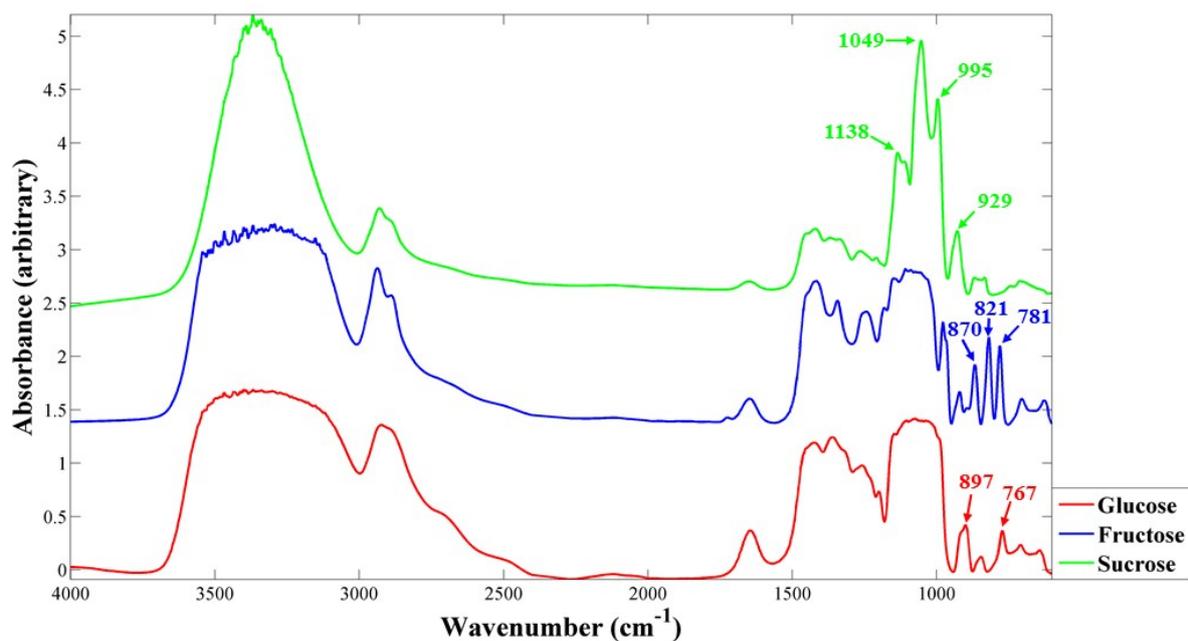
Figure S2:   FT-IR reference absorbance spectra of the pure sugars (glucose, fructose and sucrose) in water at 20% (v/v) concentration. These are collected off dried solutions on Si transmission plates. Note that the 'noise' is again due to water rotations (se Figure S3).
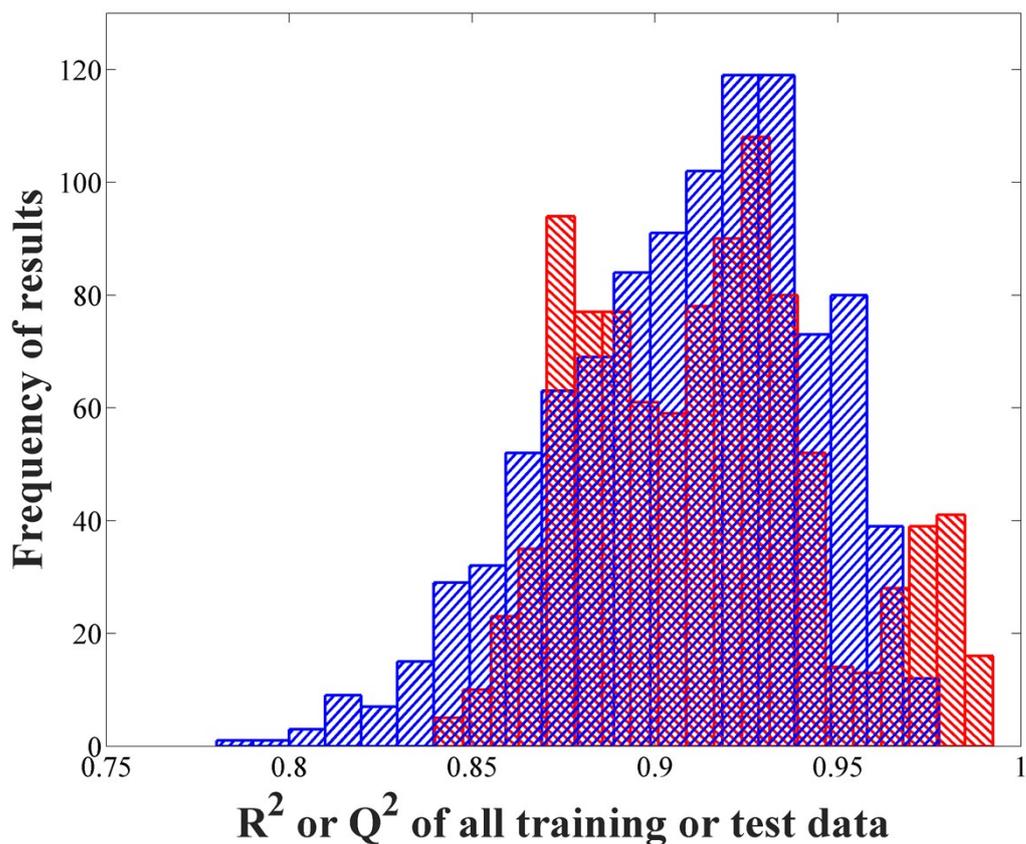
Figure S3:   Overlaid histogram frequency plots from the 1,000 bootstrap resamplings in
PLSR. The two distributions for the correlation coefficients from the training
(R^2; in red) and test sets (Q^2; in blue) are shown. These histograms illustrate
that the models are accurate as they are centred around 0.9142 for the training
set and 0.9081 for the test set. Moreover, as the two distributions for the
training and test sets overlap this indicates that the models are not overfitting.
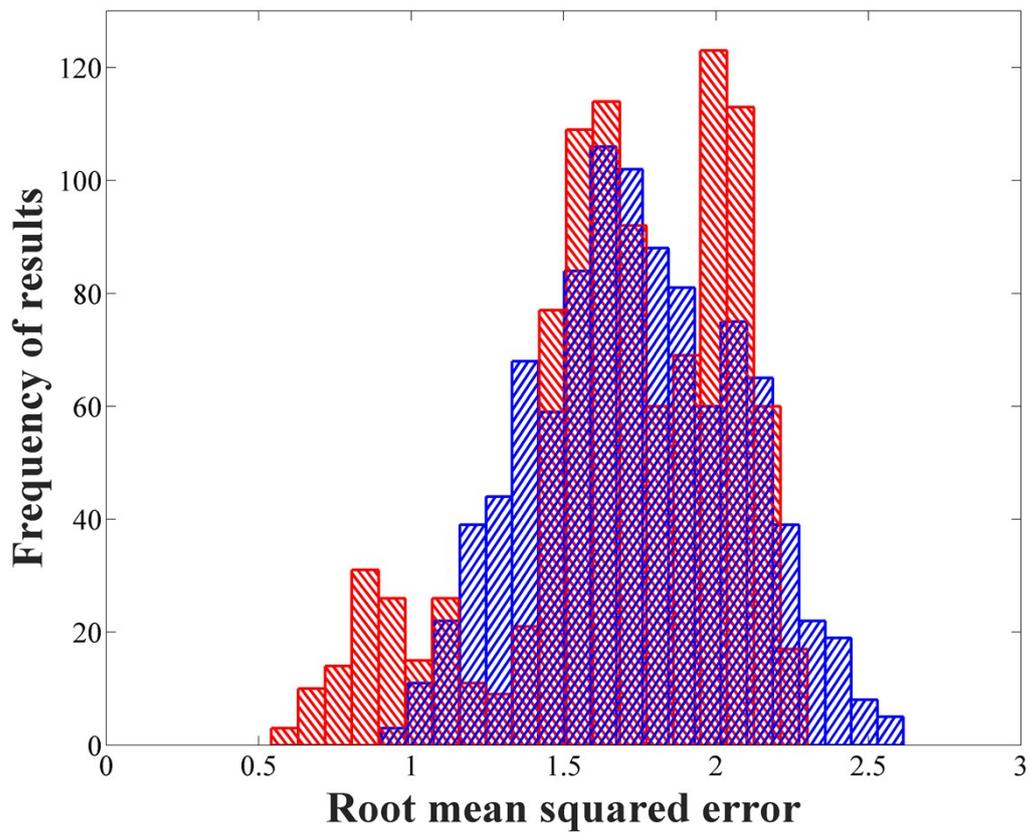
Figure S4:   Overlaid histogram frequency plots from the 1,000 bootstrap resamplings in
               PLSR. The two distributions for root mean squared (RMS) error are shown for
               the training (RMSECV; in red) and test sets (RMSEP; in blue) are shown.
               These histograms illustrate that the models are accurate as they are centred
               around 1.6851 for the training set and 1.7391 for the test set.