Raman mapping and hyperspectral data analysis

A study of in vitro cellular response to titanium dioxide and goethite nanoparticles

LINNEA AHLINDER, SUSANNE WIKLUND LINDSTRÖM,
LARS ÖSTERLUND

FOI, Swedish Defence Research Agency, is a mainly assignment-funded agency under the Ministry of Defence. The core activities are research, method and technology development, as well as studies conducted in the interests of Swedish defence and the safety and security of society. The organisation employs approximately 1000 personnel of whom about 800 are scientists. This makes FOI Sweden's largest research institute. FOI gives its customers access to leading-edge expertise in a large number of fields such as security policy studies, defence and security related analyses, the assessment of various types of threat, systems for control and management of crises, protection against and management of hazardous substances, IT security and the potential offered by new sensors.

FOI
Defence Research Agency
CBRN Defence and Security
SE-901 82 Umeå

Phone: +46 90 10 66 00
Fax: +46 90 10 68 00

www.foi.se

FOI-R--3126--SE Scientific report
ISSN 1650-1942 December 2010

CBRN Defence and Security

Linnea Ahlinder, Susanne Wiklund Lindström,
Lars Österlund

Raman mapping and hyperspectral data analysis

A study of in vitro cellular response to titanium dioxide and goethite nanoparticles

Titel	Raman-mappning och hyperspektral dataanalys: En in vitro-studie av det cellulära svaret från titandioxid och götitnanopartiklar
Title	Raman mapping and hyperspectral data analysis: a study of in vitro cellular response to titanium dioxide and goethite nanoparticles
Rapportnr/Report no	FOI-R--3126--SE
Rapporttyp Report Type	Vetenskaplig rapport
Sidor/Pages	48 p
Månad/Month	December
Utgivningsår/Year	2010
ISSN	
Kund/Customer	Försvarsdepartementet
Projektnr/Project no	A4031
Godkänd av/Approved by	Anders Norqvist

FOI, Totalförsvarets Forskningsinstitut

Avdelningen för CBRN-skydd och säkerhet

901 82 Umeå

FOI, Swedish Defence Research Agency

CBRN Defence and Security

SE-901 82 Umeå

Sammanfattning

Rapporter om att nanomaterial kan ge upphov till negativa hälsoeffekter efterlyser mer systematiska studier och utveckling av nya referensmetoder för att bedöma relevanta egenskaper hos partiklarna och för att kunna fastställa lämpliga riktlinjer. Studier av upptag och cellulärt svar från nanopartiklar är viktiga och av praktiska och etiska skäl är relevanta in vitro-baserade analyser att föredra för en första screening.

Konfokal Ramanspektroskopi är en inmärkningsfri teknik som här används för att studera nanopartikelexponerade lungepitelceller (A549). Tekniken ger kemiskt selektiv identifiering av biologiska och oorganiska föreningar med en spatial upplösning ned till ~ 1 μm^3 i levande celler.

Vi rapporterar här om det cellulära svaret från titandioxid- (TiO_2) och götitnanopartiklar (α-FeO(OH)) i A549 lungepitelceller som exponerats för partiklar under olika exponeringstider. Data har samlats in i flera områden av cellerna för att därmed skapa hyperspektrala bilder ur vilken information extraheras med hjälp av hyperspektral dataanalys. Möjligheten att skilja mellan molekylära vibrationer från DNA, proteiner och membran hos kontrollceller och partikelexponerade celler och kvantitativt klassificera det spektrala svaret genom hyperspektral multivariat dataanalys diskuteras.

Nyckelord: Raman-mappning, Konfokal Ramanspektroskopi, Hyperspektral dataanalys, PLS-DA, Nanotoxikologi

Summary

Reports that new *engineered nanomaterials* may cause adverse health effects calls for more systematic studies and development of new reference methods to assess relevant properties and define appropriate guidelines. It is important to perform studies of cellular uptake and cellular response to nanoparticles. From practical and ethical viewpoints relevant in vitro based assays is preferable for initial screening purposes.

Confocal Raman spectroscopy is a label-free technique which here is used to study nanoparticle exposed lung epithelial cells (A549). The technique provides chemically selective identification of biological and inorganic compounds and intracellular distributions in living cells down to ~1 μm^3 spatial resolution.

Here we report on cellular response to titanium dioxide (TiO_2) and goethite (α-FeO(OH)) nanoparticles in A549 cells subjected to varying times of exposures. Data is here collected in several parts of the cells thus forming a *hyperspectral image* from which information is extracted using hyperspectral data analysis. The possibility to discriminate between fundamental molecular vibrations originating from DNA, proteins and membranes on control cells and particle exposed cells and quantitatively classify the spectral response by *hyperspectral multivariate data analysis* is discussed.

Keywords: Raman mapping, Confocal Raman spectroscopy, Hyperspectral data analysis, PLS-DA, Nanotoxicology

Table of contents

1 Introduction

A popular definition of a nanoparticle is a particle with at least one dimension smaller than 100 nm. Today it is possible to make nano-sized particles, with different shape and composition with great precision, which can be used in variety of applications, including biology and medicine. The reason why nanoparticles are so useful is the small size, which gives the particles unusual properties such as high surface area, modifications of surface structure, shape, solubility, aggregation and electronic properties, which is relevant in the present context (quantum effects). The small size is however not wholly positive, since the particles easily may be taken up by humans and penetrate into cells and give rise to adverse health effects (Nel et al. 2006, p.622). It is therefore urgent to investigate the toxicity of nanoparticles. This can be done *in vivo*, but for screening purposes and ethical reasons, complementary *in vitro* methods are desirable. In this report, the possibility to study the effects of nanoparticles on living cells using Raman microspectroscopy together with hyperspectral data analysis is investigated. In this study, human lung epithelial cells (A549) were exposed to nanoparticles and compared to control cells. We choose lung epithelial cells as model systems since they represent cells found in the alveolar system which is the main entry pathway of nanoparticles in humans.

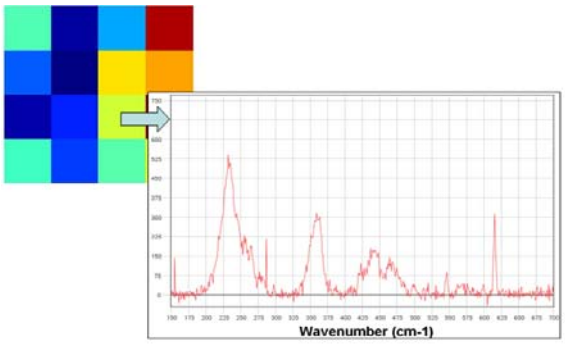

Figure 1 A hyperspectral image made up by 16 pixels. Each pixel contains a Raman spectrum.

Raman microspectroscopy is a non-invasive technique, which can be used for measurements on living cells with a minimum of sample pre-treatment. With this technique one can measure several spots in a sample in two dimensions sequentially in one run to construct a Raman "map" [1] or a *hyperspectral image* (Figure 1), i.e. an image with spectral information from many wavelength bands in each pixel (Baena & Lendl, 2004, p.534-535). With confocal Raman microspectroscopy it is in addition possible to perform three-dimensional Raman mapping. The data matrix obtained from mapping is here unfolded to a two dimensional data matrix and the data is then analyzed multivariately by principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). Multivariate methods are useful when the data to be analyzed consists of large and complex data matrices, which may contain missing data, noise or multicollinear variables. (Eriksson et al. 2006, pp.23-25)

Selected nanoparticles investigated in this project are titanium dioxide (TiO_2) and goethite (α-FeO(OH)) (Figure 2). TiO_2 has a wide range of application. It is primarily used as pigment in e.g. paints, cosmetics, sun screen, food, etc. TiO_2 has for a long time been considered as toxicologically inert and has been used as negative control in both *in vivo* and *in vitro* studies (Hext et al. 2005. pp.461-462), but there are also studies which show

[1] Sequential acquisition of spectra in different spatial location is usually called mapping to distinguish it from imaging, which is reserved for simultaneous acquisition of spectra in different spatial locations. The latter is typical done with Fourier-transform infrared spectroscopy (FTIR) and magnetic resonance (MRI) techniques.

that TiO_2 nanoparticles also cause adverse health effects. Gurr et al. have shown that TiO_2 nanoparticles cause oxidative damage of DNA in human bronchial epithelial cells (2005. pp.66-73) and Oberdörster et al. have shown that nano-sized TiO_2 causes lung inflammation in rats (1992, p.198).

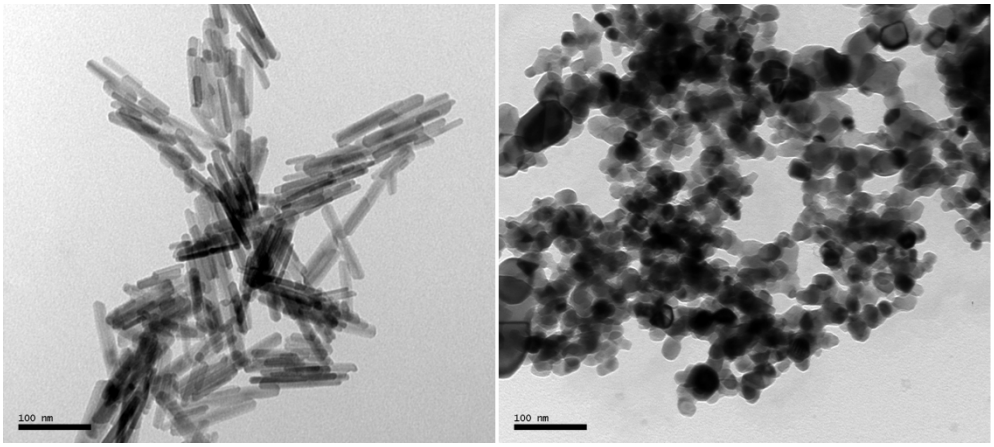

Figure 2 Goethite nanoparticles (left) and TiO2 nanoparticles (right).

Goethite is an iron oxide mineral present in soil and sediments and it is also a component of rust (Manceau et al. 2000, p.3643; Suh et al. 2009, p.153). Toxicological studies have so far mostly been concentrated on iron oxides other than goethite. However, since the structure of goethite nanorods (Figure 2) resembles the structure of carbon nanotubes and the disreputable crocidolite in amphibole asbestos, goethite may, from morphological reasons, be a suspected candidate to give adverse effect on cells. There are several studies which show that carbon nanotubes are toxic. Casey et al. (2008, p.83) have shown carbon nanotubes to induce indirect cytotoxicity in A549 cells.

1.1 Objective

The hypothesis of this project is to investigate if it is possible to discriminate between Raman signals from control cells and nanoparticle exposed lung epithelial cells and possibly also be able to quantitatively classify the spectral response by hyperspectral multivariate data analysis. The project includes data collection, investigation of pre-treatment techniques and multivariate analysis of the spectroscopic data.

2 Theory

2.1 A549 lung cells

A549 cells are a human type II alveolar epithelial cell line. Vibrational frequencies and peak assignments of Raman spectra from A549 cells are shown in Table 5, Appendix 1 (Notingher et al. 2002, p.233).

2.2 Confocal Raman microspectroscopy

In Raman spectroscopy, monochromatic light, from a laser source, passes through a sample and the scattered radiation is analyzed. Most of the light is absorbed or passes through the sample unaffected. A small part of the light (≈1/1000) is scattered in all directions without changing the energy, $E=h\nu_0$, of the light (Rayleigh radiation). A tiny fraction of this scattered light (≈1/1000 of the scattered or ≈1/106 of the total intensity) does however interact with the molecules in the sample. This changes the energy of the light such that $\Delta E=h(\nu_0\pm\nu_i)$, corresponding to the vibrational energy levels i of the molecules. The photons can either lose some of their energy (Stokes radiation) or collect energy from already excited molecules (anti-Stokes radiation) (Figure 3). The laser frequency, used to irradiate the sample, ν_0 can be in the UV, visible or infrared region, but the difference of the frequency of the irradiated light and the scatter light $|\nu_0\pm\nu_i|$ is an infrared (vibrational) frequency. Hence, Raman spectroscopy is a vibrational spectroscopy method and is sensitive to molecular vibrations. In the present study we employ either near infrared (λ=785 nm) or visible (λ=514 nm) laser light. To be Raman active, the molecules need to have a polarizability that changes because of rotations or vibrations in the molecule (Atkins & de Paula, 2005, pp.481-504). The classical explanation of the Raman process is that the electric field of light influences the electric fields within the molecule i.e. induces a dipole in the molecule. This in turn changes the frequency of light. The induced dipole moment, μ_{ind}, is proportional to the field strength:

$$\mu_{ind} = \alpha E \qquad (1)$$

where the constant α (units m3) is the polarizability of the molecule. Larger molecules typically have larger α, but depend on the specific bond that is excited. All molecules have non-zero polarizability even if they have no permanent dipole moment. For this reason, Raman scattering can measure the vibrational frequencies of molecules that will not absorb infrared light and is in this sense complementary to infrared spectroscopy. An advantage in biological studies is that water is a molecule that only gives a weak Raman signal and hence does not interfere with signals from other molecules (Pyrgiotakis et al. 2009, p.1465).

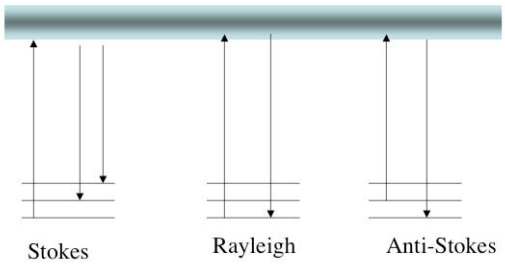

Figure 3 Raman (Stokes and anti-Stokes) and Rayleigh scattering. (Vanderkooi, 2006, p.5)

The Raman signal is however typically very weak and renders analysis of small amounts of absorbents difficult. The Raman signal is much improved if laser rejection filters, which filter the comparatively strong Rayleigh radiation, are used (McCreery, 2000, pp.1-9). This reduces contribution from notorious fluorescence in biological materials, which can be ~10^6 times stronger than the Raman signal. The Raman signal can be further increased if cooled multichannel detectors, for example charge coupled devices (CCD) are used, because they detect a range of wavelengths simultaneously, which reduces the run time and improves signal-to-noise (S/N).

In this report, confocal Raman microspectroscopy is used (Figure 4). The Raman microspectroscope consists of a microscope connected to a Raman spectroscope. The laser is focused through the objective of the microscope and the scattered photons are then collected by the same objective. The light passes through a confocal hole on the way to a CCD detector (Figure 4). The confocal hole is an adjustable hole that separates light which is in focus from light that are outside the focal point (Baena & Lendl, 2004, p.535). A small hole gives a focused signal, which increases the resolution in the Z-direction (Horiba JobinYvon, HR800 User Manual, p.35).

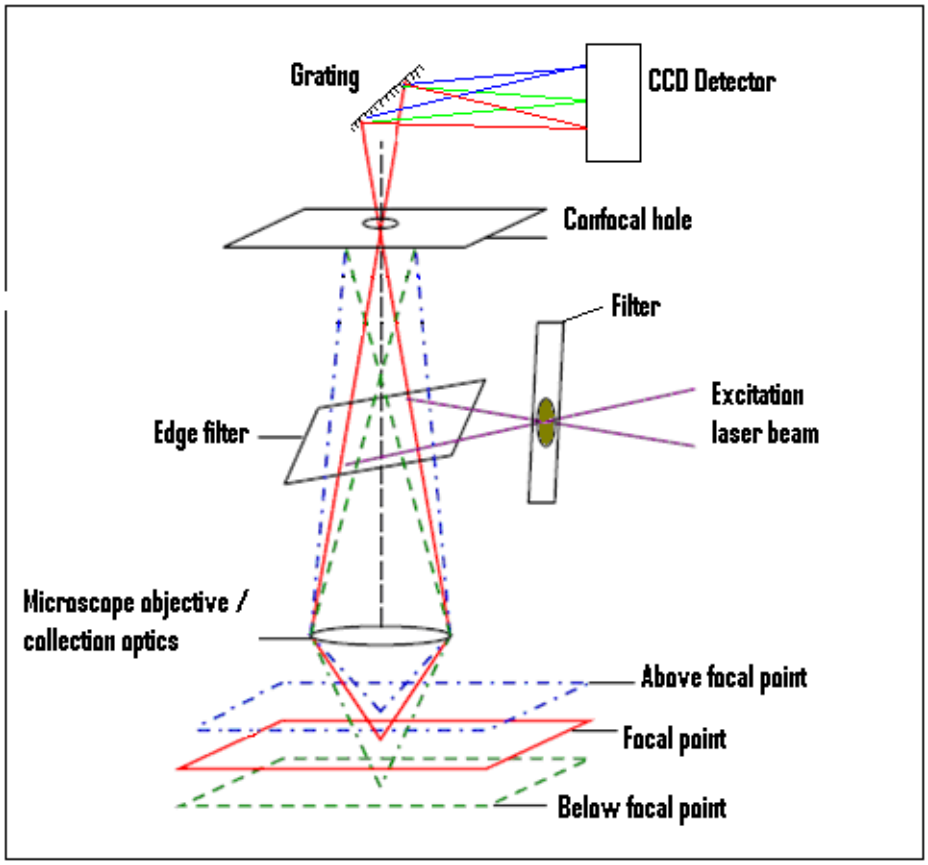

Figure 4 Schematic overview of a confocal Raman microspectroscope (modification of copyright material from Kaiser Optical Systems, 2009).

2.3 Data pre-treatment

All data contain varying amounts of noise and there are often peaks hidden in the background. If the sample contains organic molecules, the background can be especially troublesome due to fluorescence (Zhang, Z-M et al. 2009, p.1). It is therefore necessary to pre-treat the data, for example smooth the spectrum and remove the background. Other necessary pre-treatments may be normalization or frequency calibration. Spectral shifts, so called alignment problems, can for example occur when the laser or the laser rejection filter is replaced, if the temperature is decreased or increased or when the laser is switched off (Witjes et al. 2000, pp.105-106; Swierenga et al. 1999, pp.3-4). Such shifts are small, but also small shifts can give misleading results in a multivariate data analysis and it is therefore important to correct for these spectral variations (Witjes et al. 2000, p.105-116). Normalization is necessary if there are fluctuations of the laser or a drift in the laser power. Fluctuations and drift often occur because of the sensitive optics and since the Raman signal is directly proportional to the laser power, intensity variations will be seen in spectra collected on different occasions. (Cooper, 2009, p.244)

2.3.1 Background-correction

Two different methods were tested: a baseline correction and the baselineWavelet-algorithm implemented in R language by Zhang, Z-M et al. (2009). For baseline correction, a straight line is calculated between two user defined points, or the mean values between a number of user defined points. The spectrum is then projected to this line and brought to baseline.

The baselineWavelet-algorithm detects the peaks by using continuous wavelet transform (CWT) with the Mexican hat wavelet as mother wavelet and estimates the peak widths using CWT with the Haar wavelet as mother wavelet. The background is subsequently fitted using penalized least squares with binary masks. The R-code is in Appendix 2.

2.3.2 Normalization

Normalization is a procedure where the intensity of the spectra is corrected, for example by dividing peaks by the intensity of an internal standard. Here is no internal standard used and there are no peaks that can be expected to have equal intensity in all measurements, so it is not possible to use a peak in the spectra for normalization. An alternative normalization procedure is vector normalization, where all vectors are divided by the Euclidian norm:

$$\mathbf{x}_{normalized} = \frac{\mathbf{x}}{\| \mathbf{x} \|} \qquad (1)$$

In vector normalization, all vectors are normalized to unit vector length, which means that all intensity variations between spectra are eliminated, but the original shape of each spectrum is kept. (Schmid et al. 2009, p.162). Since intensity differences may reflect concentration differences, the vector normalization may remove important spectral information. It is however considered necessary to vector normalize to eliminate all intensity differences from the fluctuating laser. This process has been applied on Raman spectroscopic data before by Schmid et al. (2009) and Zhang, L et al. (2005). Swierenga et al. (1999) concluded that vector normalization gives small prediction errors in PLS models if it is used together with Savitzky-Golay smoothing (pp.14-15).

2.3.3 Calibration

The Raman microspectroscope was here regularly calibrated against the 520.7 cm^{-1}-peak from Si. Small remaining spectral shifts were corrected by internal calibration against the peak at 322 cm^{-1}, originating from the supporting CaF_2 (see section 3). This was done by finding the maximum of a parabola fitted to eleven points around the peak maximum. All spectra were shifted accordingly.

2.3.4 Savitzky-Golay smoothing

Savitzky-Golay smoothing is a smoothing method, based on a least squares procedure. A number of points before and after the number to be smoothed (the Window size) are selected and the point is then fitted to a polynomial of a selected degree. In that way, the amplitude of the noise is reduced with minimal distortion of the peaks containing information (Savitzky & Golay, 1964, pp.1627-1639).

2.4 Multivariate data analysis

In large data sets, it can be of use to apply projection methods, a sort of multivariate data analysis, which aim to reduce the number of variables without loss of important information. There are many applications and the analysis can be applied to get an overview of data, as well as for classification or prediction purposes. In contrast to classical statistics, the variables do not need to be independent and the data may also contain noise or missing values. The data in multivariate data analysis is usually arranged in matrices, where each row corresponds to an observation (measurement) and each column corresponds to a variable. (Eriksson et al. 2006, pp.8-29)

2.4.1 Hyperspectral data analysis

In traditional spectroscopic measurements, where hyperspectral data analysis has been applied, such as near-infrared spectroscopy (NIR), the measurements have been performed on single spots. This method works as long as the sample is homogenous and the sample spot represents the whole sample. Today, there are spectroscopic instruments that can record spectral information at different spatial locations on the sample, either simultaneously (imaging) or sequentially (mapping). These instruments can record several spots, or pixels, which thus make up a multispectral or hyperspectral image (Figure 1). The difference between multispectral and hyperspectral images is not clearly defined, but a multispectral image is an image with four or more wavelength channels and hyperspectral images contain usually 100 or more wavelength channels. (Burger, 2006, pp.1-4).

In hyperspectral data analysis, the information from such images is unfolded to a large data matrix and multivariate methods as principal component analysis (PCA) or partial least squares discriminant analysis (PLS-DA) can be applied. Small spots of the whole sample are, in that way, analyzed at the same time. The method has especially big advantages when the sample is complex or inhomogeneous (Burger, 2006, pp.1-4).

2.4.2 Principal component analysis

Principal component analysis is a multivariate method, which can be applied to a data matrix, **X**, to get an overview of the data and to see groups or trends in data. PCA-models can also be used for classification. The data is projected to new variables, principal components, by decomposing the data according to:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \qquad (2)$$

Here, **T** is the score matrix, **P** is the loading matrix and **E** contains the residuals (Geladi et al, 1989, p.211).

The columns in the score matrix are the principal components, of which the first is chosen to explain most of the variance in data and the following are orthogonal. (Eriksson et al. 2006, pp.46-47) A picture of the projection onto the two first principal components is shown in Figure 5. The picture also shows the distance to the model, which is the standard deviation of the residual (Eriksson et al. 2006. p.385). A critical value, which corresponds to a higher value than 95% of the observations' distance to model, is used to find outliers. If the PCA model is used for classification, observations below the critical value are considered as members of the class.

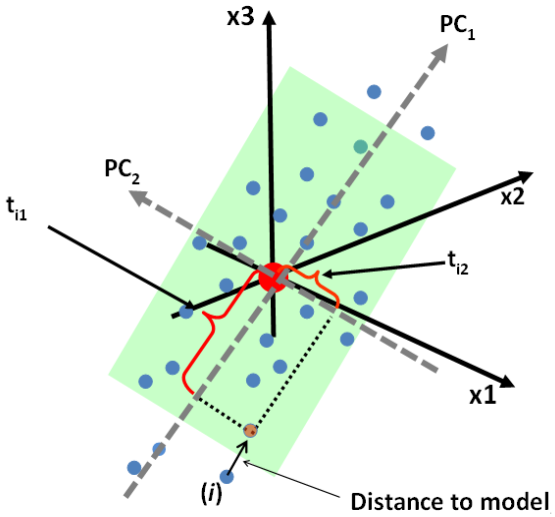

Figure 5 Principal component analysis. The picture shows the plane made up by principal component 1 (PC1) and principal component 2 (PC2) and the projection of observation (i) onto this plane. t_{i1} and t_{i2} are score values.

The projection can be visualized in a score plot, in which the observations are plotted in the new coordinate system, based on the principal components. The loading matrix contains information about the variables and its loading vectors can be plotted in a similar way, to visualize groupings among the variables. (Eriksson et al. 2006, p.33) Some of the observations may deviate in the projection. Such outliers are identified in a score plot, which often is combined with an elliptical 95% confidence interval (Hotelling T^2 ellipse) (Eriksson et al. 2006, p.391).

2.4.3 Partial least squares discriminant analysis

Partial least squares projection to latent structures (PLS) is a multivariate method, which can be used when there is not only a data matrix **X**, but also a matrix **Y** that contains response variables. PLS-models are mainly of use for prediction and classification.

The data is projected to new variables, score-vectors, which explain the main variation in data. Unlike PCA, these variables do not only describe the variation in **X**, but the variation in **X** and **Y** simultaneously. **Y** and **X** are decomposed as follows:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \quad (3)$$

$$\mathbf{Y} = \mathbf{TC}' + \mathbf{F} \quad (4)$$

E and **F** are residuals matrices, **T** is the score matrix, **P** is a loading matrix and **C** is a matrix containing the coefficients, the "weights", in the model. The scores can also be regarded as a linear combination of the original variables with coefficients **W**:

$$\mathbf{T} = \mathbf{XW} \quad (5)$$

(5) and (6) give the relationship:

$$\mathbf{Y} = \mathbf{XWC}' + \mathbf{F} \quad (7)$$

WC' is also referred to as PLS-regression coefficients (Wold et al. 2001, pp.109-115).

PLS discriminant analysis (PLS-DA) is a special case of PLS, where **Y** contains dummy variables, which assign the observations to classes. (Eriksson et al. 2006. pp.181-182) Examples of dummy variables in our study are variables which give information about particle exposure: "Control cell", "TiO_2" and "Goethite". In each variable, all observations are given discrete values, 1 or 0, depending on if the observation belongs to the class or not. Prediction cut off values are set for the classification. The lowest value was set to 0.5 and the highest value was set to 1.5. These cut off values means that all observations in the interval -0.5 – 0.5 are classified as not belonging to the class and all observations in the interval 0.5 – 1.5 are classified as members to the class. Observations outside the limits are not classified i.e. unknown class.

PLS-DA models can be evaluated by cross validation, which means observations are excluded and predicted by the model. It is common to exclude all observations one by one, i.e. leave-one-out cross validation, but since it is a time-consuming procedure, 1/7 of the data, evenly spread in the data set, were here excluded each round. The differences between the real values and the values predicted by the model are summarized in the predicted residual error sum of squares (PRESS) and used to calculate Q^2, which explains the predictive power of the model. (Eriksson et al. 2006. p.389)

$$PRESS = \sum (Y - \hat{Y})^2 \quad (8)$$

$$SS = \sum Y^2 \quad (9)$$

$$Q^2 = \frac{1 - PRESS}{SS} \quad (10)$$

Y is the observed value and $\hat{Y}$ is the value predicted by the model.

2.4.4 Pre-processing methods – mean centering and scaling to unit variance

Mean centering means subtracting the mean value for each variable from all observations. After mean centering, the center of the data is moved to origin of coordinates (Eriksson et al. 2006, pp.45-46), which makes it easier to interpret multivariate models. The interpretation is unaffected. (Wold et al. 2001, p.113).

Scaling to unit variance, UV-scaling, is a pre-processing method where all observations are divided by the standard deviations for each variable. The method can be regarded as a normalization procedure, where big differences in order of magnitude are removed and variables with high values are prevented from taking disproportionate significance in the model. This means for example that in a measurement from a complex sample, compounds with low concentrations and compounds with high concentrations will give equal contribution to the model after UV-scaling (Wiklund, 2007, p.37).

2.4.5 Variable selection

The **W**-matrix contains the weights for all variables. The uncertainty in the weights can be estimated by using the Jack-knifing procedure to calculate confidence intervals. The Jack-knifed confidence intervals are calculated from the standard error from each loading, based on cross-validation, and from a statistical value corresponding to a 95% confidence interval. These Jack-knifed confidence intervals are in some cases large and include zero. Such variables do most likely not contribute to important information and can be excluded. (Wiklund et al, 2008, pp.118-119)

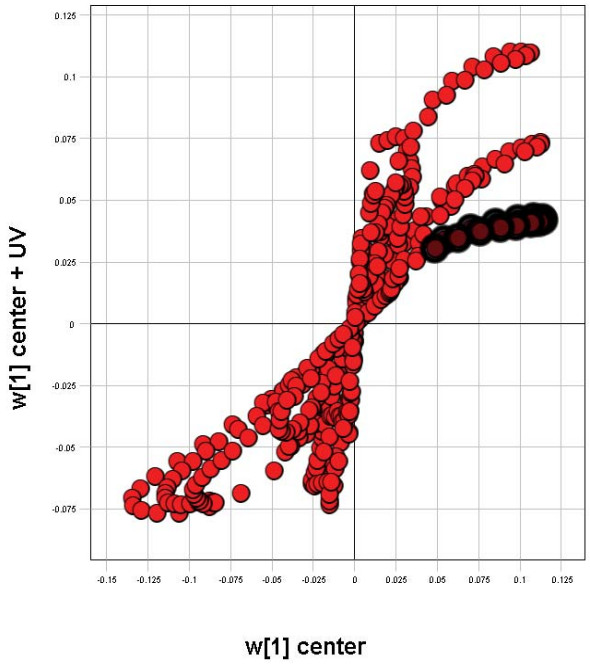

Figure 6 An alternative S-plot. Weights for PLS-model nucleus/cytoplasm, centered data on X-axis. Weights for PLS-model nucleus/cytoplasm, centered and UV-scaled data on Y-axis. Variables corresponding to the selected weights have been excluded.

Spectra may also have peaks with especially high intensities. Such peaks may have high impact in the model, but do not necessarily contain important information. These peaks can be identified in an alternative S-plot, where the weights from a PLS model, based on centered data, are plotted against the weights from a PLS model, based on centered and UV-scaled data. Variables that do not follow the ideal S-shape can often be excluded without loss of important information, which can be confirmed by high Jack-knifed confidence intervals. S-plots, where the covariance and correlation are combined in a scatter plot, have been used previously by Wiklund et al. (2008) for orthogonal PLS models. Wiklund et al. (2008) found that S-plots are useful in analysis of complex data, since they facilitate selection of variables that both have a high correlation and a low covariance, i.e. they are important in the model and do not origin from noise. An example of a S-plot is shown in Figure 6.

3 Method

3.1 Cell preparation

A549 cells (ATCC CCL-185; American Type Culture collection) were cultured in RPMI-1640 (Gibco BRL, Paisley, UK) supplemented with 10% fetal calf serum (FCS; Hyclone, Perbio Science, Aalst, Belgium) and 50 µg/ml gentamicin at 37° C in a humidified atmosphere with 5% CO_2. For Raman spectroscopy measurements, cells were seeded at density of $5*10^4$ cells/ml onto CaF_2 substrates in 12-well culture plates and allowed to attach over night before exposed to nanoparticles. Stock solutions of 1 mg/ml TiO_2 particles (P25) or goethite particles in phosphate buffered saline pH 7.2 (PBS) was performed. The samples were sonicated at +4 (Bransonic 221 sonicator) for 45 min and vigorously vortexed before diluted further to 10 µg/ml in cell medium. This was done to produce a stable and well dissolved nanoparticle suspension. After 24-72 h exposure to nanoparticles, the cell cultures were washed 5 times with 1 ml PBS to remove detachable nanoparticles. The CaF_2 substrates were subsequently transferred to 6-well culture plates prior to Raman spectroscopy measurements. Figure 7 shows a schematic picture of the sample presentation. A summary of the measured cells are in Table 4, Appendix 1.

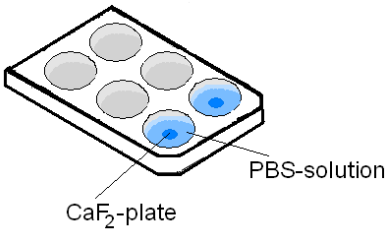

Figure 7 Schematic picture of the sample presentation. The A549 cells are placed on CaF2-substrates and covered with phosphate buffered saline (PBS) solution in 6-well plates.

3.2 Nanoparticles

Two types of nanoparticles were used: Titanium dioxide (TiO2) and an iron hydroxide, goethite (α-FeO(OH)). The TiO2 sample, denoted P25, was obtained from Degussa AG, Germany and contains mainly anatase nanoparticles with a primary particle size dp ≈ 21 nm. The α-FeOOH was prepared and characterized as described by Boily et al. (2001, pp.12-27) and Mäkie et al. and consist of approximately 11-16 nm wide and 62-120 nm long elongated nanoparticles (nanorods).

3.3 Raman mapping

Confocal mapping was performed on living cells on CaF_2 substrates in sample wells as shown in Figure 7. A schematic drawing of the confocal Raman microspectroscopy set-up is shown in Figure 4. Unless otherwise stated, each cell was measured at 16 different spots at a fixed focal plane (z-axis) which was defined by a 10.5*10.5 µm square grid. The scan time was set to 90*3 s in each spot resulting in a total measurement time of 72 min for all 16 points. The measured spots were chosen to cover as much as possible of the cell and to cover both the nucleus and cytoplasm regions. All spectra were recorded using a Horiba JobinYvon LabRam HR800 Raman microscope with a 60× water immersion objective and a thermoelectrically air-cooled CCD detector. The laser source employed was an Ar^+ laser (514 nm) operated at 12.5 mW. A 600 lines/mm-grating was used in the measurements and the confocal hole was set to 150 µm.

3.4 Data pre-treatment

Spikes, i.e. peaks that contain no spectral information from the sample and originate from spurious background radiation or bad pixels in the CCD, were removed manually by replacing their intensity values with the mean value for the intensity for the points on both sides of the spike.

A number of pre-treatment methods were tested on the spike-eliminated data: baseline correction, baselineWavelet, Savitzky-Golay smoothing, vector normalization and calibration. Data analysis was done in Evince Image, version 2.4.0. (UmBio, Umeå, Sweden), in MatLab 7.10.0. (The MathWorks, Nattick, USA) and in R 2.8.1 (The R Foundation for Statistical Computing, Vienna, Austria).

3.5 Hyperspectral data analysis

The multivariate data analysis was performed in Evince Image, version 2.4.0. (UmBio, Umeå, Sweden).

In the data analysis, all measurements from the cells were included – here referred to as *observations*. Furthermore, all data points in the spectral region between 730 cm^{-1} and 1800 cm^{-1} (which does not contain absorption bands due to nanoparticles) were included and are referred to as *variables*. The chosen spectral region contains 622 variables as dictated by the spectral resolution (1.72 cm^{-1}).

Of all measured cells, 24 were randomly picked out and included in the models, while the remaining 6 measured cells were used as an external test set for evaluation. All data is here mean centered before modelling and variables are selected using alternative S-plots and Jack-knifed confidence intervals.

4 Results and Discussion

4.1 Raman mapping of living cells – experimental considerations

In Raman mapping experiments, it is critical to optimize scan time and laser power before data collection. It is important to maintain high intensity without applying too high photon dose, especially when a small confocal hole is used, since a small confocal hole also limits the amount of light that reaches the detector and results in poor S/N. It is however not possible to employ too long run time because biological samples are sensitive to irradiation damage and laser heat dissipated may dehydrate the cells or induce irreversible thermal modifications.

The A549 cells used in this study are known to be robust. Notingher et al. have shown that they survive for at least 60 min exposure to λ=785 nm irradiation at 115 mW laser power (Notingher et al. 2002, pp.231-232). The effect of the laser used in this study is much lower (maximum 30 mW, λ=785 nm). Employing the 785 nm laser diode yielded however not good quality Raman spectra in our case. Instead, a 514 nm Ar-ion (0-50 mW) laser was used. The lower wavelength implies a higher energy which is potentially more likely to damage the cells. It is therefore necessary to carefully inspect the cells before and after measurements. For example, the cell adherence to the CaF_2 substrate is sensitive to the cell condition. Bad cells do often come off the substrate. Another important test is to inspect the activity inside the cell. A lot of movements imply activities such as cell division or cell death, and can be an indication that the cell does not feel well. Figure 8 shows an example of a cell that has come off the substrate. The membrane did in some cases burst (Figure 9), which clearly indicates damage.

A 12.5 mW laser power was chosen and 90*3 s in each measurement point was considered sufficient scan time to give good quality spectra with interpretable peaks.

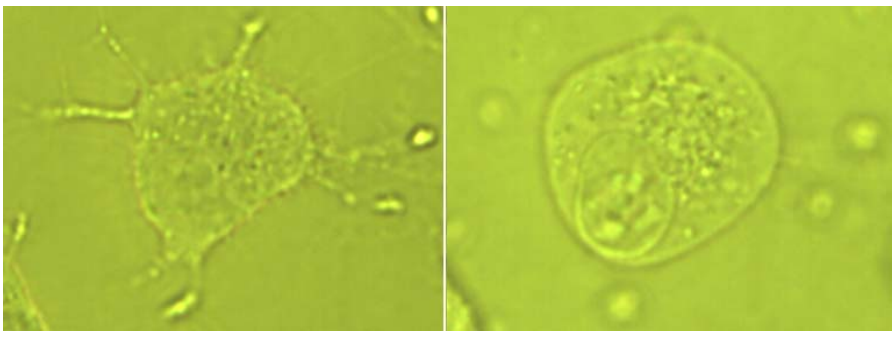

Figure 8 A549 cell before (left) and after (right) 2 h exposure to □=514 nm laser, 50 mW.

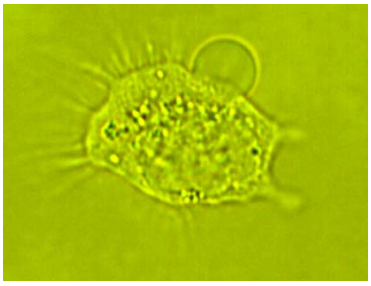

Figure 9 Cell with a damaged cell membrane after exposure to □=514 nm laser, 5-25 mW (photon dose: 5.0 * 10-4 E).

The measured points in the Raman mapping were chosen not to overlap. The diameter of an A549 cell is about 20 to 30 µm and, since a 4*4 grid was measured in each cell, a 3.5 µm distance between each measurement point was regarded necessary. To make sure the illuminated part of the sample is not larger than this distance, a confocal hole of maximum 165 µm must be used according to the relationship between the confocal hole and the illuminated part of the sample (Horiba JobinYvon, HR800 User Manual, p.38):

$$D_{\text{hole}} = D_{\text{iilluminated}} \cdot M_{\text{objective}} \cdot 1.4 \cdot 0.56 \qquad (11)$$

Here, D_{hole} is the diameter of the confocal hole, $D_{\text{illuminated}}$ is the illuminated part of the sample and $M_{\text{objective}}$ is the magnification of the objective (60 in this case). The size of the confocal hole was set to 150 µm in all measurements presented here.

The depth resolution for λ = 514 nm and confocal hole = 150 µm is about 4.57 µm (Horiba JobinYvon, Quality Control). To make sure the measurement was not performed outside the cell and to make sure the cell nucleus was in the mapping, the Z-value was optimized by measuring at different Z in the nucleus and choosing the Z that gave the largest intensity for the peaks at 669 cm^{-1}, 782 cm^{-1} and/or 788 cm^{-1}, which corresponds to signals from DNA/RNA. A comparison between a spot inside and a spot outside a cell nucleus is shown in Figure 10, where a clear difference can be seen at 782 cm^{-1}.

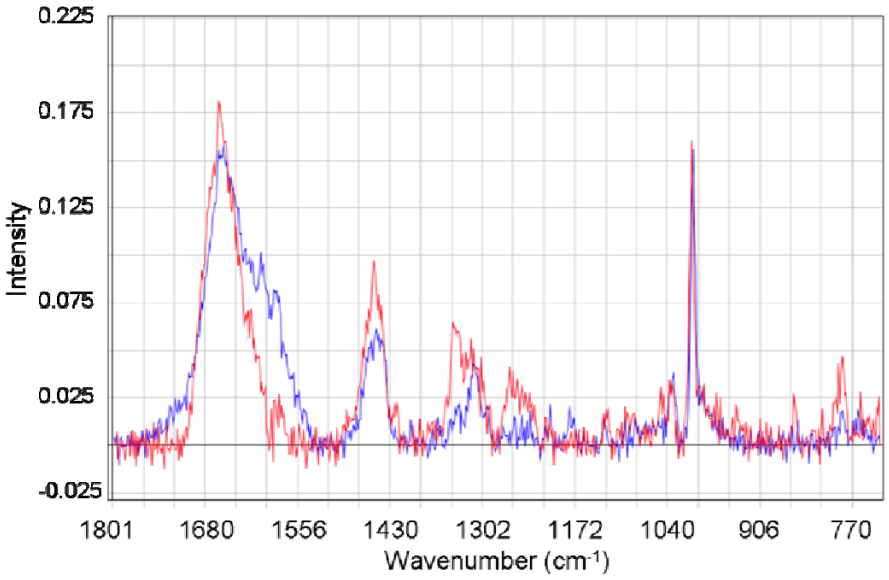

Figure 10 Comparison between spectrum from a sample spot inside the cell nucleus (red) and outside the cell nucleus (blue).

4.2 Data pre-treatment

4.2.1 Background-correction

The baseline correction procedure did not give adequate result in our case, mainly because of the difficulty to identify representative start- and end points to be used in the calculation of the correction line. Figure 11 shows an example of the baseline correction. It is evident from Figure 11 that reliable baseline correction is difficult.

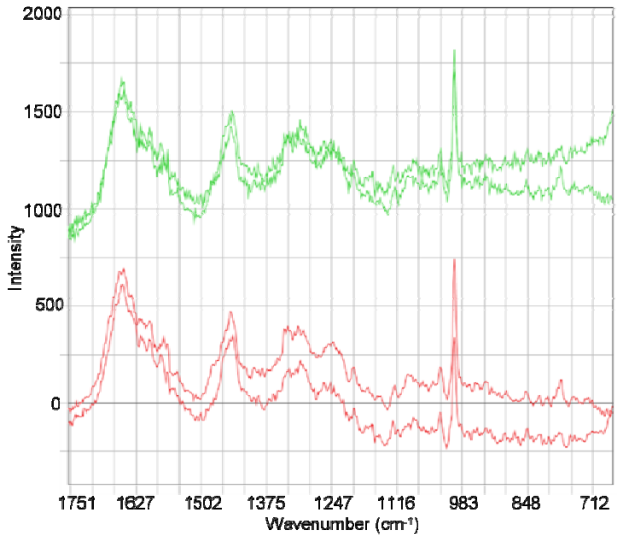

Figure 11 Spectra before (green) and after (red) baseline correction.

Instead the background was corrected by using the baselineWavelet-algorithm implemented by Zhang et al. (2009). The result is shown in Figure 12. It is found that the baselineWavelet algorithm yields a good and reproducible background correction for all

spectra without significant peak broadening and loss of spectral information. This method was applied on all spectra used for hyperspectral data analysis. R-code for the baselineWavelet background correction is in Appendix 2.

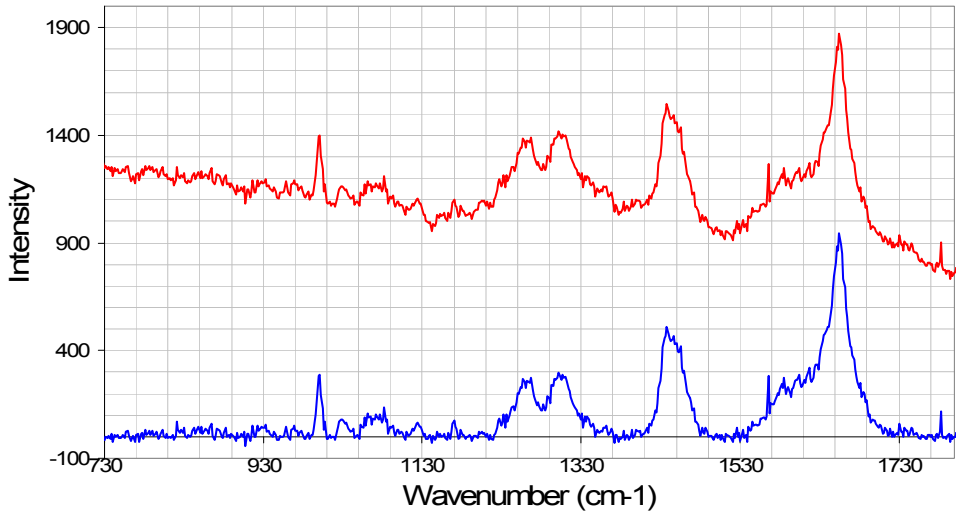

Figure 12 Background correction using the baselineWavelet-algorithm (Zhang, Z-M et al. 2009). Red – raw data. Blue – data after background correction.

4.2.2 Normalization

Spectra before and after vector normalization are shown in Figure 13 and Figure 14, respectively. Normalized data was used in all models, since it was considered necessary to normalize data to eliminate intensity differences caused by sample inhomogeneities (particles exposure/distribution and inter-cell variations), date of acquisition (laser fluctuations), cell culturing, and other factors that could not be held constant.

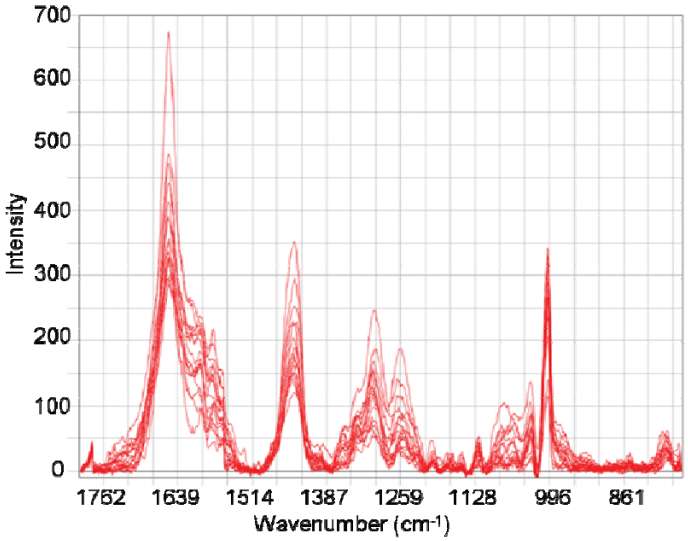

Figure 13 Spectra before vector normalization.

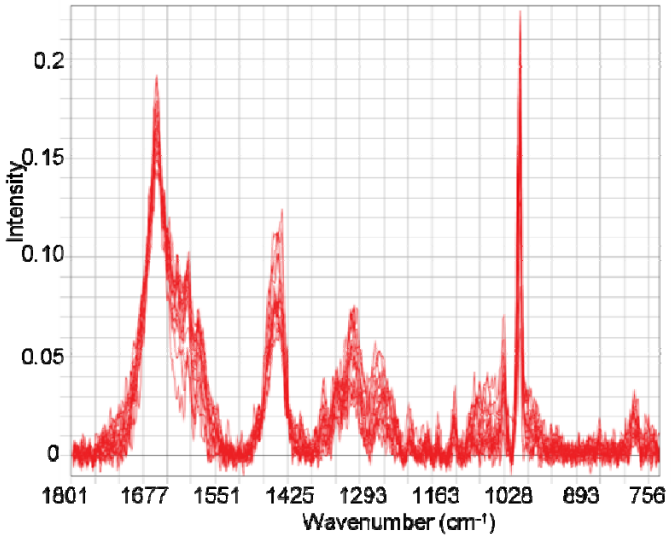

Figure 14 Spectra after vector normalization.

4.2.3 Calibration

An example of internal CaF_2 frequency calibration is shown in Figure 15. The picture shows the phenylalanine peak, which is a well isolated peak in the region used for hyperspectral data analysis. The correction is not perfect due to the finite frequency resolution (1.72 cm^{-1}), but a clear improvement is seen. All spectra were calibrated in this manner before data analysis.

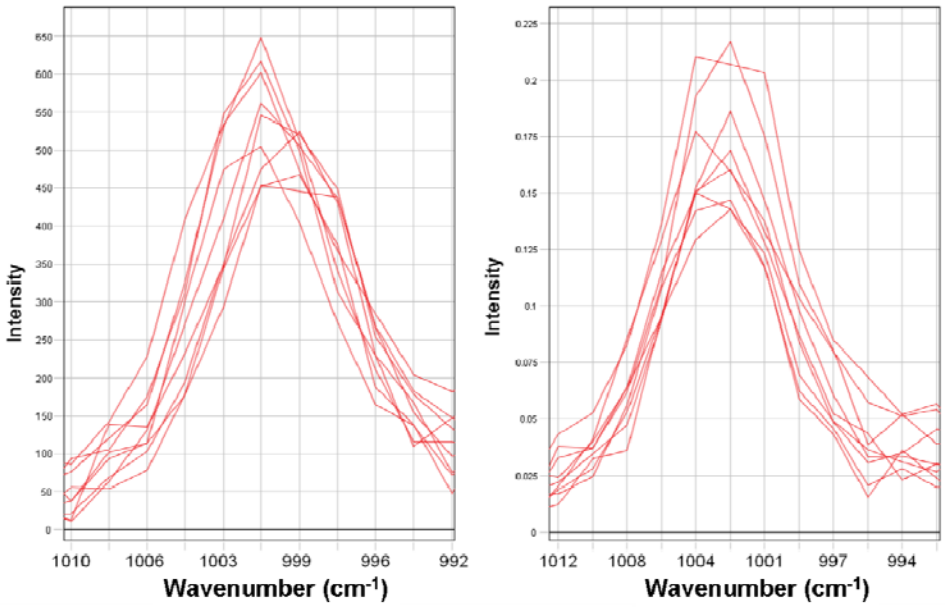

Figure 15 The phenylalanine peak before (left) and after (right) calibration and vector normalization.

4.2.4 Savitzky-Golay smoothing

The window size, the number of points before and after the point to be smoothed, was varied and the optimum window size was considered to be the size that gave minimum distortion of peaks. Optimum window size was found to be 5 points. Figure 16 shows

spectra before and after smoothing. Savitzky-Golay smoothing gives spectra that are easier to visual interpret, but models based on smoothed data had generally lower Q^2 compared to models based on non-smoothed data, probably because of distortion of peaks, especially sharp peaks, which are sensitive to smoothing due to the spectral resolution. Spectra were not smoothed before modelling, unless otherwise stated.

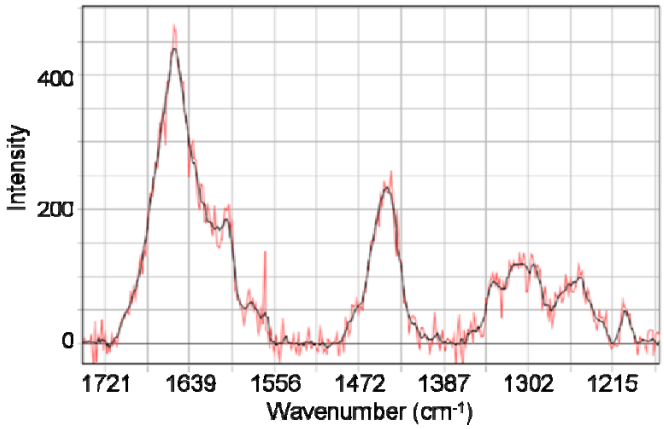

Figure 16 Spectrum smoothed by Savitzky-Golay algorithm. Red superposed spectrum shows raw data.

4.3 Identification of the cell nucleus

It is sometimes possible to distinguish the nucleus from the cytoplasm in optical microscopy (OM) images merely by inspecting the images. However, the OM images give no depth resolution and the precise 3D location is not possible to deduce based on OM only. Employing Raman spectroscopy, the nucleus can be identified by comparing the intensities in the vibrational bands that are unique for DNA/RNA. In particular the band at 782 cm^{-1} is comparable strong and well-isolated from other absorption bands. This peak has here been compared in Raman intensity maps and the measurements that showed the highest intensity were considered to belong to the nucleus. In most cases, these observations were found to correspond well to the nucleus identified in OM. Figure 17 shows an example of an optical image and corresponding intensity map, which shows the amplitude of the 782 cm^{-1} peak.

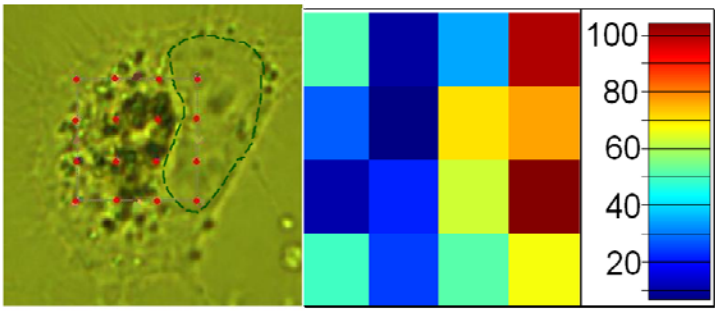

Figure 17 Left: OM image of Cell L, cell exposed to TiO_2. Cell nucleus is marked with a green, dashed line and red dots corresponds to the measured spots at fixed Z. Right: An intensity map, which shows the pixels in the hyperspectral image, colored according to relative intensity for the peak at 782 cm^{-1}.

Observations that belong to nucleus and cytoplasm, respectively, were identified by using both intensity maps and OM images. In cases where both the intensity map and OM image indicated that a certain measurement belongs to the same class (nucleus or cytoplasm), the

measurement was included in a PLS-DA model. 144 observations, of which 84 were classified as measurements from nucleus and 60 were classified as measurements form cytoplasm, were finally included in the PLS-DA. The data was smoothed with Savitzky-Golay algorithm (5 points, second-order polynomial). The model yielded a relatively low Q^2, 0.58, (see Table 7, Appendix 1) but score scatter plot (Figure 18) showed a clear trend in data. Most of the observations from nucleus were found to have positive values in component 1, while most of the observations from cytoplasm were found to have negative values in component 1.

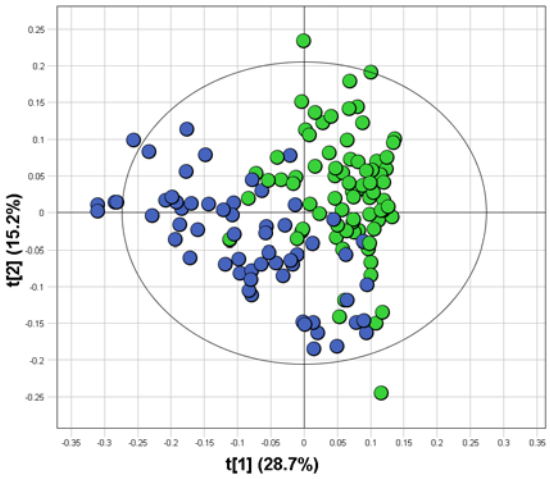

Figure 18 Score plot. PLS-DA model, nucleus (green)/cytoplasm (blue).

The weights in the final model are shown in Figure 19. As expected, the peak at 788 cm^{-1} has a large impact in the model, but an advantage of a multivariate model is that it includes all spectral features. In Figure 19 it is evident that the model finds important contribution to the model due to the peaks at 1680 cm^{-1} and 1342 cm^{-1}. Indeed, these absorption bands are found to have about the same weight as the one at 788 cm^{-1}, which is readily distinguished and identified in the raw spectra (see e.g. Figure 10). Guided by Table 5, Appendix 1 we see that the peak at 1342 cm^{-1} corresponds to CH deformations, but also contain contributions from adenine and guanine in DNA and RNA. This peak is not easy to distinguish in the raw spectra. It can however readily be revealed in a PLS-DA model.

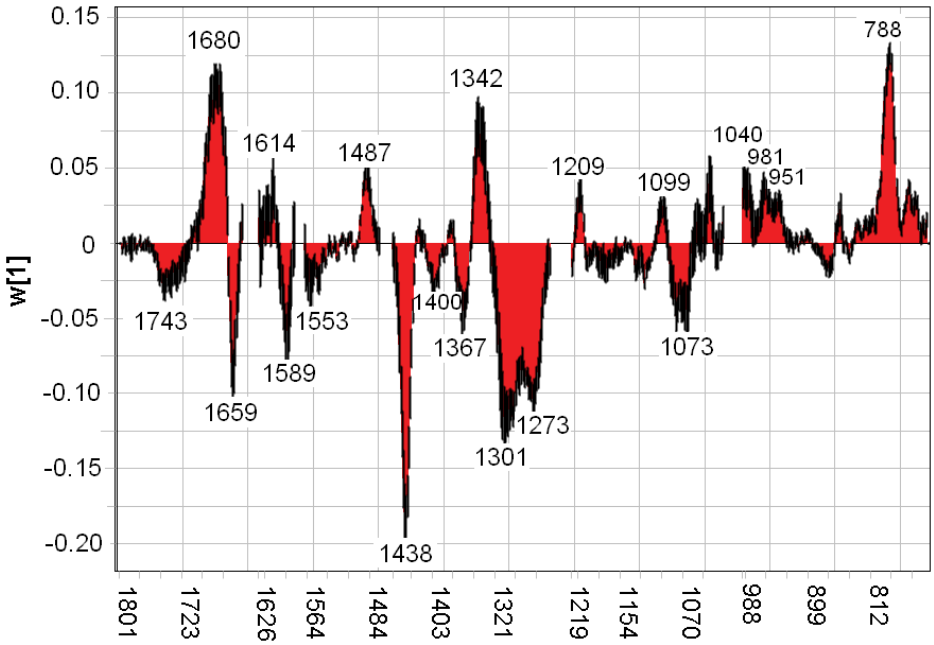

Figure 19 Weight plot (W[1]) for PLS-DA model, nucleus/cytoplasm. Positive values are correlated to observations from nucleus and negative values are correlated to observations from cytoplasm. 95% confidence interval is marked in black.

Other important absorption bands associated with the nucleus are located at 1614 cm^{-1}, 1487 cm^{-1}, 1209 cm^{-1}, 1099 cm^{-1}, 1040 cm^{-1}, 981 cm^{-1} and 951 cm^{-1}. The 1209 cm^{-1} peak originates from C-C_6H_5 stretching in phenylalanine and from tryptophan. Other peaks do not correspond to peaks assigned in Table 5, Appendix 1. The 1614 cm^{-1} peak is close to peaks due to tyrosine, tryptophan and phenylalanine. The 1099 cm^{-1} peak is close to peaks due to PO_2^- stretch (DNA/RNA) and C-C stretch in lipids and carbohydrates. The 1040 cm^{-1} peak is close to peaks due to C-C and C-O stretch in lipids and carbohydrates, respectively.

We find that the 1438 cm^{-1} peak is important to describe the cytoplasm. This spectral feature is a part of a region that contains signals from CH deformations. The region between 1248 cm^{-1} and 1329 cm^{-1} is also important to describe differences between cytoplasm and nucleus. This region contains mainly information from amides, CH deformations and =CH deformation in lipids. The 1320 cm^{-1} peak overlaps with DNA/RNA absorption due to guanine. Other important peaks are located at 1743 cm^{-1}, 1659 cm^{-1} and 1367 cm^{-1}, respectively, which mainly originate from lipids. Peaks from lipids can be expected to be important since some of the measurements from cytoplasm may contain information from the cell membrane. The model also shows that the peaks at 1589 cm^{-1}, 1400 cm^{-1} and 1073 cm^{-1} also describe important spectral features associated with the cytoplasm.

Based on this PLS-DA model, all measurements were either classified as belonging to the cytoplasm or the nucleus region. Of these 310 observations, 10 were not classified because these observations had distance to model-values above the calculated critical distance to model value.

The model was evaluated by using the test set. However, since no independent calibrated test set exists, which quantifies the full spectral response, this evaluation can only be made on the assumption that the nucleus and the cytoplasm, respectively, may be determined solely from analysis of optical images and from selected specific spectral features, which

here is the 783 cm^{-1} peak. Figure 20 shows a comparison of the observations classified from the 783 cm^{-1} peak and the observations classified by the PLS-DA model

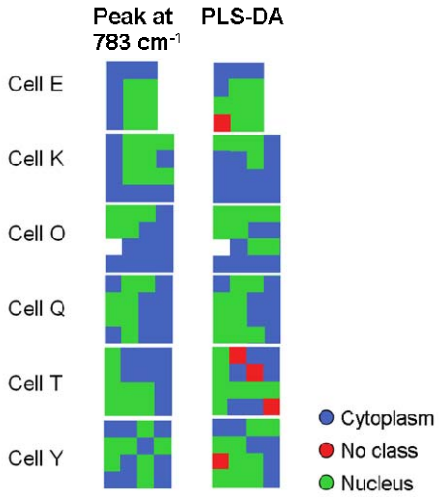

Figure 20 Comparison of the classification based on the 783 cm-1 peak and the classification made by PLS-DA. Cell E – control cell. Cell K – 24 h exposure to TiO2. Cell O – 48 h exposure to TiO2. Cell Q – control cell. Cell T – 24 h exposure to goethite. Cell Y – 48 h exposure to goethite.

The classification of Cell E, Cell O, Cell Q and Cell T is quite similar for the two classification methods, even though there are some deviations. The classification of Cell K and Cell Y differ however significantly. Figure 21 shows OM images of Cell K and Cell Y. The nucleus is difficult to distinguish, but the dashed green line indicates where shadows from the nucleus membrane may be seen. Only four measurements in Cell K are classified as measurements from the nucleus by the PLS-DA model, but these seems to correspond to the cell nucleus in OM image and to be more reliable than the classification based only on the 783 cm^{-1}-peak. Since the nucleus is difficult to distinguish in the OM image for Cell Y, the classification by the PLS-DA model is hard to evaluate, but the observations in the lower left corner seems to belong to the nucleus and they are in that case correctly classified.

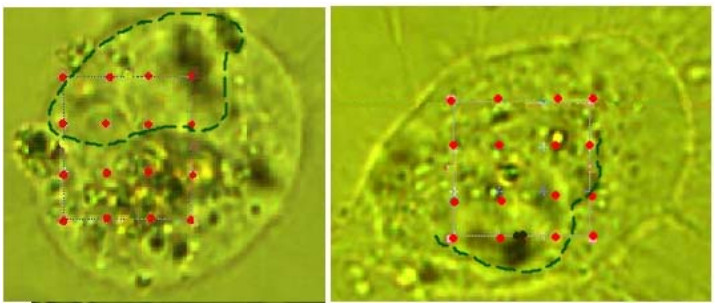

Figure 21 OM micrograph of Cell K, cell exposed to TiO2 (left), and OM micrograph of Cell Y, cell exposed to goethite (right). Spots measured with Raman microspectroscopy are marked by red dots. The position of the nucleus, inferred by inspecting the OM image, is marked by green dashed line.

4.4 Particle distribution

4.4.1 Titanium dioxide

The anatase TiO_2 nanoparticles appear mainly as large 400-700 nm agglomerates – much larger than their primary particle sizes – inside the cells after 24 h exposure in agreement with previous reports (Andersson et al.) and are therefore readily observed in OM. Raman mapping directly proves that the agglomerates inside the cells consist of anatase TiO_2. In Figure 36, Raman spectra in the 279-2030 cm^{-1} region, obtained from a cell exposed to TiO_2, are shown. The Raman vibrational frequencies and peak assignments are shown in Table 6, Appendix 1. The particle distribution in cells from the test set is depicted in Figure 23. The picture shows the relative intensity of the peak at 513 cm^{-1}. Intensities above 500 eps are considered as “high” and are colored in red/brown. Figure 23 shows that particles can be found in cytoplasm as well in nucleus, in varying amounts. Cell O has low content of particles.

4.4.2 Goethite

Goethite nanoparticles can also be seen in OM, see Figure 22. In Figure 36, Raman spectra in the 279-2030 cm^{-1} region, obtained from a cell exposed to goethite, are shown. The Raman vibrational frequencies and peak assignments are shown in Table 6, Appendix 1. Figure 23 shows the relative intensity of the peak at 478 cm^{-1} for the cells in the test set. Intensities above 1000 eps are considered as “high” and are colored in red/brown. Particles can be found in cytoplasm as well in nucleus, in varying amounts.

Figure 22 OM micrograph of Cell V: Cell exposed to goethite during 24 h.

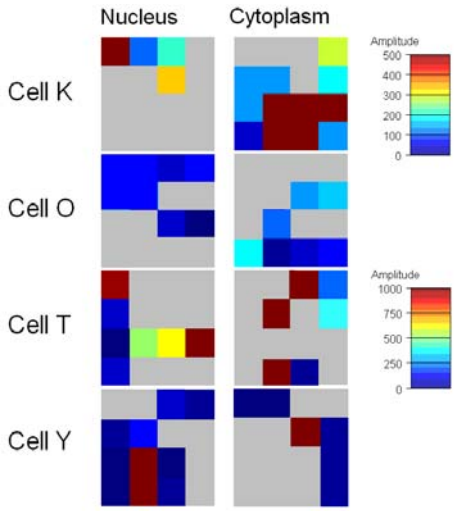

Figure 23 Intensity map for particle exposed cells in the test set. Cell K: 24 h exposure to TiO2. Cell O: 48 h exposure to TiO2. Cell T: 24 h exposure to goethite. Cell Y: 48 h exposure to goethite. High intensities (> 500 cps for the 513 cm-1-peak from TiO2 or > 1000 cps for the 478 cm-1-peak from goethite) are red/brown and other are colored according to relative intensities for the 513 cm-1-peak and the 478 cm-1-peak. Grey denotes area outside nucleus or cytoplasm, respectively.

4.5 Groupings and classes

PCA of raw data shows a trend, which at a first glance seems to be a separation between cells exposed to nanoparticles and control cells, but in fact seems to be correlated to the date when the data was collected. Figure 24 and Figure 25 show score plots which highlight this. This shows the importance of randomization of studies, to eliminate the effects from differences between different days and other differences that are not possible to hold constant. It also shows the importance of awareness of which relationship that is modelled and the importance of keeping track of all variables that may affect a study, not only the investigated variables.

After pre-treatment and collection of more data, groupings based on different days could not clearly be seen in a PCA model. A PLS-DA model, with the different days set as Y-variables, gave a model which could not predict which day the observations from the test set were measured, which also indicates that these differences are eliminated. No other clear groupings were seen in PCA models based on pre-treated data.

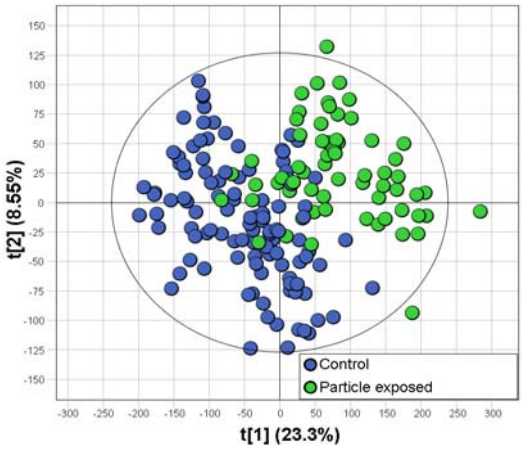

Figure 24 Score plot. PCA of raw data. Green – control cells. Blue – particle exposed cells.

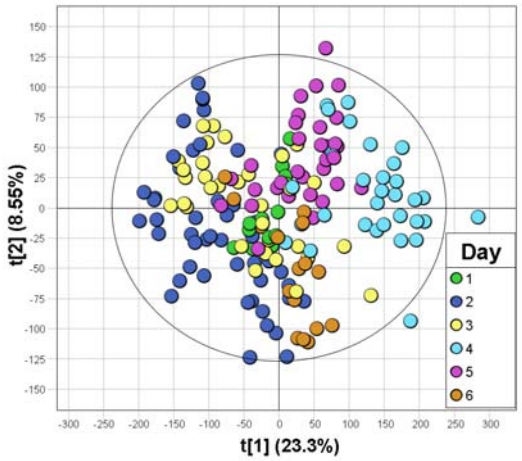

Figure 25 Score plot. PCA of raw data. Same score plot as in Figure 23, but with observations colored depending on day of measurement. Green – day 1. Blue – day 2. Yellow – day 3. Light blue – day 4. Purple – day 5. Brown – day 6.

4.6 Classification of particle exposed cells

PLS-DA models were constructed based on the whole data set as well as separate models based on the observations that the PLS-DA model had classified as either observations from nucleus or observations from cytoplasm. A PLS-DA model was also made based on control cells and observations with high particle concentration (intensity > 500 cps for the 513 cm^{-1}-peak from TiO_2 or intensity > 1000 cps for the 478 cm^{-1}-peak from goethite). A summary of the number of components and Q^2-values is in Table 1 and in Table 7, Appendix 1. The Q^2 varies between ~0.42-0.50, which can be considered as poor. Score plot for principal components 1, 2 and 3 showed a weak separation between classes in all four models. The model based on observations with high particle concentrations was slightly better than other (Figure 26). Large overlaps are evident but there are also some outliers among the observations from particle exposed cells in component 1 and from control cells in component 3.

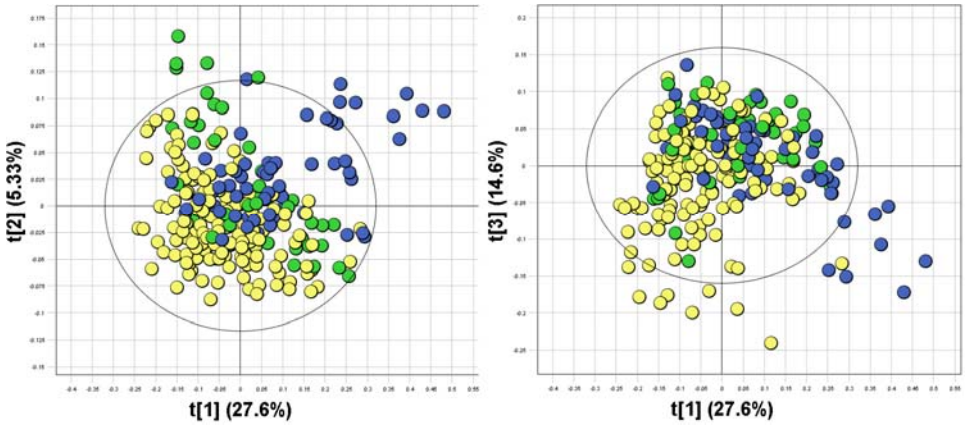

Figure 26 Score plots. PLS-DA model, control cells/particle exposed cells. The model is based on observations from control cells and observations with high particle concentration. Left: component 1 and 2. Right: component 1 and 3. Yellow – control cells. Blue – cells exposed to goethite. Green – cells exposed to TiO2.

Table 1 Summary of number of components and Q2 for PLS-DA models.

Model	No. components	Q^2
All observations	9	0.41579
Cytoplasm	8	0.46608
Nucleus	7	0.42062
High particle concentration	7	0.50489

The models were also evaluated by using the test set defined in Table 4, Appendix 1. Figure 27 shows a classification of the observations in the test set. The model based on all observations classifies most of the observations well except the observations from Cell O, in which more than 50% of the observations are false classified. Note that there is only a low concentration of particles in this cell (see Figure 23). Cell O can therefore be expected to show more similarities to control cells than particle exposed cells. The model based on observations with high particle concentrations show worse classification of Cell O and Cell Y. This model classifies more observations as control cells, which is expected. If there are no particles, or a very low concentration of particles, in the measured spot, the observation will probably be more similar to observations from control cells. The model based on observations from cytoplasm has on the whole a similar classification to the model based on all observations. The model based on observations from nucleus has many false classified observations, particularly in Cell E, Cell O and Cell Y.

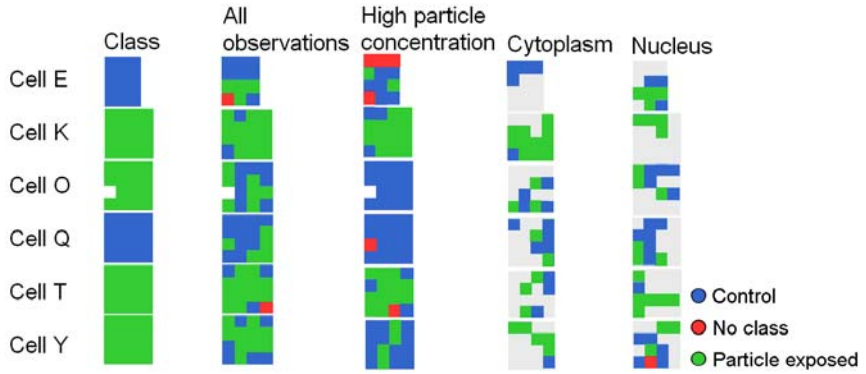

Figure 27 Comparison of the classification. Class denotes the correct classes. Cell E – control cell. Cell K – 24 h exposure to TiO2. Cell O – 48 h exposure to TiO2. Cell Q – control cell. Cell T – 24 h exposure to goethite. Cell Y – 48 h exposure to goethite.

It is however not to be expected that any model is able to correctly classify all measurements of cells, since a cell not necessarily show "particle exposed properties" throughout the whole cell. The observed particle distributions are spatially inhomogeneously distributed and are expected to chemically affect only the local environment significantly. Thus a cell cannot be expected to show "particle exposed" properties in the cell nucleus if particles are not detected there.

Figure 28 shows the weights for the model based on control cells and observations with high particle concentrations. The corresponding plot for all observations and observations from nucleus or cytoplasm are quite similar. The important spectral regions which describe the control cells are: (i) 1005 cm^{-1} (phenylalanine) and 1033 cm^{-1} (C-C stretch in lipids), (ii) the 1600-1640 cm^{-1} region which includes signal from C=C tyrosine and tryptophan, and (iii) the region between 730-820 cm^{-1}, which originates from tryptophan, uracil, cytosine, thymine and O-P-O stretch (DNA/RNA). The important peaks that describe particle exposed cells are peaks in the 1200-1350 cm^{-1} region. Here, peaks from amides, =CH deformations, CH deformations, phenylalanine, tryptophan, CH_2 twist, adenine and guanine can be found. The peak at 1578 cm^{-1}, from guanine and adenine, and the peak at 1013 cm^{-1}, from C-O deoxyribose (DNA/RNA), are also important.

The PLS-DA models show that particle exposed cells exhibit higher Raman intensity from some of the peaks that originate from DNA/RNA and also higher intensity in the region containing peaks due to amides. This may be an indication of an increased production of m-RNA. The 788 cm^{-1}-peak from O-P-O stretch (DNA/RNA) is however connected with control cells, but overlaps on the other hand with the 760 cm^{-1}-peak from tryptophan.

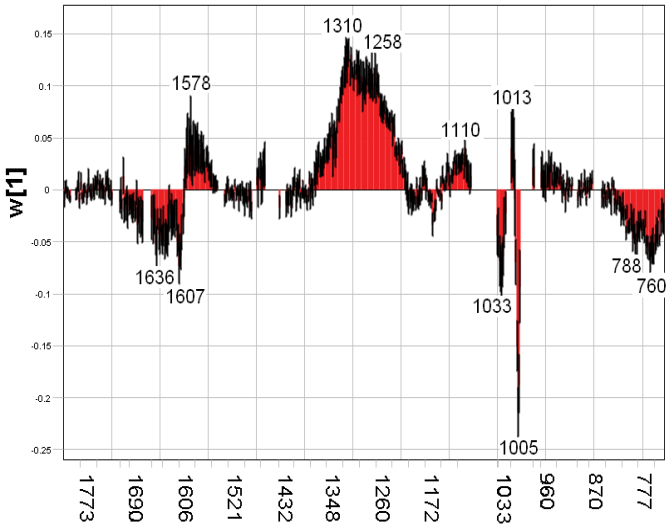

Figure 28 Weight plot (w[1]) for PLS-DA model, control cells/particle exposed cells. The PLS-DA model is based on control cells and observations with high particle concentration. Positive values are correlated to particle exposed observations and negative values are correlated to observations from control cells. 95% confidence interval is marked in black.

4.7 Effects of titanium dioxide

Based on previous studies, the biological effects of TiO_2 after 24 h exposure is expected to be small (Andersson et al.; Hext et al, 2005, p.471). Thus, spectral modifications in the 1000-1800 cm^{-1} region, where biological molecules are seen, are expected to be small. Here we constructed PLS-DA models based on all data as well as separate models based on the observations that the previously described PLS-DA model had classified as either belong to nucleus or cytoplasm. A PLS-DA model was also made based on control cells and observations with high particle concentration (intensity > 500 cps for the 513 cm^{-1}-peak from TiO_2). A summary of the number of components and Q^2-values is in Table 2 and in Table 7, Appendix 1. They show Q^2-values, which varies between ~0.40 and 0.62. The observations were not well separated in score plots, but the model based on observations from cytoplasm demonstrated the best separation and is shown in Figure 29.

Table 2 Summary of number of components and Q2 for PLS-DA models.

Model	No. components	Q^2
All observations	8	0.61773
Cytoplasm	5	0.51866
Nucleus	6	0.39523
High particle concentration	9	0.65803

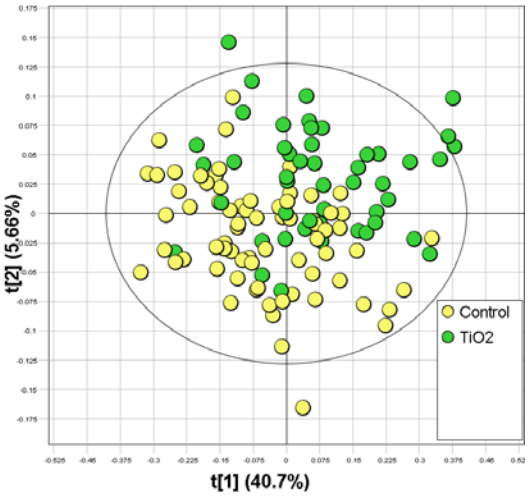

Figure 29 Score plot. PLS-DA, control cells/cells exposed to TiO2. The PLS-DA model is based on observations from cytoplasm. Green – cells exposed to TiO2. Yellow – control cells.

Classification by PLS-DA models are shown in Figure 30. Overall the classification is good, except for Cell O. As before, Cell O is generally classified as a control cell. Cell Q has in three out of four models some observations classified as particle exposed.

Figure 31 shows a weight plot for the PLS-DA model based on observations from the cytoplasm. The phenylalanine peak at 1005 cm^{-1} seems to be the most important spectral feature which describes control cells. Other important regions are 1600 cm^{-1} – 1650 cm^{-1}, which consists of signals from C=C tyrosine and C=C phenylalanine, tryptophan, tyrosine, guanine and adenine. The most important peaks that describe cells exposed to TiO_2 are those found between: (i) 1650– 1670 cm^{-1}, which originates from amides and C=C stretches in lipids, (ii) 1425 – 1480 cm^{-1}, which originates from CH deformations in proteins and lipids, and (iii) 1245 – 1315 cm^{-1}, which describe amides and =CH deformations and CH_2 twists in lipids. Thus, the particle exposed cells appear to have a higher concentration of proteins and/or higher protein activity. This can maybe be explained by production of small proteins, for example cytokines, due to the onset inflammation (Ekstrand-Hammarström et al.; Singh et al. 2007. p.149), and aggregation of protein on particles.

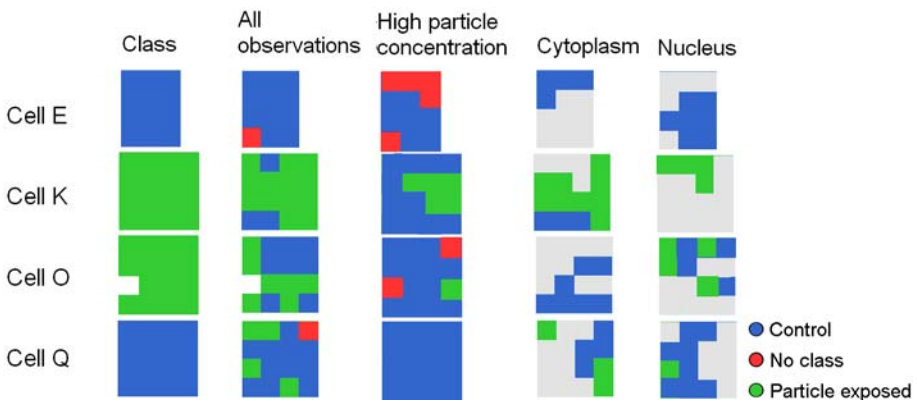

Figure 30 Comparison of the classification. Class denotes the correct classes. Cell E – control cell. Cell K – 24 h exposure to TiO2. Cell O – 48 h exposure to TiO2. Cell Q – control cell.

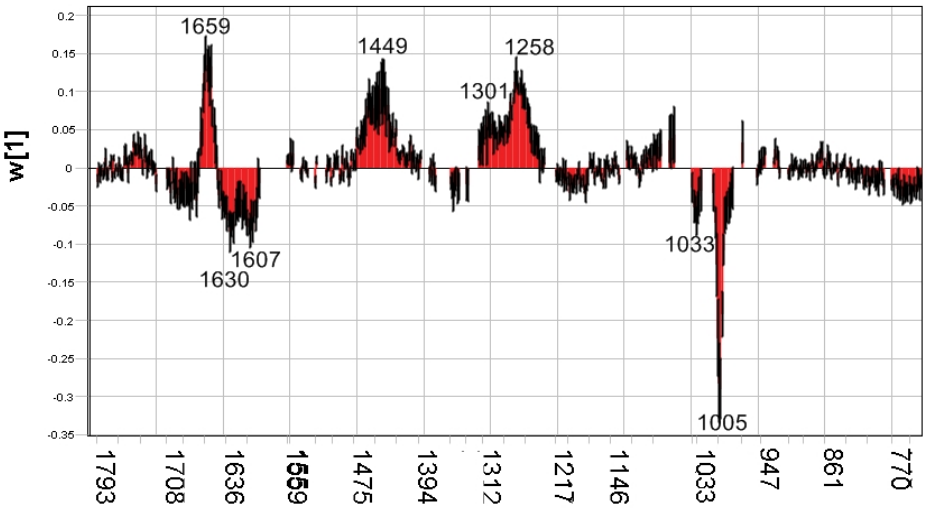

Figure 31 Weight plot (w[1]) for PLS-DA model, control cells/particle exposed cells. The PLS-DA model is based on observations from cytoplasm. Positive values are correlated to particle exposed observations and negative values are correlated to observations from control cells. 95% confidence interval is marked in black.

4.8 Effects of goethite

PLS-DA models were constructed from all data as well as separate models based on the observations that the previously described PLS-DA model had classified as either belonging to nucleus or cytoplasm. A PLS-DA model was also made based on control cells and observations with high particle concentration (intensity > 1000 cps for the 478 cm^{-1} peak due to goethite). A summary of the number of components and Q^2-values is shown in Table 3 and in Table 7, Appendix 1. The Q^2 varies between ~0.45 and 0.65, which is somewhat better than Q^2 for models that separates control cells from cells exposed to TiO_2. The separation in the score plot is significantly improved (Figure 32). There are however still overlap between particle exposed cells and control cells. The models based on observations from cytoplasm and selected observations with high particle concentration showed the best separation (Figure 32). In both models, there are more observations from control cells than observations from particle exposed cells. The groups in PLS-DA models should ideally be of same size and the fact that the control groups are larger may negatively affect the model.

Table 3 Summary of number of components and Q2 for PLS-DA models.

Model	No. components	Q^2
All observations	8	0.65394
Cytoplasm	5	0.55824
Nucleus	4	0.44869
High particle concentration	5	0.53013

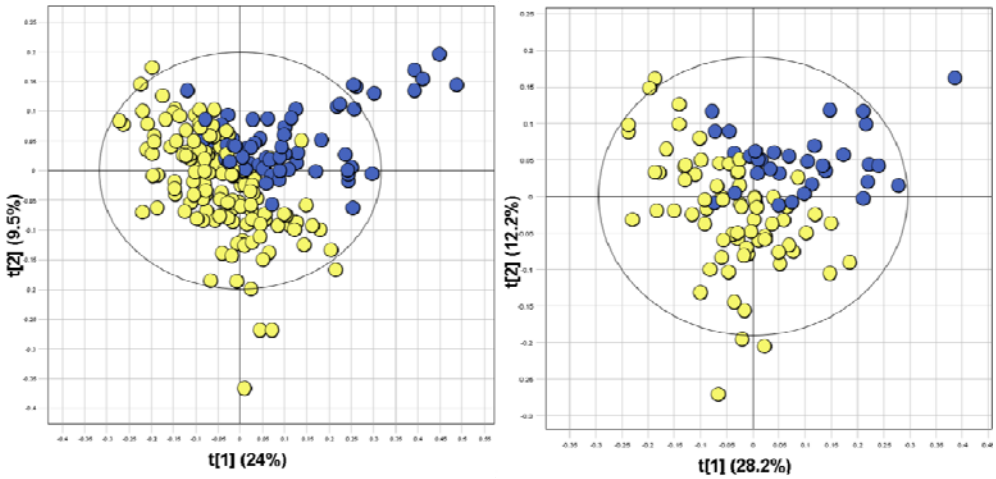

Figure 32 Score plots. PLS-DA, control cells/cells exposed to goethite. Control cells and observations with high particle concentration (left), observations from cytoplasm (right). Yellow – control cells. Blue – cells exposed to goethite.

The same score plots are also shown in Figure 33, where the observations are colored depending on exposure time. The observations from the cells that have been exposed to goethite for 48 h or more are in general not better separated from control cells than observations from the cells that have been exposed to goethite during a shorter time. A better separation could not be seen in other components either. Thus, the exposure time is not important to describe the explained differences due to goethite exposure.

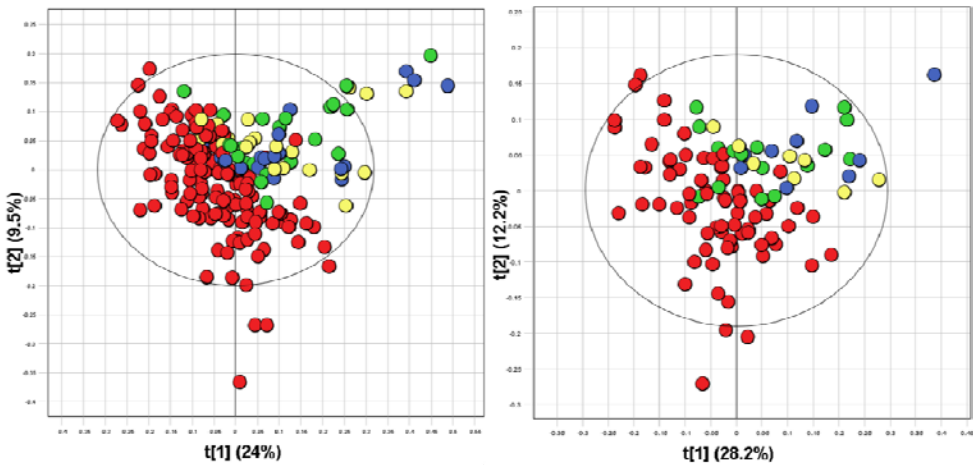

Figure 33 Score plots. PLS-DA, control cells/cells exposed to goethite. Control cells and observations with high particle concentration (left), observations from cytoplasm (right). Green – 24 h exposure time. Blue – 48 h exposure time. Yellow – 72 h exposure time. Red – control cells.

Figure 34 shows the classification of the test set. The two control cells, Cell E and Cell Q, are in general well classified, but with many wrongly classified observations by the model based on observations from nucleus. Cell T is also well classified in other models than the model based on observations from nucleus, which classifies only 50% of the observations as particle exposed. Most of the observations in Cell Y are classified as control cells, especially by the model based on observations from cytoplasm and the model based on observations with high particle concentration, which only classify the three observations with high particle concentration as particle exposed.

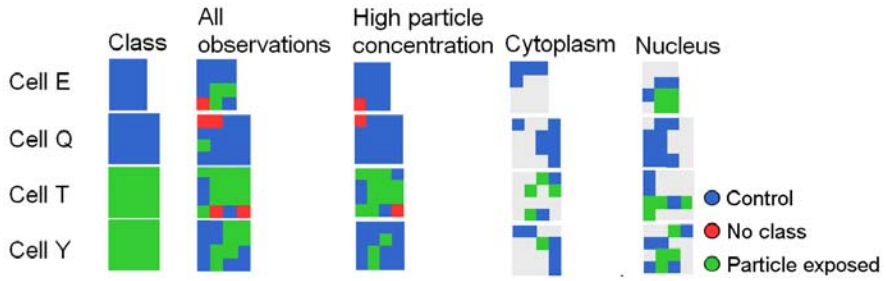

Figure 34 Comparison of the classification. Class denotes the correct classes. Cell E – control cell. Cell Q – control cell. Cell T – 24 h exposure to goethite. Cell Y – 48 h exposure to goethite.

The weight plot in Figure 35 shows the important spectral bands in the PLS-DA model based on observations from cytoplasm. The most important region, which describes the cells exposed to goethite, is 1200–1374 cm^{-1}. Here, peaks mainly due to amides and =CH deformations in lipids are present, but also CH deformations in proteins and signals from adenine and guanine. The latter assignment is supported by the high weight of the peak present at 1578 cm^{-1} which occurs in the same region where adenine and guanine exhibit strong absorption. We speculate that an increased protein and DNA/RNA activity may be explained by a production of m-RNA to synthesize cytokines due to ensuing inflammatory response. However, we do not have independent cytometric data that support this conclusion, but merely based on the much stronger separation in the model for goethite compared to the model for TiO_2, we may predict that goethite induces inflammatory responses in the lung epithelial cells. We note however that this spectral interpretation is complicated by small and overlapping peaks at 782 cm^{-1} and 788 cm^{-1}, which are also signals from DNA/RNA, but according to the model, are important to describe control cells. Since these peaks are small and the confidence intervals are so high, this correlation is uncertain, but we cannot rule out protein aggregation on the nanoparticles as a cause for the spectral modification.

Other peaks that are important to describe control cells are peaks between 1020 cm^{-1} and 1090 cm^{-1} (C-C stretching in lipids and C-N stretching in proteins), the peak at 760 cm^{-1}, which originates from tryptophan, a region around 1445 cm^{-1}, which originates from CH deformations, and peaks between 1600 cm^{-1} and 1670 cm^{-1}, which originates from C=C stretching from lipids, phenylalanine, tyrosine, tryptophan and signals from amides.

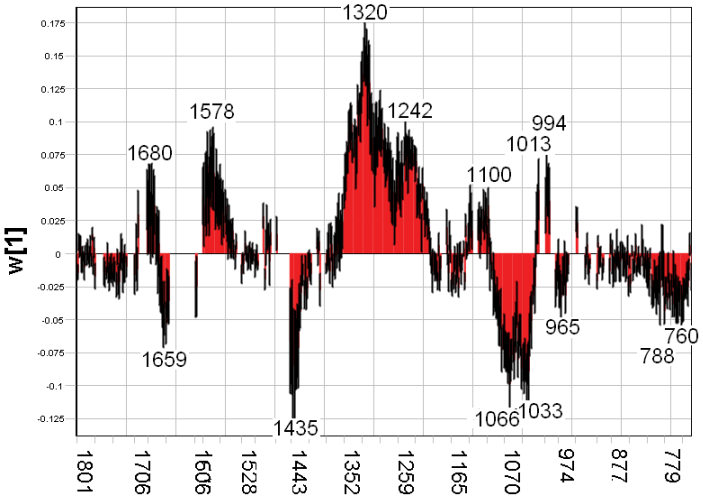

Figure 35 Weight plot (w[1]) for PLS-DA model, control cells/ cells exposed to goethite. The PLS-DA model is based on observations from cytoplasm. Positive values are correlated to particle exposed observations and negative values are correlated to observations from control cells. 95% confidence interval is marked in black.

5 Conclusions

Nanoparticle exposed cells and control cells are complex and inhomogeneous samples that are difficult to compare if only a few measurements are collected from each sample, in particular since measurements from different parts of cells have spectral differences (e.g. peaks from DNA/RNA in cell nucleus), which may be much larger than possible spectral differences caused by nanoparticles. In this study, hyperspectral images have been collected and whole cells or measurements from a certain location in cells have been analyzed and compared.

For Raman measurements, we used an Ar-ion laser, $\lambda = 514$ nm, the confocal hole was set to 150 μm and the total measurement time was set to 72 min/cell. Cells may be sensitive to the laser irradiation and cannot be measured during a too long time, but the cells were inspected after measurement and they appeared to be unaffected. The collected spectra were also found to have sufficient quality for hyperspectral data analysis. A longer measurement time, however, had been advantageous because we had been able to collect more measurements, e.g. measurements from three dimensions instead of a plane.

Before hyperspectral data analysis, it was considered necessary to pre-treat the data. Background correction is important because fluorescence gives a complicated background and makes a multivariate data analysis more difficult. Intensity variations, which originate from fluctuating laser power, were eliminated by vector normalization, which also can eliminate possible important differences, which originate from concentration differences in the sample. Internal standard calibration could have been a better option, but was not used here. Another pre-treatment method, smoothing, was tested. However, smoothing was found to worsen the analysis and spectra were hence not smoothed. An advantage with hyperspectral data analysis is that it is robust to noise.

PLS-DA was here used to identify the cell nucleus. A PLS-DA model was based on observations that showed evidence from both optical microscope images and Raman spectra (peak at 783 cm^{-1}) to belong to either the cytoplasm or nucleus region. The weight-plot obtained from this analysis shows correlation between the studied DNA/RNA-peak and other DNA/RNA-peaks, which are difficult to distinguish before hyperspectral data analysis.

PLS-DA models were also made to analyze differences between control cells and particle exposed cells. Separate models were made for the two different nanoparticles in the study. The separate models for TiO_2 and goethite are better than the model based on all data, because they show better separation between particle exposed cells and control cells in score plots and the Q^2-values are improved. The model based on control cells and cells exposed to goethite shows the best separation, which indicates that goethite spectrally affects the cells more than TiO_2 nanoparticles. The separation between control cells and particle exposed cells can also be improved if the observations from the nucleus are excluded. This can be explained by small differences between the cell nucleus in control cells and in particle exposed cells. About the same improvement is seen if the observations that contain high amount of particles are selected and other observations among the particle exposed cells are excluded.

A test set has been used for evaluation. One of the particle exposed cells in the test set contained only small amounts of particles. This cell was overall classified as a control cell by all PLS-DA models. Other cells were on the whole correctly classified. The models based on observations with a high particle concentration classify most of the observations as control cells, which is expected since observations from sample spots with low amount of particles may have more similarities to control cells. It is however not likely that a model can classify all observations correctly. The simple explanation to this is that a cell cannot be expected to show "particle exposed properties" throughout the whole cell.

Weight plots give information about variables that are important to describe the modelled differences. The plots differ somewhat between different PLS-DA models, but in general they show the same pattern. Important peaks to describe the particle exposed cells are peaks in the region from ca 1200 cm^{-1} to 1400 cm^{-1}, which originates from amides and lipids, but also DNA/RNA. In the PLS-DA model based on observations from control cells and cells exposed to goethite, as well as the PLS-DA model based on all data, the DNA/RNA-peak at 1578 cm^{-1} also seem to be important. This can maybe be explained by an increased production of m-RNA. However, this hypothesis is contradicted by peaks around 780 cm^{-1}, which originates from signals from DNA/RNA, but here is correlated to control cells. Thus we cannot rule out that an explanation of the particle-induced spectral modifications is due to protein aggregation. The peaks around 780 cm^{-1} are part of a region with overlapping peaks, which also hold information from proteins (tryptophan) and lipids. Other important regions to describe control cells are the phenylalanine peak at 1005 cm^{-1}, and a region between ca 1600 cm^{-1} and 1650 cm^{-1}, where peaks from lipids, amides and some amino acids can be found.

A final conclusion is that hyperspectral data analysis is very suitable for analysis of data from Raman mapping projects, such as the mapping of lung cells in this study. Small spectral differences, which are impossible to find by merely inspect a few spectra, can here be found by using multivariate projection methods.

6 References

Andersson, P.O; Lejon, C; Ekstrand-Hammarström, B; Akfur, C; Ahlinder, L; Bucht, A; Österlund, L. 2010. Polymorph and size dependent uptake and toxicity of TiO_2 nanoparticles in living lung epithelial cells. Submitted to Small.

Atkins, P.; de Paula, J. 2005. Elements of Physical Chemistry 4th ed. Oxford: Oxford University Press.

Baena, J.R.; Lendl, B. 2004. Raman Spectroscopy in chemical bioanalysis. Current Opinion in Chemical Biology, 8, pp.534-539.

Boiley, J.; Lützenkirchen, J.; Balmès, O.; Beattie, J.; Sjöberg, S. 2001. Modeling proton binding at the goethite (α-FeOOH)–water interface. Colloids and Surfaces A: Physicochemical and Engineering Aspects 179, pp.11–27.

Burger, J.E. 2006. Hyperspectral NIR Image Analysis: Data Exploration, Correction and Regression. Ph. D. Umeå: Swedish University 0f Agricultural Sciences.

Casey, A.; Herzog, E.; Lyng, F.M.; Byrne, H.J.; Chambers, G.; Davoren, M.; 2008. Single walled carbon nanotubes induce indirect cytotoxicity by medium depletion in A549 lung cells. Toxicology Letters 179, pp.78-84.

Cooper, J.B. 1999. Chemometric analysis of Raman spectroscopic data for process control applications. Chemometrics and Intelligent Laboratory Systems 46, pp.231-247.

Ekstrand-Hammarström, B et al., submitted to Toxicological Sciences.

Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikström, C.; Wold, S. 2006. Multi- and Megavariate Data Analysis: Part I: Basic Princples and Applications, 2nd edition. Umeå: Umetrics Academy.

Geladi, P.; Isaksson, H.; Lindqvist, L.; Wold, S.; Esbensen, K. 1989. Principal Component Analysis of Multivariate Images. Chemometrics and Intelligent Laboratory Systems 5, pp.209-220.

Gurr, J-R.; Wang, A.S.S.; Chen, C-H.; Jan, K-Y. 2005. Ultrafine titanium dioxide particles in the absence of photoactivation can induce oxidative damage to human bronchial epithelial cells. Toxicology, 213, pp.66-73.

Hext, P.M.; Tomenson, J.A.; Thompson, P. 2005. Titanium Dioxide: Inhalation Toxicology and Epidemiolgy. The Annals of Occupational Hygiene 49, pp.461-472.

Horiba JobinYvon. HR800 User Manual.

Horiba JobinYvon. Quality Control. 2007.

Kaiser Optical Systems Inc. 2009. Raman Products Technical Note 1350: Confocal Raman Microscopy.

Manceau, A.; Schlegel, M.L.; Musso, M.; Sole, V.A.; Gauthier, C.; Petit, E.; Trolard, F. 2000. Crystal chemistry of trace elements in natural and syntethic goethite. Geochimica and Cosmochimica Acta 64. pp.3643-3661.

McCreery, R.L, 2000. Raman Spectroscopy for Chemical Analysis. Columbus, Ohio: Wiley-Interscience.

Mäkie, P.; Westin, G.; Persson, P.; Österlund, L. Manuscript in preparation.

Nel, A.; Xia, T.; Mädler, L.; Li, N. 2006. Toxic Potential of Materials at the Nanolevel. Science, 311, pp.622-627.

Notingher, I.; Verrier, S.; Haque, S.; Polak, J.M.; Hench, L.L. 2002. Spectroscopic Study of Human Lung Epithelial Cells (A549) in Culture: Living Cells Versus Dead Cells. Biopolymers (Biospectroscopy) 72, pp. 230-240.

Oberdörster, G.; Ferin, J.; Gelein, R.; Soderholm, S.C.; Finkelstein, J. 1992. Role of the Alveolar Macrophage in Lung Injury: Studies with Ultrafine Particles. Environmental Health Perspectives 97, pp.193-199.

Pyrgiotakis, G.; Kundakcioglu O.E.; Finton K.; Pardolos P.M.; Powers K.; Moudgil P.M. 2009. Cell Death Discrimination with Raman Spectroscopy and Support Vector Machines. Annals of Biomedical Engineering, 37(7), pp.1464-1473.

Savitzky, A.; Golay, M.J.E. 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36(8), pp.1627-1639.

Schmid, U.; Rösch, P.; Krause, M.; Harz, M.; Popp, J.; Baumann, K. 2009. Gaussian mixture discriminant analysis for the single-cell differentiation of bacteria using micro-Raman spectroscopy. Chemometrics and Intelligent Laboratory Systems 96, pp. 159-171.

Singh, S.; Shi, T.; Duffin, R.; Albrecth, A.; van Berlo, D.; Höhr, D.; Fubini, B.; Martra, G.; Fengolio, I.; Borm, P.J.A.; Schins, R.P.F. 2007. Endocytosis, oxidative stress and IL-8 expression in human lung epithelial cells upon treatment with fine and ultrafine TiO_2: Role of the specific surface area and of methylation of the particles. Toxicology and Applied Pharmacology 222, pp.141-151.

Suh W.H.; Suslick, K.; Stucky, G.D.; Suh, Y-H. 2009. Nanotechnology, nanotoxicology and neuroscience. Progress in neurobiology 87, pp.133-170.

Swierenga, H.; de Weijer, A.P.; van Wijk, R.J.; Buydens, L.M.C. 1999. Strategy for constructing robust multivariate calibration models. Chemometrics and Intelligent Laboratory Systems 49, pp.1-17.

Vanderkooi. 2006. Vibrational spectroscopy.

Wiklund, S. 2007. Spectroscopic data and Multivariate Analysis: Tools to Study Genetic Perturbations in Poplar Trees. Ph. D. Umeå: Umeå University.

Wiklund, S.; Johansson, E.; Sjöström, L.; Mellerowicz, E.J.; Edlund, U.; Shockcor, J.P.; Gottfries, J.; Moritz, T.; Trygg, J. 2008. Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models. Analytical Chemistry 80, pp.115-122.

Witjes, H.; van den Brink, M.; Melssen, W.J.; Buydens, L.M.C. 2000. Automatic correction of peak shifts in Raman spectra before PLS regression. Chemometrics and Intelligent Laboratory Systems 52, pp.106-116.

Wold, S.; Sjöström, M.; Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. Chemometrics and intelligent laboratory systems 58, pp.109-130.

Zhang, L.; Henson, M.J.; Sekulic, S.S. 2005. Multivariate data analysis for Raman imaging of a model pharmaceutical tablet. Analytica Chimica Acta 545, pp.262-278.

Zhang, Z-M.; Chen, S.; Liang, Y-Z.; Liu, Z-X.; Zhang, Q-M.; Ding, L-X.; Ye, F.; Zhou, H. 2009. An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. Journal of Raman Spectroscopy, early view, published online 9 Oct 2009.

7 Appendix 1

Table 4 Summary of measured cells

Cell name	Particle (time of exposure)	Day measured	Test set/training set
Cell A	-	1	Training set
Cell B	-	2	Training set
Cell C	-	2	Training set
Cell D	-	2	Training set
Cell E	-	2	Test set
Cell F	-	3	Training set
Cell G	-	3	Training set
Cell H	TiO_2 (24 h)	4	Training set
Cell I	TiO_2 (24 h)	4	Training set
Cell J	TiO_2 (24 h)	5	Training set
Cell K	TiO_2 (24 h)	5	Test set
Cell L	TiO_2 (24 h)	5	Training set
Cell M	TiO_2 (48 h)	6	Training set
Cell N	-	6	Training set
Cell O	TiO_2 (48 h)	7	Test set
Cell P	TiO_2 (48 h)	7	Training set
Cell Q	-	7	Test set
Cell R	-	7	Training set
Cell S	-	8	Training set
Cell T	Goethite (24 h)	8	Test set
Cell U	Goethite (24 h)	8	Training set
Cell V	Goethite (24 h)	8	Training set
Cell W	Goethite (24 h)	8	Training set
Cell X	-	9	Training set
Cell Y	Goethite (48 h)	9	Test set
Cell Z	Goethite (48 h)	9	Training set
Cell AA	Goethite (48 h)	9	Training set
Cell AB	-	10	Training set
Cell AC	Goethite (72 h)	10	Training set
Cell AD	Goethite (72 h)	10	Training set

Table 5 Assignment of peaks in A549 cells (Notingher et al. 2002, p.233).

Peak (cm^{-1})	DNA/RNA	Proteins	Lipids	Carbohydrates
1743			>C=O ester	
1659		Amide I α helix	C=C stretch	
1617		C=C tyrosine Tryptophan		
1607		C=C phenylalanine Tyrosine		
1578	Guanine Adenine			
1460		CH deformation	CH deformation	CH deformation
1449		CH deformation	CH deformation	
1367			Symmetric stretch CH_3	
1342	Guanine Adenine	CH deformation		
1320	Guanine	CH deformation		
1301			CH_2 twist	
1284		Amide III α helix	=CH deformation	
1258		Amide III β sheet	=CH deformation	
1242		Amide III β sheet		
1231		Amide III random coils		
1209		C-C_6H_5 stretch phenylalanine, Tryptophan		
1176		C-H in-plane bending tyrosine		
1158		C-C/C-N stretch		
1128		C-N stretch		
1095	PO_2^- stretch		Chain C-C stretch	C-C stretch
1080		C-N stretch	Chain C-C stretch	C-O stretch
1066		C-N stretch	Chain C-C stretch	
1049			Chain C-C stretch	C-O stretch
1033		C-H in-plane phenylalanine		
1013	C-O deoxyribose			C-O stretch
1005		Symmetric ring breathing phenylalanine		
985			C-C head groups	
937		C-C backbone stretch A helix		
897	Backbone Deoxyribose			

853		Ring breathing tyrosine		
828	O-P-O stretch	Out of plane ring breathing tyrosine		
811			O-P-O	
788	O-P-O stretch			
782	Uracil Cytosine Thymine ring breathing			
760		Ring breathing tryptophan		
728	Adenine	Ring breathing tryptophan	C-N head group	
669	Thymine Guanine			

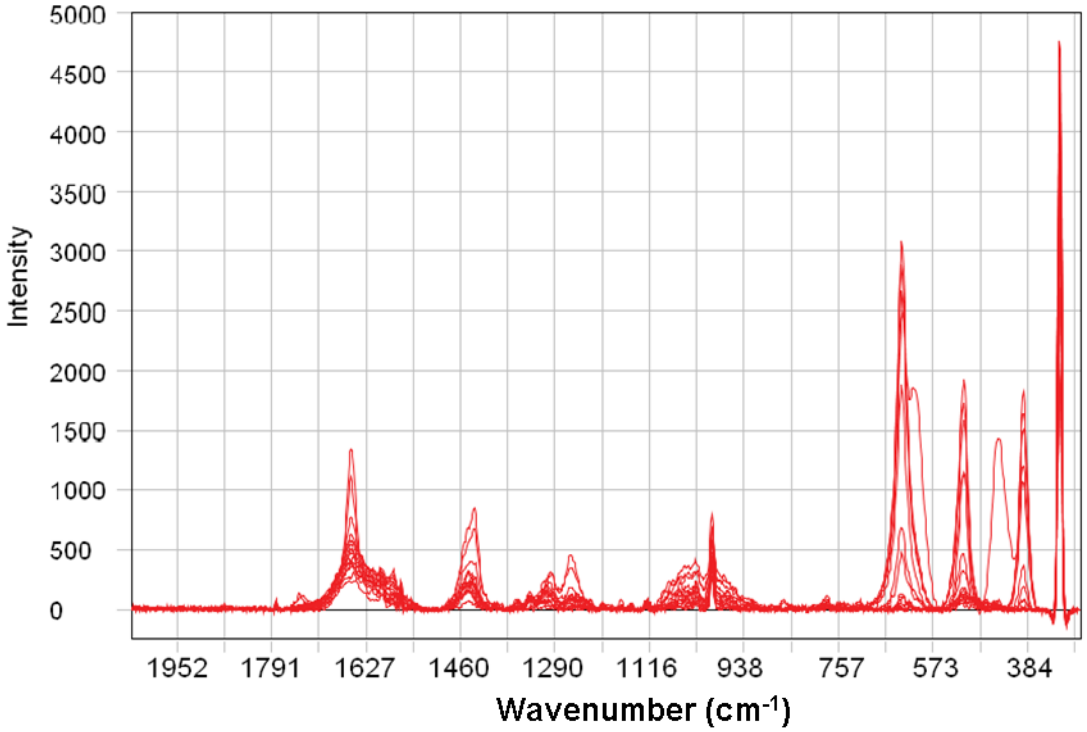

Figure 36 Superposed spectra from mapping of Cell H, A549 cell exposed to TiO2 during 24 h.

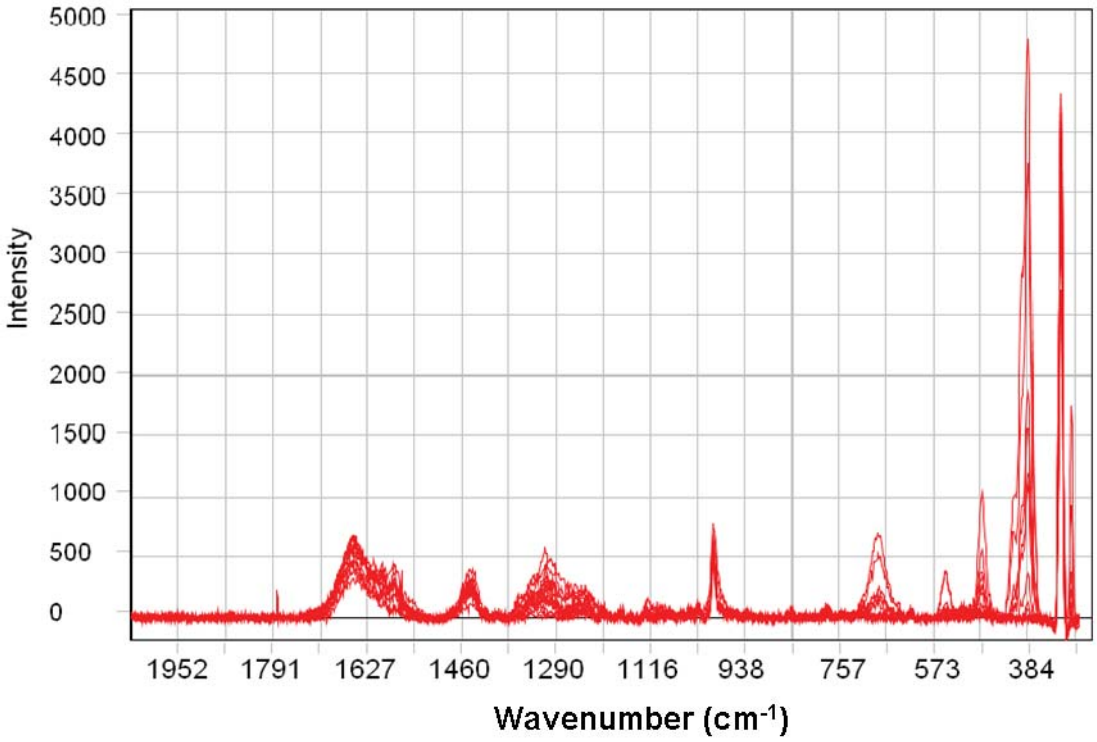

Figure 37 Superposed spectra from mapping of Cell W, A549 cell exposed to goethite during 24 h.

Table 6 Peak assignment for TiO2 and goethite in the measured spectral region

Peak (cm^{-1})	**Particle**
296	Goethite
388	Goethite
393	TiO_2
478	Goethite
442	TiO_2
513	TiO_2
552	Goethite
612	TiO_2
634	TiO_2
684	Goethite

Table 7 Summary of number of components and Q^2 for PLS-DA models.

Model	**No. components**	**Q^2**
Cytoplasm/nucleus		
	4	0.57640
Particle exposed/control, all data		
All observations	9	0.41579
Cytoplasm	8	0.46608
Nucleus	7	0.42062
High particle concentration	7	0.50489
TiO_2/control		
All observations	8	0.61773
Cytoplasm	5	0.51866
Nucleus	6	0.39523
High particle concentration	9	0.65803
Goethite/control		
All observations	8	0.65394
Cytoplasm	5	0.55824
Nucleus	4	0.44869
High particle concentration	5	0.53013

8 Appendix 2

R-code for baselineWavelet background correction (Zhang, Z-M et al. 2009).

```
library(baselineWavelet)
library(MASS)

# Reads row 2-17 from in.txt. in.txt contains a 17*1007 matrix, where
the first row holds the wavenumbers and each row 2-17 corresponds to a
Raman measurement. Column 1-2 contains spatial information([1,1] and
[1,2] is empty).
Cell <- matrix(scan("in.txt",n=16*1007,skip=1),16,1007,byrow=TRUE)

# Creates a vector that contains 63 scales
scales <-seq(1, 63, 1)

# Loops over all rows
q=1
for(q in 1:16) {
            x=Cell[q,3:1007]

            # Performs continuous wavelet transform with Mexican hat
            wavelet. The return is a matrix that holds the CWT
            coefficients for each scale.
            wCoefs <- cwt(x, scales=scales, wavelet='mexh')

            # Finds the local maxima among the CWT coefficients
            localMax <- getLocalMaximumCWT(wCoefs)

            # Identifies ridges from the local maximum of the CWT
            coefficients. The local maximum corresponds to the peak
            center.
            ridgeList <- getRidge(localMax, gapTh=3, skip=2)

            # Identifies the peaks by using the ridgelist and the
            estimated signal-to-noise.
            majorPeakInfo = identifyMajorPeaks(x, ridgeList, wCoefs,
            SNR.Th=1,ridgeLength=5)

            # Estimates the peak width of the identified peaks by
            using continuous wavelet transform with Haar wavelet.
            peakWidth=widthEstimationCWT(x,majorPeakInfo)

            # Fits the background by using penalized least squares
            with binary masks. threshold: peak shape threshold.
            lambda: adjustable parameter; high values give smoother
            fitted background. differences: the order of the
            difference of penalties of Whittaker Smoother method.
            backgr =
            baselineCorrectionCWT(x,peakWidth,threshold=0.3,lambda=100
            ,differences=1)

            # Subtracts the background
            corrected=x-backgr

            # Replaces the original spectrum by the background
corrected spectrum
            correctedt <- t(corrected)
            Cell[q,3:1007] <- correctedt

}

# Constructs a matrix that contains the background corrected data and
writes the file out.txt
correctedmatrix <- matrix(nrow=16,ncol=1007,dimnames=NULL)
correctedmatrix[1:16,1:1007] <- Cell
write.matrix(correctedmatrix,file="out.txt",sep="\t"
```