GÖTEBORGS UNIVERSITET

INSTITUTIONEN FÖR MARINA VETENSKAPER

Practical Application of Machine Learning for the Analyses of complex biological Matrices and environmental Phenomena

Alexandra A. Walsh

Akademisk avhandling för filosofie doktorsexamen i Kemi, som med tillstånd från Naturvetenskapliga fakulteten kommer att offentligt försvaras fredag den 2 oktober 2020, kl 10:00 i Hörsalen, institutionen för marina vetenskaper, Carl Skottsbergsgatan 22 B, Göteborg. Avhandlingen kommer att försvaras på engelska.

Institutionen för marina vetenskaper
Naturvetenskapliga fakulteten
2020

Practical Application of Machine Learning for the Analyses of complex biological Matrices and environmental Phenomena

alexandra.walsh@chem.gu.se

ISBN 978-91-8009-022-3 (PRINT)
ISBN 978-91-8009-023-0 (PDF)
Tillgänglig via http://hdl.handle.net/2077/66075

Printed in Borås, Sweden 2020
Printed by Stema Specialtryck AB, Borås

SVANENMÄRKET
Trycksak
3041 0234

ABSTRACT

This thesis presents research aimed at forwarding an understanding of machine learning methods as a method of studying complex matrices and environmental phenomena. A number of machine learning methods in the form of linear projection algorithms and statistical experimental designs were applied for qualitative analysis of different matrices. The used linear projection algorithms included principal component analysis (PCA), partial least squares (PLS), orthogonal partial least squares (OPLS), and transposed orthogonal partial least squared (T-OPLS). Several different statistical designs of experiments (DoE) were also implemented, including face-centred composite design (CCF), simplex mixture design, and definitive screening (DS) design. The analysed matrices included mammalian cells, wood, and a protein mixture. In addition to biological matrices, this work also presents research aimed at forming a multivariate understanding of a specific environmental phenomenon, namely the biogenic production of volatile halogenated organic carbons. Through the above enquiries, several challenges that exist in machine learning were examined.

The application of several linear projection algorithms for the spectral interpretation of hyperspectral images of human blood cells and PC12 cell line from rats was investigated when applied for spectral interpretation close to the detection limit. The achieved results revealed the benefits and the shortcomings of T-OPLS under such conditions. A deepened understanding of the T-OPLS algorithm was achieved by examining a protein-buffer mixture. The thesis provides therefore the first extensive examination of this algorithm and its performance in the analysis of nonlinear, co-dependent data. Also, the research presented here provided an extensive report on how linear projection algorithms with or without DoE may contribute to qualitative interpretation of nonlinear spectroscopic data.

A simplex mixture design and PLS were used to successfully quantify polyethylene glycol (PEG) in waterlogged archaeological wood. This study contributed both to the field of wood conservation and to the understanding of the performance of the used machine learning methods. Lastly, the biogenic production of volatile halogenated organic compounds (VHOCs) was examined. The reported research in this thesis was the first of its kind to involve DoE in the field of biogenic VHOC production. The acquired results indicate that previously reported formation mechanisms of VHOC were dependent on several abiotic factors, making the connection between those factors and the formation of VHOCs more complicated than had been previously assumed. By examining the biogenic VHOC formation multivariatly for the first time thus contributed to a deeper understanding of the formation of VHOCs and also emphasized the need for multivariate approaches, in particularly DoE, in any future examinations.

Key words: surface enhanced Raman spectroscopy, doxorubicin, acute lymphatic leukaemia, waterlogged archaeological wood, volatile halogenated organic carbons, marine algae, design of experiments, orthogonal partial least squares, principal component analysis.

TABLE OF CONTENTS

CONTRIBUTION LIST

Part I

Chapter 1

The author has solely performed all of the planning and experimental work in this chapter, as well as all of analysis with machine learning methods, evaluation of results, and writing.

Chapter 2

This work was published in *Holzforschung*:

Henrik-Klemens, Å., Abrahamsson, K., Björdal, C., Walsh, A. (2019). An *in situ* Raman spectroscopic method for quantification of polyethylene glycol (PEG) in archaeological waterlogged wood.

This author contributed to Raman expertise, the evaluation of results, supervision of the work, and writing the article. The chapter contains additional results not published in the article.

Part II

Chapter 3

This work was published in *Journal of Chemometrics*:

Walsh, A., Josefson, M., Abrahamsson, K. (2020). Method development for *in situ* study of marine vanadium peroxidase based on SERS and chemometrics

The author performed all experimental work, data evaluation, and had the main responsibility for the work involving writing the article. The chapter contains additional results not published in the article.

Chapter 4

The author planned, created, and analysed the executed statistical design and contributed with supervision. The author also interpreted and evaluated the acquired results. The results reported in this chapter will later be submitted as an article to the journal *Marine Chemistry*.

OTHER CONTRIBUTIONS

Below is a list of scientific and popular scientific contributions by the author not included in this thesis:

- Co-author to the chapter:
    - Josefson, M., Walsh, A., Abrahamsson, K. (2015). Imaging and identification of marine algal bioactive compounds by surface enhanced Raman spectroscopy (SERS). In: Stengel, D. and Connan, S. (eds) Natural products from marine algae. Methods in Molecular Biology, vol 1308. Humana Press, New York, NY.

- Participated in the Arctic Ocean 2018 expedition to the geographic North Pole.

- Author of the popular science publication about Raman spectroscopy:
    - Walsh, A. and Abrahamsson, K. (2015). Robust och bred analysteknik. Kemivärlden Biotech: kemisk tidskrift. *3*, 22-23.

LIST OF ABBREVIATIONS

airPLS – asymmetric iterative reweighted penalized least squares

ALL – acute lymphatic leukemia

AML – acute myoblastic leukemia

ANN – artificial neural networks

AuNP – gold nanoparticle

AW – archaeological wood

BSA – body surface area

BThB – bromothymol blue

CCD – charges coupled device

CCF – face-centred composite design

C-VBPO – *C. officinalis* vanadium bromoperoxidase

DA – discriminant analysis

DFT – density functional theory

DMEM – Dulbecco's modified Eagle's medium

DMS – dimethyl sulphide

DMSP – dimethylsulfoniopropionate

DoE – design of experiments

DOP - dopamine

DOX – doxorubicin

DS – definitive screening design

ECD – electron capture detector

EDA – exploratory data analysis

$FADH_2$ – flavin adenine dinucleotide hydroquinone

FF – full factorial design

F-HG – flavin-dependent halogenase

GC – gas chromatography

HEPES – 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

HI-HPO – heme iron dependent haloperoxidases

HPO – cofactor-free haloperoxidase

IS – internal standard

IUPAC – International Union for Pure and Applied Chemistry

LOD – limit of detection

LSPR – localised surface plasmon resonance

MBN – 4-mercaptobenzonitrile

MBN – 4-mercapto-benzonitrile

MCR – multiple curve resolution

MIMS – membrane inlet mass spectrometry

MLR – multiple linear regression

MSC – multiplicative scatter correction

MWL – milled wood lignin

NADPH – nicotinamide adenine dinucleotide phosphate

NAS – net analyte signal

NET – neutrophil extracellular network

NI-HG – nonheme iron-dependent halogenase

NIR – near infrared spectroscopy

OPLS – orthogonal partial least squares

PAH – polycyclic aromatic hydrocarbon

PBS – phosphate buffer saline

PC – principal component

PCA – principal component analysis

PDL – poly-D-lysine

PEG – polyethylene glycol

PL – photoluminescence

PLL – poly-L-lysine

PLS – partial least squares

PT – purge and trap

RCF – rolling circle filter

RMSE – root mean square error

ROS – reactive oxygen species

RSD – relative standard deviation

RW – recent wood

SAM – self-assembled monolayer

SAM-S-HG – S-adenosyl-methionine dependent methyl halogenase

SEM – scanning electron microscopy

SERS – surface-enhanced Raman spectroscopy

SNR – signal-to-noise ratio

SNV – standard normal variance

TEM – transmission electron microscopy

ThB – thymol blue

TOF-SIMS – time-of-flight secondary ion mass spectroscopy

T-OPLS – transposed orthogonal partial least squares

TP – target projection

T-PLS – target partial least squares

TrB – trypan blue

UV – unit variance

V-BrPO – vanadium bromoperoxidase

V-ClPO – vanadium chloroperoxidase

VHOC – volatile halogenated organic carbons

V-HPO – vanadium dependent haloperoxidases

V-IPO – vanadium iodoperoxidase

GLOSSARY

Allelopathy– a phenomenon in which an organism produced biochemical compounds to influence the growth, survival, and reproduction of other organisms.

Anexic – the state of a culture where the species present is uncontaminated by any other organisms.

Apoplast – in plant cells, apoplast is the name of the cell walls combined with the waterfound in them. It is found outside the plasma membrane.

Apoptosis – "programmed" cell death, or cell "suicide".

Cytoplasm – all components within a cell which are enclosed by cell membrane. Cell nucleus is not included into the definition.

Cytosol – the liquid inside the cells and is a part of the cytoplasm.

Cytostatic drug – a drug that inhibints cell proliferation and growth.

Efflux pump – protein responsible for moving unwanted compounds out of cells. They are present in the cell membrane.

Endocytosis – engulfment of extracellular matter. The term encomapasses phagocytosis (uptake of solids) and pinocytosis (uptake of liquids).

Genotype – the organim's genetic traits and is one of the factors comprising a phenotype. Commonly refers to a specific characteristic, e.g. metabolism.

Genus (plural genera) - taxonomic rank used in biological classification of organisms. It is followed by classification into species.

Granulocytes - polymorphonuclear leukocytes, which are characterised by a multi-lobed nucleus. Granulocytes include neutrophils, basophils, and eosinophils.

Heteroscedascicity – unequal variability scatter. It is described by residuals.

Heterotrophic – an organism which cannot produce its energy resources.

Leukocytes – a collective denomination of white blood cells that include granulocytes and mononuclear leukocytes.

Meristoderm – a layer on the surface of brown algae.

Mesocosm – a field experimental setup where natural phenomena are examined under controlled conditions. It is a compromise between controlled laboratory experiments and field surveys.

Mononuclear leukocytes – white blood cells that contain a one-lobed, non-segmented nucleus. This class of leukocytes includes monocytes and lymphocytes.

Morphology – an organism's form and structure.

Motility – an organism's movement.

Neoplastic – description of abnormal tissue growth, i.e. a tumour.

Order – taxonomic rank used in biological classification of organisms. The classification by order is then followed by a classification by family, genus, and species.

Peripheral blood – the blood circulating through heart and blood vessles. Contains, apart from leukocytes, red blood cells, thrombocytes, and plasma.

Phagocytosis – engulfment of solid materials by cells, e.g. nanoparticles or bacteria.

Pharmacogenomics – the discipline studying how genes influence the respose to drugs.

Phenotype – the variation of observable characteristics of an organism, with reference to the organism as a whole or a specific trait. Is insfluenced both by heriditary traits, i.e. genotype, and also environmental factors.

Phyllym (plural phyla) – taxonomic rank used in biological classification of organisms. It is followed by class and then order (see definition of order above).

Plasma (blood) – the component of blood that carries blood cells.

Polymorphism – the occurance of different phenotypes in a population of species.

Strain – a genetic subtype in microbiological organisms.

Taxon (plural taxa) – a unit formed by a group of at least one population of an organism.

Thallus (plural thalli) – undifferentiated tissue of a multicellular non-moving organism, e.g. algae.

Turnover number – the number of enzymatic conversion of substrates per second at a single catalytic site

Viability – within this chapter the term refers to the survivability of cells.

Zygote – eukaryotic cell formed through fertilisation two gametes, e.g. a sperm and an egg.

1. INTRODUCTION

During the Faraday Discussions held in Edinburgh in 2019, Johan Trygg[1] stated:

> *The challenge is not in data collection but in maximising information in data and transforming data into information, knowledge and wisdom.*

With this quote, he emphasises that the challenges analysts face today have primarily to do with the extraction of relevant information from data. Although data acquisition, especially that of high-quality data, remains a challenge, it has nevertheless become easier to generate over the last decade, both with respect to quality and abundance. Traditionally, various univariate statistical approaches have been (and still are) the most basic tool used by analytical chemists to analyse and validate data. However, as the analytical methods become more sophisticated and generate more complex data, classical statistics comes short. The decomposition and interpretation of high-dimensional data has therefore created a need for tools that can accommodate an increase in complexity. Thus, machine learning methods provided some of the tools required for the interpretation of complex data.

Machine learning (which also encompasses multivariate data analysis and chemometrics) can be seen as the study of algorithmic and statistical model-based solutions that aim to classify the information within data into patterns or to predict behaviour in a system based on *a priori* information. Interest in the application of machine learning methods in chemistry arose from the realisation that traditional univariate statistics were inadequate to describe chemical systems, which often were multivariate[2]. This paradigm shift occurred in the late 1960s, resulting in the first analytical publication dedicated to pattern recognition[3]. Finally, Svante Wold coined the term 'chemometrics'[4-6][*] for these machine learning methods for extracting chemical information from complex data – a term with which the reader will perhaps be more familiar with.

Chemometric methods were introduced from several sources into analytical chemistry. The first historical development occurred in the early twentieth century when quantitative analysis and analytical figures of merit (i.e. accuracy, sensitivity, etc.) became integral parts of the analytical discipline. The second push for chemometrics came in the 1960s through 1970's, when a number of theoretical chemometric papers appeared[3, 8-12], some of which were dedicated to the determination of the number of components in spectroscopic data. A third influence came from the pioneers of applied statistics in the 1920s and 1930s, Pearson and Fisher[13], who inspired the modern way of thinking about multivariate analysis. For instance, Pearson[14], and later Fisher and McKenzie[15], were among the first to formulate the modern definition of what is today called principal component analysis (PCA)[16] – a data exploratory analysis method broadly applied in analytical chemistry and

---

[*] Chemometrics, although a part of the machine learning methods and statistics in analytical chemistry, is a much narrower definition[7]. For instance, chemometrics focus largely on multivariate computational methods. Despite that, this author will be using the terms machine learning and chemometrics interchangeably from this point.

other disciplines[2, 17].

Eventually, these numerous influences converged, and became chemometrics as an independent discipline in the 1980s[2], with the appearance of the first dedicated journals, such as the *Journal of Chemometrics*[7]. Last but not least, the advent of increased computation power since the 1960's[18] has been, and still is, expanding the scope of data which machine learning can handle. This, in turn has allowed chemometrics to expand beyond the quantitative analytical chemistry into other disciplines such as forensics[19], pharmaceutics[20], metabolomics and metabonomics[21-22], proteomics[23], and cultural heritage studies[24-25].

Another integral part of modern chemometrics is the design of experiments (DoE). As more overlaps between statistics and chemistry occurred, the idea of applying experimental design was promoted, originally focusing on process optimisation. The principles behind statistical DoE had already been established already in the 18th and 19th centuries[2]; however, the first appearance of formalised DoE in chemistry occurred during the early years of the Second World War[7], with first mentions in literature occurring after the war[26]. The rationale for introducing DoE into analytical chemistry was, again, based on the advantage of considering chemical systems as multivariate rather than as univariate. To be precise; the different factors in a chemical system interact, so, designing experiments where one factor was varied at a time could lead to erroneous optima. Another advantage lies in the reduction of resource consumption by extracting more information from fewer experimental runs[2, 27].

In this thesis, the machine learning has been applied to the analysis of biological matrices (see Chapters in Part I) and of an environmental phenomenon (see chapters in Part II). The qualitative aspect of machine learning presented here relates to data mining, data exploratory analysis, and to discriminatory analysis of multivariate data. In the case of quantitative enquiries, machine learning was applied as a means of multivariate calibration. However, multivariate calibration methods were also utilised in an exploratory capacity. Furthermore, DoE has been applied in several cases (see Chapters 2, 3, and 4). In the sections that follow, the reader will be made acquainted with the rationale behind the presented research as well as introduced to the theory related to the methods applied throughout this work.

1.1 ASSUMPTIONS BEHIND THE APPLICATION OF MACHINE LEARNING

As machine learning approaches have their roots in statistics, there are several ways of looking at how data analysis ought to be carried out. From a statistical point of view, many machine learning methods, as they are used in chemistry, are perceived not to hold up to scrutiny as the application of machine learning in analytical disciplines is of practical character[28]. It is therefore important to clarify which assumptions guided the work behind this thesis.

Statistical approaches in modern statistics can be roughly divided into three groups; classical, Bayesian, and exploratory data analysis (EDA). All approaches start in a similar way with postulating a scientific enquiry and all end in a conclusion. However, the intermediate steps differ[29].

The so-called classical data analysis approach first postulates a problem, then collects data, creates a model, analyses the model, and then draws conclusions. Put differently, the next step after data collection is the imposition of a model, followed by estimation and analysis.

Another approach is Bayesian data analysis. First, a problem is defined and data is collected. Then, a model is generated based on the collected data, followed by the application of a prior distribution. Prior distribution means the application of the analyst's own *a priori* knowledge to the data, thus the models are shaped into what they 'ought' to be. For example, the variables in the data can be weighted prior to modelling. Finally, analysis is performed and conclusions are drawn.

The last type of approach is called exploratory data analysis (EDA) and is the 'philosophy' used in this thesis. EDA starts with posing a scientific enquiry, followed by data collection, analysis, and the generation of models, from all of which conclusions are drawn. The most significant difference between Bayesian and classical approaches compared to EDA is therefore that the former make *a priori* assumptions, meaning that the conclusions drawn become dependent on the validity of the *a priori* assumption. Put differently, the data is not manipulated in any way before it is examined. Instead, the EDA approach is more direct allowing data to display its inherent structure, making it less objective[♠], yet more intuitive and suggestive for modelling data[29].

In this thesis, the guiding assumptions for analysis are those of EDA. Although the research presented here will use algorithms which are inherently classical (e.g. such as partial least squares, PLS), all of the modelling is guided by the data itself. This means, for instance, that the application of exploratory algorithms always comes before the application of algorithms which need *a priori* assumptions. In addition, multivariate calibration algorithms have been, for the most part, used as tools for qualitative analysis rather than for the prediction of unknowns.

1.2 THE ADVANTAGES OF MULTIVARIATE METHODS

In chemometrics, the data is often considered as multivariate instead of univariate, which brings about several advantages for an analyst. If the data input is complex and large, univariate approaches may give an oversimplified view of the system being studied, giving rise to false positives and false negatives. Further, univariate methods cannot detect important relationships and synergies between variables that may be hidden in the data. This is due to the fact that univariate approaches tend to treat variables as being independent of each other, ergo, co-dependency is ignored. In contrast, multivariate methods do allow for the isolation of correlating variables and also help identify which

---

[♠] To call upon Gaukroger[30], what sciences require from the notion of objectivity is not absolute verisimilitude of reality; rather, what is sought is the reliability of interpretation. In that sense, all three approaches provide tools for such trained 'objective' judgements[31].

variables contribute most to the variability in the data[32-33]. In addition, some machine learning methods allow for dimensionality reduction, i.e. the many variables in the data set can be reduced to a few new variables called latent variables, which carry the main information from all original variables[34]. This reduction in turn makes it easier to visualise large data sets and facilitates a deeper understanding of the experimental data. Another advantage lies in the noise reduction achieved by using more redundant measurements of the same phenomenon.

Further, machine learning encompasses techniques focusing on multivariate calibration. One of the objectives is to reduce the number of dependent variables that need to be measured[33], i.e. responses that can be predicted from independent variables. The ultimate goal of multivariate calibration is to model a relationship between a set of measured variables and the property one wishes to predict[33]. The most common quantitative example of a dependent variable used in chemistry is the concentration of the analyte in unknown samples. Lastly, in similarity to exploratory analyses, multivariate calibration models allow for detection of outliers based on both graphical means and on a set of statistical assumptions[35].

2. THESIS DISPOSITION

In the sections below, the reader will be made acquainted with the instrumental and machine learning methods used by this author.

Thereafter, this thesis is split into two parts. Two of the chapters contained in Part I of the thesis concern themselves with Raman spectroscopic analysis of, arguably, one the most complex matrices an analyst may work with - those of biological origin; human blood cells (Chapter 1), and waterlogged archaeological wood (Chapter 2). What unifies the research presented in Part I is the aim of resolving and clarifying a number of issues in the Raman spectroscopic analysis of complex matrices and mixtures, such as nonlinearity and non-selectivity. Chapter 2 applies DoE of different complexity and all chapters in Part I include the utilisation of multivariate projection algorithms.

In Part II, the methods applied were instead focused on examining of the biogenic production of volatile halogenated organic carbons (VHOCs) by temperate marine algae (Chapter 4). In addition, Chapter 3 examines the enzyme responsible for VHOC production with Raman spectroscopy, thereby creating a bridge between the challenges faced in Chapters 1 and 2. VHOC production has been studied extensively in earlier research, especially with gas chromatography, but the research field concerned with these compounds largely focused on univariate approaches. The chapters in Part II therefore attempt to capture the complex nature of VHOC production via the application of multivariate machine learning methods.

The thesis closes with a discussion of what has been achieved during the author's research, and how it paves way for the future societal and scientific contributions. At the beginning of this thesis, the reader may also find a list of abbreviations and a glossary, the latter containing vocabulary uncommon to the field of analytical chemistry.

3. METHODOLOGY OVERVIEW

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*
>
> R. A. Fisher[36]

In the sections that follow, the reader will be introduced into the chemometric methods used throughout this thesis. The analytical methods applied as well as the theory concerning specific analytes can be found in the introductory sections of Part I and Part II.

3.1 DESIGN OF EXPERIMENTS

A design can be the very first step towards the creation of reproducible and robust methods. A designed mathematical space in which experiments are positioned can therefore provide information on how sensitive various responses are to variation in experimental factors, as well as being of assistance in identifying factors that have the greatest influence on the studied system. In a design, all of the studied factors are varied simultaneously in such way as to capture as much variation as possible while reducing the number of samples[37]. This maximization of variation makes the design of experiments a suitable first step if one is to work with multivariate analysis at a later stage.

A prerequisite for a successful design is the analysts' familiarity with the subject matter, a process often overlooked in literature. The familiarisation process may include steps such as trial runs to evaluate if an experiment can be run. Another important part of the process is to be able to set the relevant ranges for the levels in a design, by, for example, using mathematical estimations or by establishing lower and upper limits of detection (LOD_{min} and LOD_{max}) for the selected instrument.

The aim of performing an experimental design can have several purposes. In the early stages of an investigation, it is possible to screen for important factors that affect the system and processes. In such designs, the focus lies on identifying (without *a priori* knowledge) key factors. Within the work the author of this thesis did, the majority of the implemented designs were of the screening type as such design usually requires fewer experimental runs to generate detailed information. An optimization design, on other hand, is often applied after a screening design, which helped to select the factors affecting one's system. Lastly, when optimization has been performed, robustness testing is applied to minimize variations that stem from sources beyond the control of the chemist, i.e. noise. In other words, robustness testing aims to minimize random errors.

3.1.1 FULL FACTORIAL DESIGN (FF)

The simplest screening designs are the so-called 2^k factorial design, where 2 represents the two levels assigned to each factor – high and low (+1, -1) – and *k* corresponds to the number of factors considered in the design. Thus if the design considers three factors, the resulting number of experimental runs will be eight (2^3). The factors can be qualitative or quantitative[27]. The full factorial 2k (FF) design allows for the computation of both the main effects (linear terms, X_i) and the two-factor interaction effects (cross-terms, X_iX_j) without any *a priori* assumptions[38], and in turn avoids confounding between factors (**Equation 1**). FF design is an orthogonal design method, which means that the scalar product of the columns of any two factors is zero[38].

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_{1x2}X_1X_2 + b_{1x3}X_1X_3 + b_{2x3}X_2X_3 \qquad \text{(Eq. 1)}$$

In the equation above, the response *Y* is acquired as the sum of one constant term (b_0), three linear terms (b_1X_1, b_2X_2, b_3X_3), three two-factor cross-terms ($b_{1x2}X_1X_2$, $b_{1x3}X_1X_3$, $b_{2x3}X_2X_3$), which leads to a 2^3 design. An example of a 2^3 design is illustrated graphically in **Figure 1**, which is a representation of **Equation 1**.

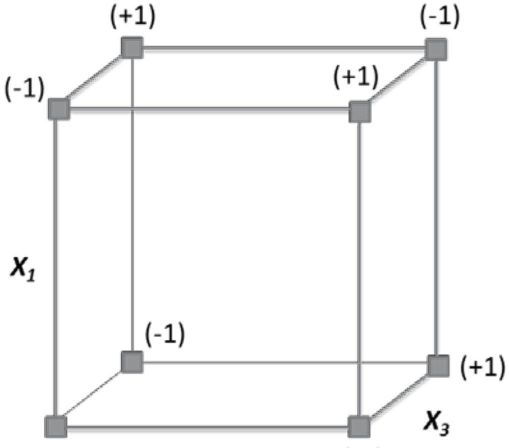


Figure 1. Graphical representation of a full factorial (FF) design space with two levels and three factors, i.e. a 2^3 design. The factors are X_1, X_2, and X_3. The (-1) and (+1) stand for the two levels of the design, where (-1) is the low level of a factor and (+1) is the high level of a factor.

However, if the relationship between the response and factors is not linear, the 2^k designs cannot adequately describe the behaviour of the studied system. For that purpose there are designs presented below which account for quadratic terms (X_i^2).

The issue with FF arises when the number of factors in the design space starts to increase. For instance, a 2^7 FF design will end up in 128 experiments, which may prove impractical. This factor redundancy can be amended with fractional FF designs. By using a fractional FF one assumes that not all factors and factor interactions are relevant because they are of negligible size. It is common, for instance, that linear terms have more impact than two- and –three factor interactions. Fractional FF designs reduce therefore the number of runs and consider primarily main factor effects and two-factor interaction effects. An example of a fractional FF is one-half fractional FF, 2^{k-1}. The disadvantage that may arise is confounding effects, for example the effects of three-factor interactions may become confounded with main linear effects[38]. An alternative way to minimize the number of experiments is to calculate quadratic design spaces (see examples in sections **3.1.2** and **3.1.4**).

3.1.2 Face-centred Composite Design (CCF)

Modelling nonlinear data produces curved design spaces. In such cases, designs based on nonlinear functions are given preference and include quadratic terms (X_i^2), in addition to the linear terms (X_i), and cross-terms (X_iX_j) present in linear designs. Quadratic terms allow for the identification of variables causing the design curvature, while cross-terms highlight synergy between different variables[27]. It could be said that with a design that allows for quadratic modelling, the nonlinearities are revealed directly, if they are present, due to the mathematical layout of the design.

In general, the central composite designs enable the estimation of the constant term, linear terms, cross-terms, and quadratic terms[27] between two or more factors (**Equation 2** represents a two-factor design built on X_1 and X_2)

$$Y = b_0 + b_1X_1 + b_2X_2 + b_1b_2X_1X_2 + b_{1x2}X_1X_2 + b_{1x1}X_1^2 + b_{2x2}X_2^2 \qquad \text{(Eq. 2)}$$

With standardized data (as in the software used in this thesis), the linear coefficient X_i describes the positive or negative slope at the origin, i.e. the middle of the standardized scales. The quadratic coefficients X_i^2 describe instead the depth of the valley or height of the peak. When the curve is pulled down to form a valley (apex at bottom) this is the convex state. The convex state has positive coefficients for quadratic terms. When the curve is pulled up to form a peak (apex at top) this is the concave state. The concave state has negative coefficients for quadratic terms. An x-coefficient around 0 with a positive X_i^2 component indicates a valley in the middle of the data space. A larger or smaller x-coefficient with a positive X_i^2 component indicates that the middle of the data space is on a slope down to the valley and it is likely that the bottom of the valley is not within the designed data space. When the data is plotted in the original, non-standardized coordinate system, the above discussion does not hold. When the original scale is desired instead it is recommended to plot a contour plot.

An example of this can be found in Chapter 3, where multivariate analysis performed on face-centred composite (CCF) designed space assisted in highlighting the nonlinear behaviour of the studied two-component system.

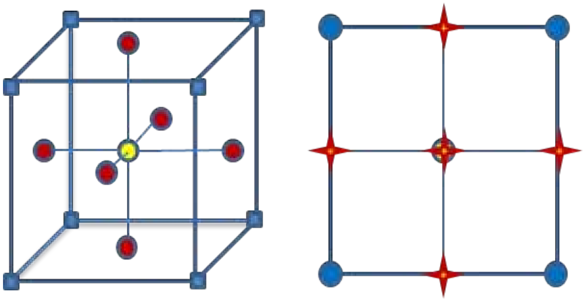


Figure 2. Graphical representation of a face-centred composite (CCF) design in three dimensions (left) and two dimensions (right). The CCF design space is composed of points corresponding to corners in a two-level factorial design (blue), symmetrical star points positioned on the factor axes (red), and a repeated centre point (yellow). Factor points are positioned in the corners of the factorial square, and star points are located on the face of each side of the factorial space so that the distance from the centre of the design space (α) equals 1.

A CCF design is created from two designs; a FF design and a star design. In addition to the minimum and maximum levels (2^k) assigned in a FF, the star design adds centre points and star points, which are the mean level of a factor. The values of the maximum and minimum levels are assigned as the distance from the star point. The FF backbone of this design reduces the number of necessary samples to achieve linear interpolation, which in turn allows for curvature estimation[38]. Each factor has therefore three levels assigned to it; high, low, and mean. Because of this, there are some experiments outside the original design space and some in the middle[33, 38], with centre points providing a built-in estimate of pure error and an estimation of the uniformity of precision[39].

3.1.3 SIMPLEX MIXTURE DESIGN

The designs presented so far were constructed using independent variables, i.e. each factor could be assigned a value range independently of other factors. Mixture designs, on other hand, contain proportions of different factors, rather than independent factors[27]. This means that all mixture factors must sum to 1 when expressed as fractions of the mixture. In this way, it is assumed that the measured response depends solely on the relative proportions, making the amount of the mixture irrelevant[29], and this makes it possible to study larger systems[37]. This means that the effect of the absolute quantities of factor variation is not in focus; instead, it is the effect of the factor ratios[27].

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + b_{1x2}X_1X_2 + b_{1x3}X_1X_3 + b_{2x3}X_2X_3 + b_{1x2x3}X_1X_2X_3 \qquad \text{(Eq. 3)}$$

Equation 3 illustrates that there is no constant and no quadratic factor, and is graphically represented by **Figure 3**. A mixture design is applied in Chapter 2.

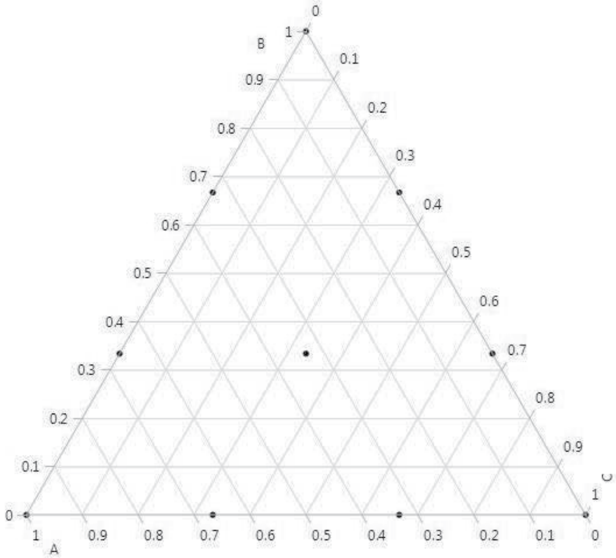


Figure 3. Graphical representation of a simplex mixture design for 3 factors (corners of the pyramid).

3.1.4 DEFINITIVE SCREENING (DS)

Definitive screening design (DS) is a screening design introduced by Jones and Nachtsheim[40]. It is distinguished from other screening designs in that it includes the estimation of quadratic effects, but with fewer experiments necessary. Most screening designs are 2^k designs, where the main effects could become confounded with cross-terms. DS is instead a three level design which presents some advantages, such as the ability to study quadratic effects while allowing for cross-terms. However, DS requires that all experimental runs are present for successful modelling; should there be a need to remove an outlier experiment or if that experiment fails, the DS design confounds the different terms.

3.2 Principal Component Analysis (PCA)

Principal component analysis (PCA)[16] is a projection method of exploratory analysis, which enables an overview of the data and projects the variation in data into fewer dimensions. More simply put, PCA is a multivariate analysis that focuses on simply 'looking' at the data and thus enables an easier visualisation of hidden structures[33].

The principle behind PCA was originally defined in 18th century by two separate mathematicians, Jordan and Beltrami[41], but the method fell into obscurity until it was mentioned again by Pearson in 1901[14], and Fisher and McKenzie in 1923[15]. PCA became visible again in the 1960's thanks to the work of Herman Wold[42]. During the 'boom' in chemometrics, PCA has become an established machine learning method in analytical chemistry, developed and advocated by, among others, Svante Wold[16], and Bruce Kowalski[43]. PCA is an eigenvector based multivariate analysis[♣] that makes it possible to see the intrinsic structure of the data in such way that it highlights the variation and co-variation in the data.

An input matrix X, also called an experimental matrix, is created that consists of variables (column vectors) and observations (row vectors). The observations are usually measured samples while the variables are the parameters describing the observations, such as pH, temperature, or *m/z* ratios[♦]. In other words, PCA, as well as other projection methods mentioned later on, starts with a two-way matrix (*K* x *N*), which gives rise to a first-order multivariate model[41]. PCA is an unsupervised method, which means that hidden structures can be found without the need of *a priori* assumption to execute calculation. As such, PCA is best suited for co-linear data to provide a broad picture of the data structure, but may miss small unique variation, especially if the sampling size is small.

PCA projects the data to extract one or several latent variables or principal components (PCs), which are linear combinations that give an approximation of the variation in the data. This extraction is achieved with the help of the column vectors contained in the X matrix. The first PC represents the maximum variation in the data and is established through a least squares fit (**Figure 4**).

[♣] The algorithms used in this thesis (PCA, OPLS, OPLS-DA, and T-OPLS) are all eigenvector based methods. Such methods also are called 'factorial methods', 'two-way modelling' and 'linear projection algorithms'. The denomination 'linear projection algorithms' will be used predominantly in this text.

[♦] The column vectors in this thesis are instrument responses, either chromatographic (retention times, min) or spectroscopic (wavenumbers, cm^{-1}). In the literature, several denominations of these column vectors can be found: 'variables', 'predictor variables', 'real variables', and 'manifest variables'.

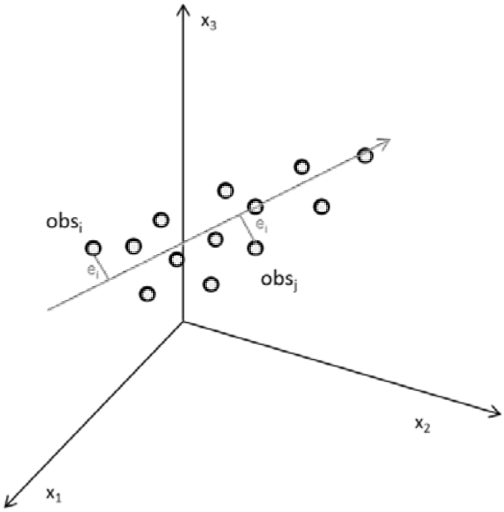


Figure 4. Fitting of a PC on raw data. Observations, or the row vectors, of the X matrix are plotted in a space created by three variables (x_1, x_2, and x_3), which are also called predictor variables. Obs$_i$ and obs$_j$ are observation in the data set, projected perpendicularly onto an arbitrary axis. The distance from the observation point to the axis represents the residual (e_i) for that observation.

In **Figure 4**, an arbitrary axis is drawn through the observation swarm. Each observation is the projected perpendicularly onto this axis. The perpendicular projected distance from the observation onto the line represents the residual e_i. Through a least square fit, it is possible to find an axis that gives the best fit to all points, i.e. minimising the sum of all squared e_i, $\Sigma(e_i)^2$. This line represents the first PC describing the maximum variance in the data swarm[33]. Therefore, all observations contribute to the determination of a PC though their squared projection distances. The second PC is orthogonal to the first PC and attempts to approximate the remaining variation, while the third PC is orthogonal to the second, and so on. As the calculation proceeds, PCs keep describing less and less of the remaining variation until they contain only noise[44].

The interpretation of the PCs is done through projections comprised of score (**t**) and loading (**p**) vectors. Each PC has corresponding **t** and **p** vectors, which together describe the variation. The scores are coordinates of observations in a *K*-dimensional coordinate space defined by PCs. The loadings on other hand are coefficients representing the linear combinations of variables for each PC[33]. The scores show how observations are located in relation to each other, while loadings illustrate the significance of each variable in relation to the directions in the score plot. The latter serves thus as a link between a PC and the variables[41]. By examining the loadings and scores in tandem, it becomes more feasible to represent many variables and their correlation[44].

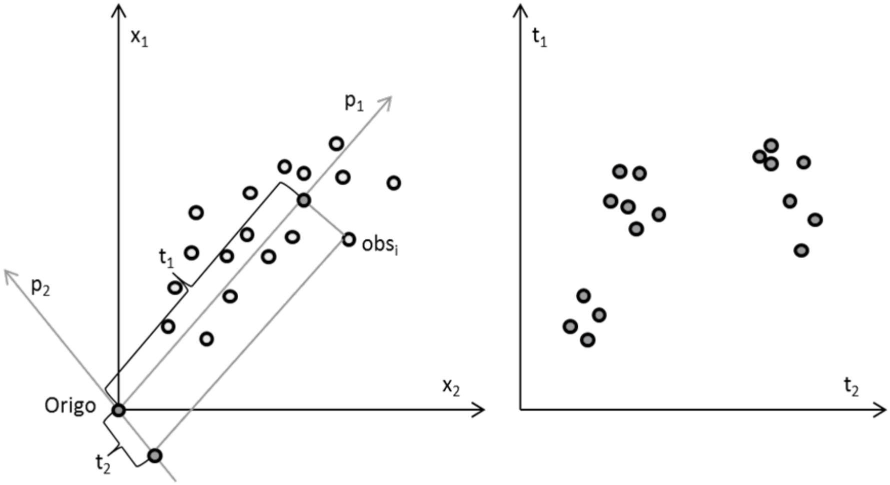


Figure 5. The projection of score (**t**) onto loading (**p**) vectors in a two dimensional space created from variables $\mathbf{x}_1$ and $\mathbf{x}_2$. (Left): Loading vectors $\mathbf{p}_1$ is drawn through the data swarm in the direction of maximum variance, then, a second loading vector $\mathbf{p}_2$ is drawn explaining the second-largest variation. Each observation (obs_i) in the data swarm is then projected onto the $\mathbf{p}_1$ and $\mathbf{p}_2$. The distance from projection to origin become an observation's new coordinates ($\mathbf{t}_1$, $\mathbf{t}_2$) and represent the score vectors. (Right): Through perpendicular projection onto loading vectors $\mathbf{p}_1$ and $\mathbf{p}_2$, each observation (obs_i) in the data swarm is assigned a new coordinates ($\mathbf{t}_1$, $\mathbf{t}_2$) which can be plotted in a new coordinate system called a score plot. The represented data swarm is not the same as in **Figure 4**.

The projection of **t** and **p** in a two dimensional coordinate space is illustrated in **Figure 5**. The first loading vector **p1** is drawn in the direction of maximum variation, and **p2** is drawn orthogonally to **p1**. Then, the score vectors **t** are projected perpendicularly onto the **p1** and **p2**. The resulting score plot shows the observations with new coordinates – the resulting groupings in the score plot reveal the structure of the data, where observation with similar properties governed by the **p** vectors are clustered together. The loading plot shows instead the weights by which each original variable should be multiplied in order to be transformed into the new coordinate system. The proximity of points in the loadings plot also reveals the level of correlation between the variables within the components selected for plotting.

The projection of matrix X into a new vector space, i.e. assigning **t** and **p** vectors to each observation and variable, involves a degree of information loss. Since the projection process is an approximation of the X matrix, large squared residuals e_i indicate a poor model fit, i.e. is not a good representation of the original data[33]. This residuals for each point can be calculated as $e_i = \mathbf{x}_i - \mathbf{t}_i\mathbf{p}_i$, where e_i is the residual, $\mathbf{x}_i$ is the corresponding column vector a.k.a. variable, and t_ip_i is the product of the score and the loading vector for $\mathbf{x}_i$. Thus, PCA decomposes the experimental data matrix X into a model part containing the structure of the data and a noise part[33], resulting in the following equation (**Equation 4**);

$$\mathrm{X} = \mathrm{TP}^T + \mathrm{E} = \vec{t}_1\vec{p}_1 + + \vec{t}_2\vec{p}_2 + \ldots + \vec{t}_i\vec{p}_j + E \qquad \text{(Eq. 4)}$$

Where X is the input experimental matrix, T is the matrix containing all the score ($\mathbf{t}_i$) vectors, P^T is the transposed matrix of all loading ($\mathbf{p}_j$) vectors, and E is the residual matrix containing the residuals for observations ($\mathbf{e}_i$) and variables ($\mathbf{e}_j$). The matrix E is in other words variation that did not fit well into the model matrix TP^T.

3.3 PARTIAL LEAST SQUARES (PLS)

A second common projection algorithm is partial least squares (PLS), which is a regression method based on the similar principles as PCA. Although PLS is primarily known as a tool for prediction of unknowns, it can also be used in interpretive manner similar to PCA.

Instead of one data matrix there are two matrices used, X and Y. The treatment of the two matrices can be conceptually seen as producing two separate PCA models that are connected through scores from both X and Y sets. One of the aims of PLS is to correlate the X and Y matrices with the purpose of prediction of the analyte concentration (or any other quantifiable parameter) in the X matrix. Moreover, PLS also aims to determine variables contributing most to the correlation between the two matrices. In this way it becomes possible to utilise PLS in order to calculate the concentration of the analyte from spectra of similar samples. New Y values are predicted from the X matrix; this estimation is called a prediction. Due to this, PLS and its derivatives are considered to be supervised methods.

$$\mathrm{X} = \mathrm{TP}^T + \mathrm{E} \qquad \text{(Eq. 5)}$$

$$\mathrm{Y} = \mathrm{UQ}^T + \mathrm{F} \qquad \text{(Eq. 6)}$$

Equations 5 and **6** show that the data decomposition for the X and Y matrices occurs in the same fashion. However, since, generally speaking, two PCA models are calculated. PLS generates two types of loadings; the **p** loadings projected from the X matrix and **q** loadings contained in matrix Q. The weight loadings, **w**, that constitute the connection between the X and the Y sides, are slope coefficients calculated from regression of each column in X to the score vectors of the Y matrix, **u**. In this way, columns, i.e. variables, in X that are strongly correlated to the **u** vector have large **w** values. A similar process is done for rows to acquire the **t** corresponding to the current **w** vector. In total, regular PLS provides five vectors – the vectors **t** and **p,** modelling the X matrix, and the vectors **u** and **q** modelling the Y matrix, together with **w** that describes the connection between the Y and X.

The difference is that PLS aims to correlate X and Y matrices to each other through scores from both X and Y for the purpose of prediction a response in the X matrix. Additionally, PLS aims to determine variables contributing most to the correlation between the two matrices. Thus, the aim of PLS is threefold:

1. Find the best description of trends in the X matrix (given by scores, **t**).
2. Find the best description of trends in the Y matrix (given by scores, **u**).
3. Find the largest covariance between X and Y matrices (correlation of scores **t** ***and*** **u**).

Covariance is the measure of how much the variation in two entities such as single variables or set of variables aggregated as vectors vary together.

If the aim of PLS is prediction, then the calibration is commonly divided into three stages; a calibration set, a test set, and a validation set. First, a calibration set, a.k.a. training set, is built consisting of standards measurements (which are as close to real samples as possible) measured in the same way that will be used for real samples, i.e. the calibration set is the X matrix. These known measurements have also their corresponding Y values measured[33]. A test set is represents future predictions against which the models will be tested[33]. In other words, a test set is a set of standards the concentrations of which are to be predicted by the calibration set. A validation set contains real samples[37]. For instance, a calibration set would be a set of standard measurements of all components of interest, while the validation set would contain the mixtures of components of interest.

3.4 ORTHOGONAL PARTIAL LEAST SQUARES (OPLS)

Orthogonal partial least squares (OPLS) is a modified version of PLS introduced by Trygg and Wold[45] that further simplifies the interpretation of correlation between X and Y matrices. It is designed to separate the variability in X, which is orthogonal to Y[46], from the variation that is correlated with Y. This gives two sets of scores and loading from the X matrix: one in which variation in X is correlated to Y and the other where the variation in X is not correlated to Y. The former give rise to predictive PCs with loading **p** and the latter to orthogonal PCs with loading **po**[47]. Predictions generated by OPLS models produce the same results as PLS models. With a lack of a statistical design of the calibration set, OPLS may suffer from bleed-over effect, where features that should be contained to the predictive loading **p** bleed-over to orthogonal loadings **po** and vice versa. This bleed-over effect has been illustrated in literature for near-infrared spectroscopy (NIR)[48], as well as in the research presented herein (Chapter 1 and Chapter 3). In both cases, the bleed-over was identified as a consequence of nonlinear response from instrumentation and properties of the samples. Other reason for the bleed-over can stem from poor experimental design or lack thereof.

3.5 OPLS COMBINED WITH DISCRIMINANT ANALYSIS (OPLS-DA)

Both PLS and OPLS can be used together with discriminant analysis (DA). In this mode, the model is used to explain differences between properties of classes selected in the dataset. For instance, the experimental matrix X contains measurements from different sampling sites. Then, the different sampling sites are set into minimum of two classes (1 and 2). The class designation becomes a Y matrix coded as one Y variable with ones and zeros marking representing membership in the respective classes. This Y is then correlated against the X matrix. A similar operation can be done by using PLS-DA, but since PLS lacks separation between predictive and orthogonal explanations of variation, the interpretation tends to become more complicated as the number of classes increases. OPLS therefore is usually more optimal choice to combine with DA. In addition, OPLS-DA is highly suitable for classification of data having a lot of noise and collinear variables.

3.6 TRANSPOSED ORTHOGONAL PARTIAL LEAST SQUARES (T-OPLS)

Transposed orthogonal partial least squares (T-OPLS) was used in this thesis as a tool in analysis of Raman spectroscopic data. T-OPLS has its origins in the work of Feudale and Brown[49], who have demonstrated a modification of PLS which they called target PLS (T-PLS) adding another specialized tool to spectral interpretation. This modification entailed inversion of the experimental matrix, i.e. all the spectra were transposed, and then applying a standard PLS algorithm to the experimental matrix. This transposition allowed for adding a spectrum of a sought substance as a Y matrix, thus enabling isolating variation based on a specific spectral pattern. The results of T-PLS differ from standard PLS; the loadings contain what was previously in scores and vice versa.

When it comes to spectral interpretation, T-PLS faces the same challenges as standard PLS; the contribution to scores comes both from correlation connected to the Y matrix and correlation that is not related to Y. To avoid this, T-PLS has been modified into T-OPLS by Abbas et al.[50] for analysis of hyperspectral images of algal carotenoids. In that paper, T-OPLS was denominated as transposed OPLS in contrast to Feudale and Brown's original target PLS. This was done to avoid confusion with target projection (TP) introduced by Kvalheim and Karstang[51], which is highly similar to OPLS and does not include any matrix transposition.

As OPLS, T-OPLS generates a set of predictive and orthogonal components both of which can be projected as score and loading vectors. The predictive scores (**t**) should resemble the spectral shape of the reference spectrum, the Y matrix, while the orthogonal scores (**to**) account for spectral features unrelated to Y. The predictive loadings (**p**) display the spectral variation with correlation to the known reference spectrum. The remaining variations in spectra which do not correlate to the Y matrix are explained by orthogonal loadings (**po**).

T-OPLS can thus be utilised to extract both desirable and undesirable patterns, where the latter can be seen as a more sophisticated approach to background subtraction. This subtraction is exemplified in Chapter 3, where T-OPLS is used to remove unwanted background variation from the matrix by averaging the blank signal and setting it as Y. In this way, the predictive component should be isolated as a variation based on only Y. All remaining variation in orthogonal components should then explain everything else that has no dependency on matrix signal. In Chapter 1, T-OPLS is instead utilized to compensate for the low reproducibility of SERS signal by attaching an internal standard (IS) to the enhancement surface and then setting the spectrum of the IS as Y. This, in theory, would allow for reproducible intracellular SERS quantification of analytes in the cell.

T-OPLS is a relatively new method in chemometrics and has not been developed and evaluated much outside the research group[50, 52-53] that originally proposed the algorithm, save for few applications for hyperspectral imaging[54-55]. In the scope of this thesis, it has been shown that T-OPLS tends to suffer bleed-over effects[48] between its predictive and orthogonal components. This shortcoming was corroborated for hyperspectral imaging[54], suggesting instead to use methods such as weighted OPLS and multivariate curve resolution-alternating least squares (MCR-ALS) to avoid the influence from uncorrelated variation. A part of modelling is the validation of models, to verify that the anticipated features can be retrieved from models through loadings or, in case of T-OPLS, scores. In case of T-OPLS the situation is trickier, since each prediction requires a building of a new T-OPLS model. Therefore, it becomes more challenging to employ a test set. However, the issue of cross-validation in T-OPLS is not addressed in this work.

3.7 MODEL VALIDATION

In order to determine the quality of a model and increase its reliability one must possess skill as well as the right tools of validation. Together with outlier measures such as Hotelling's T^2[56], cross-validation is fundamental to avoid random correlation.

A matrix with the dimensions 100 x 100 can, in theory, give rise to 100 PCs. However, not all components are relevant. One way to select an appropriate amount of informative PCs is cross-validation. Cross-validation is often used to determine the number of components that will be used in a model. It is achieved by dividing the matrix into groups, until each observation or group of observations has been left out from calculation once. When those missing values are predicted their sum becomes the first diagnostic of cross-validation – the sum of all prediction errors, Q^2. Q^2 is an evaluation of model's prediction capability or more simply the goodness of prediction (**Equation 7**).

$$Q^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{i,crosspred} - y_{i,reference}\right)^2}{\sum_{i=1}^{N}\left(y_{i,reference} - \bar{y}\right)^2} \qquad \text{(Eq. 7)}$$

Equation 7. The goodness of prediction, where N is the number of observations in the dataset, $y_{i,\ crosspred}$ is the predicted y for the sample, i from the cross-validation, and y is the mean for all y values in the calibration set.

Finally, the full matrix is used in a model that makes a prediction for all samples and those prediction errors are summed into R^2, which describes the goodness of fit (**Equation 8**) with the respect to the original dataset [44, 57].

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{i,calibpred} - y_{i,reference}\right)^2}{\sum_{i=1}^{N}\left(y_{i,reference} - \bar{y}\right)^2} \qquad \text{(Eq. 8)}$$

Equation 8. The goodness of prediction, where N is the number of observations in the dataset, $y_{i,\ calibpred}$ is the predicted y for sample i from the calibration of the whole calibration set, and y is the mean for all y values in the calibration set. R^2 can be calculated with respect to both X and Y matrices.

The Q^2 values are then used to select the number of components to include in the model. The R^2 value is increased when more components are added, but with more components noise will also be modelled thus decreasing the value of Q^2. Therefore the aim is to concentrate on Q^2 while not accepting an R^2 that is much higher than the Q^2 value[57].

Both R^2 and Q^2 have values between 0 and 1, with 1 representing a perfect fit and prediction, respectively. Q^2 values serve as a selection criterion for the number of components to include in the model. The R^2 value is increased when more components are added because more of the variation in the model is explained with each component, but with more components comes more noise and thus the value of Q^2 decreases.

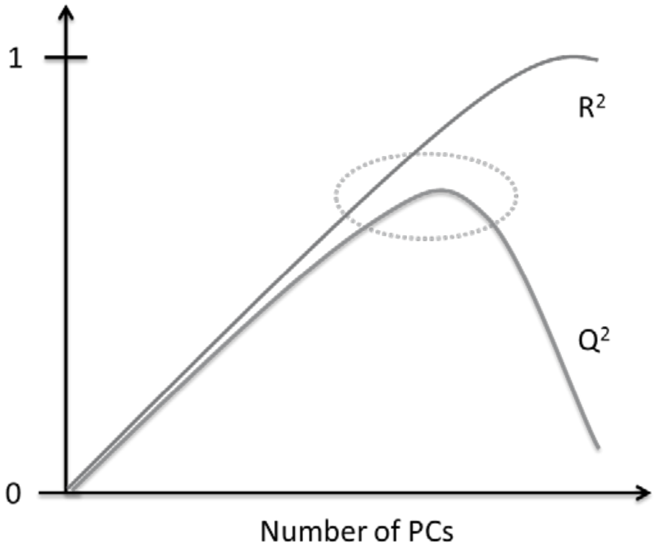


Figure 6. The image of R^2 vs Q^2. As the number of component increases, so does R^2, approaching to a value of 1. However, the prediction power of the model decreases as more noisy components are calculated: thus, Q^2 drops.

Validation is also helped by examination of score and loading vectors to detect outliers. Outliers are observations and variables with high leverage[32], i.e. have a strong influence on the model. Depending on what type of influence the outliers exert upon the model, they are classified as either strong or moderate. Strong outliers deviate from the maximum variation in the data swarm, while moderate outliers indicate observations that poorly fit the model, being thus connected to residuals. Strong outliers often have unique features that separate them from the rest of the observations or have the same features, but grossly exaggerated. It is said that strong outliers have high leverage, which can be illustrated by is Hotelling's T^2 [56], a multivariate generalisation of Student's t-test . Such outliers create a risk of model becoming biased; recognising and mitigating outliers by either complemented sampling or removal is thus essential in generating a reliable model[47, 57]. The outlier mitigation process may also give insights and unexpected information about the system under measurement.

In difference from strong outliers, which are a part of the model matrix (TP^T), moderate outliers can be identified from examining residuals of the E matrix. Many moderate outliers do not have the same profound effect on the models as strong outliers have, but their abundance can serve as an indication of lack of homogeneity in the X matrix. It is also common that strong outliers that break the trend of your data also have high residual values, but not vice versa.

3.8 SPECTRAL DATA PRE-PROCESSING

Spectral pre-processing is a crucial step in the analysis of Raman spectroscopic data. The aim in pre-processing spectral data before multivariate linear modelling is to reduce the variation between spectra that may occlude qualitative spectral analysis or to reduce the precision and accuracy of the regression models. For instance, pre-processing may assist in removing baseline effects caused by the sample matrix. In this thesis, baseline effects include the interfering florescence and background from biological matrices. If pre-processing is successful, the number of features that describe the spectral response would be reduced. In terms of linear projection methods, this could mean fewer amounts of components containing noise. This in turn leads to better quality regression models and interpretation of loading vectors. The choice of methods varies dependant on the dataset and/or the purpose of model calculation.

In Chapters 1 through 3, Raman spectroscopy was used in combination with a variety of linear projection methods. All data sets have been subjected to one or several data pre-processing strategies, which are explained below.

3.8.1 COSMIC RAY REMOVAL

Cosmic rays arise due to high-energy particles that pass through a Raman detector, which is interpreted as a signal. These signals are random and appear as sharp peaks, usually not following the expected Raman peak shapes (Gaussian, Lorentzian, Voigt, etc.)[58]. The cosmic rays can usually be easily amended by accumulating several spectra measured from the same sample/spot and then calculate their average. This was done in all research presented here where Raman spectroscopy was involved.

3.8.2 NORMALIZATION

Normalisation is a standard part of spectral pre-processing, as different spectra of a sample recorded at different times can be fluctuating in, for instance, intensity[58]. When the spectral profile is more important than the spectral magnitude, or when the spectral magnitude variations provide additional non-quantitative information that disturbs qualitative interpretation and predictions from multivariate regression models, it may be advantageous to normalise the spectra before multivariate modelling. A simple example is the measurement of a liquid placed in a cuvette with non-constant length. If all components in the sample contribute to the spectral profile, it is still possible to make an accurate calibration model and this model will get fewer components with normalisation.

Normalisation can involve dividing the individual spectra with its own sum or its own vector length. More elaborate schemes are also used, such as the standard normal variance (SNV) normalization. SNV normalization involves both a normalisation based on the standard deviation of the magnitudes in the individual spectrum combined with a subtraction of the spectral mean. Each observation (row) is centred and then divided by its standard deviation[32]. This helps with circumventing multiplicative and additive effects in spectra.

Another normalisation strategy is multiplicative scatter correction (MSC)[59], which also minimises the multiplicative and additive effects in spectra. The correction is done by calculating a mean of all observations of interest. The next step is to calculate linear regression slope and intercept for every observations and plot them against the observation mean. Each pair of values (slope and intercept) is then used for correcting each observation[32].

In this thesis, centring and auto-scaling[60-61] were used and are two of the simplest normalisation strategies available. By centring the dataset, mean of a centred variable is calculated and subtracted from all values of that variable, thus making it possible to analyse data based on covariance (**Equation 9**).

$$\tilde{x}_{ij} = x_{ij} - \bar{x}_i \qquad \text{(Eq. 9)}$$

In case of linear regression, centring may not be sufficient due to heteroscedasticity of the data, i.e. when the variability of a variable is not equal within the range that the variable used for prediction[62].

Auto-scaling is a combination of centring and unit variance scaling (variable is divided by its standard deviation, S_i). This way, the data is can be analysed based on correlation instead of covariance. The calculation is similar to SNV but is applied to columns in the dataset instead of rows (**Equation 10**).

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \qquad \text{(Eq. 10)}$$

All variables become equally important, which also means that weak signals and noise are amplified. This can lead to the unfortunate consequence of inflating measurement error and other undesirable signals[62]. For the sake of convenience, this operation is abbreviated 'UV' from this point onward.

Outlier removal is also a part of normalisation process and can be easily achieved by using dimensionality reduction with, for example, PCA.

3.8.3 DERIVATIVES

Derivatives have often been the first method tried to reduce sloping baselines in spectra. Most common are the use of first and second derivatives (**Figure 7**). The second derivative has the merit that it actually has a peak, albeit negative, at the original peak position while the first derivative has a crossing of the zero-magnitude level along with more confusion due to additional peaks compared to the original spectra (see below in **Combinations** section). It is also possible to use derivatives in order to reduce the influence of highly sloping baselines, such as when there is a large fluorescence component in Raman spectra. However within this thesis the occurrence of fluorescence was rare: the few observations that contained fluorescence were simply removed from calculations of the model.

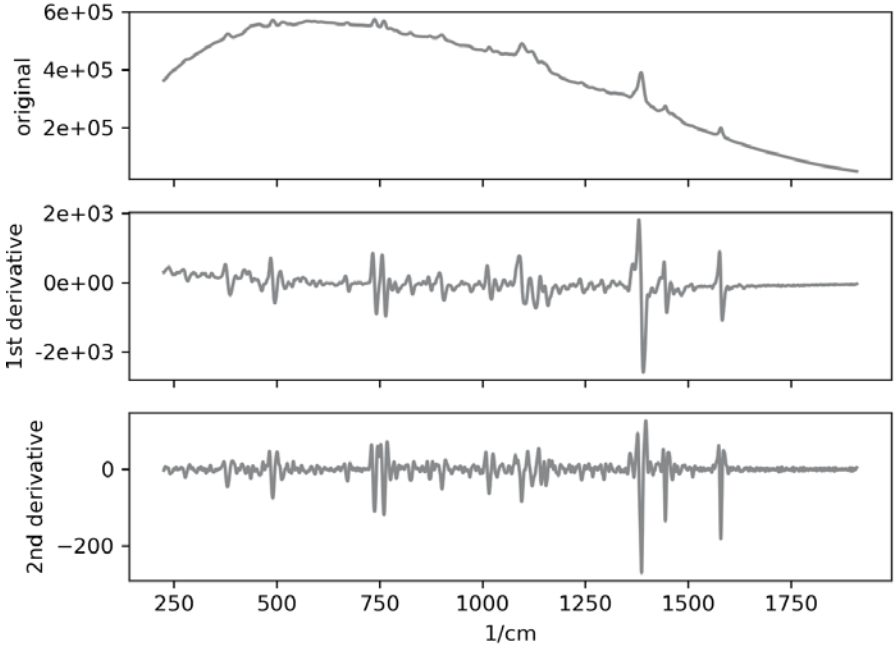


Figure 7. A regular Raman spectrum from a powder mixture of an organic analyte and microcrystalline cellulose causing the fluorescence background; 1st derivative of the same spectrum; 2nd derivative of the same spectrum.

3.8.4 BASELINE CORRECTION

While the derivatives affect the baseline in a general way, the following methods are algorithmically focused on direct adjustments of the baseline by explicitly providing an estimated baseline for subsequent subtraction.

3.8.4.1 Rolling Circle Filter

The rolling circle filter (RCF) is based on the difference the radii of curvature of the baseline and the Raman peaks[63-64]. The RCF is applied as the name suggests. A circle is rolled along the baseline from below the spectrum and the diameter of the circle determines how intrusive the circle will be into the bases of the peaks (**Figure 8**). Peaks that are narrow compared to the circle diameter will be retained but peaks that are wider than the circle diameter will be subtracted by from the baseline shape calculated from the rolling operation.

3.8.4.2 Asymmetric Least Squares Smoothing

Asymmetric least squares smoothing employs a series of smoothing operations on the spectrum that results in a calculated baseline shape that can be subtracted from the original spectrum (**Figure 8**). The degree of smoothing is governing the peak width of the retained peaks. With less smoothing the baseline will grow into the peak bases while with more, there will be a flat baseline below the peak. This effect can be achieved with ordinary least square method. However, in such a case, equal weight is given to positive and negative residual errors. As negative residuals are undesirable for spectral interpretation, the least square method can be altered to have a preference for positive residuals[58]. Different implementations of this type of baseline correction also goes under the names Whittaker filtering[65], and asymmetric iterative reweighted penalized least squares (airPLS)[66], both using the 'perfect smoother' by Eilers[67].

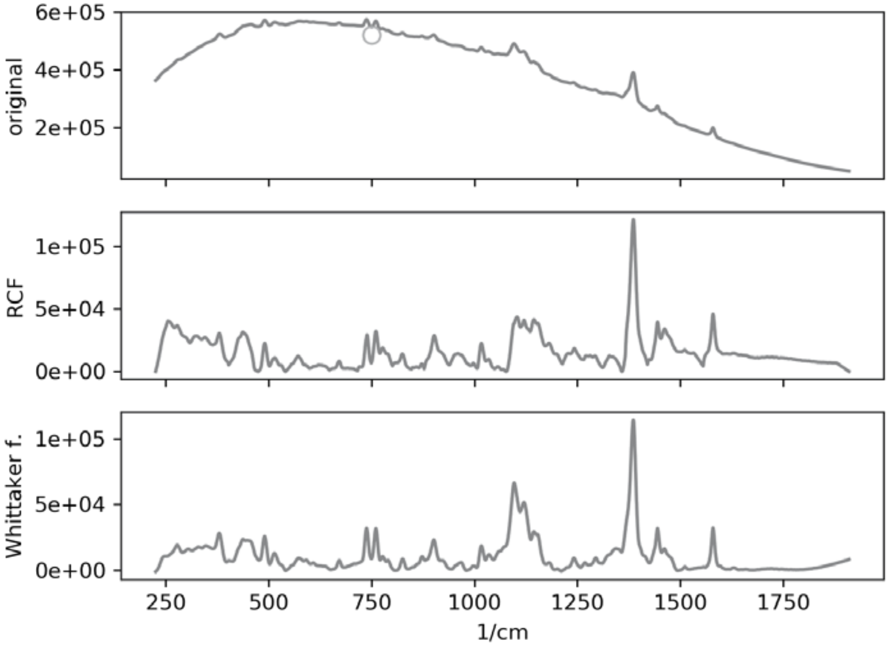


Figure 8. A regular Raman spectrum from a powder mixture of an organic analyte and microcrystalline cellulose causing the fluorescence background; the same spectrum after rolling circle filtering, a circle representing applied circle diameter is shown in the original spectrum plot; the same spectrum after application of asymmetric least squares smoothing (Whittaker).

3.8.5 PEAK FINDING

Another approach is to do regular peak finding (**Figure 9**) on the Raman spectrum in the same way as applied liquid chromatography but with the difference that the peaks are not integrated but preserved in the same frequency scale as the original data. This leaves the significant peaks and their original peak shapes intact in the spectrum while the spectral profile between these selected peaks is set to zero or down-weighted. Especially for T-OPLS, this is a way to simplify the interpretation of spectra.

Admittedly, this method is a step away from providing quantitative multivariate calibrations where it is usually a good idea to keep as close to the original variation in spectra and keeping all relevant variables in the calibration model.

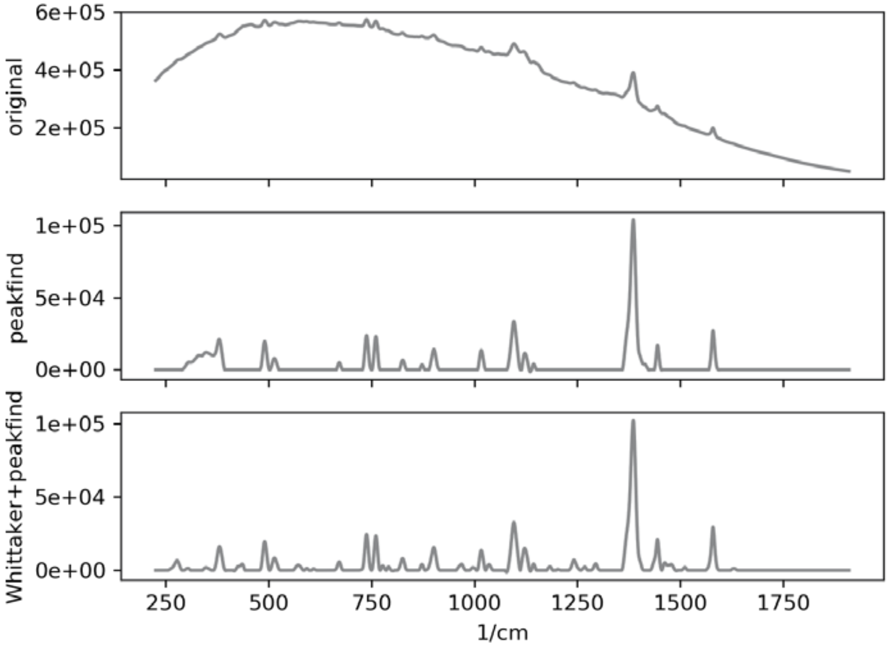


Figure 9. A regular Raman spectrum from a powder mixture of an organic analyte and microcrystalline cellulose causing the fluorescence background; the same spectrum after keeping only larger sharp peaks; the combined effect of asymmetric least squares smoothing followed by peak finding.

3.8.6 COMBINATIONS

Sometimes it is also useful to combine pre-processing methods. For example, the first-derivative calculation followed by SNV is a viable pre-treatment for Raman spectra of powders and tablets. The combination of RCF and peak picking is also useful, especially in the interpretation mode of chemometric modelling.

To reduce the noise that is enhanced by, for instance, derivatives compared to the magnitude of the analyte signal, derivative calculations are often combined with smoothing. One common approach is to use Savitsky-Golay filters for this purpose[58, 68]. If long term robustness of a multivariate calibration is important, derivatives are also increasing the sensitivity for wavelength or frequency shifts in the instrument used, which means that other pre-processing methods may be more favourable in this case.

4. CHALLENGES ADDRESSED IN THIS THESIS

Although application of machine learning methods is far from new within analytical chemistry (see Introduction), many of the challenges encountered at the birth of the discipline still await their resolution. In 1979, Wehry and Mamantov[69] stated that one of the most significant challenges in analytical chemistry at the time was the development of techniques for qualitative and quantitative analysis of trace organic compounds in complex samples. Although their definition was aimed at organic samples, the issue of separating the analyte signal from its matrix is, arguably, universal for many endeavours in analytical chemistry. Forty years later, the Faraday Discussions[1] had the topic "*Challenges in the analysis of complex matrices*", which signifies that the analysis of complex matrices remains to this day a central challenge in the analytical field. Faraday Discussions of 2019 concluded with recognizing that there seems to be little background knowledge on the samples tested, emphasizing the importance of sampling and experimental design.

In vibrational spectroscopy, there is an ongoing development of methodologies for addressing common issues arising in analysis of multidimensional data, such as interfering/overlapping signals[34], low signal-to-noise ratio (SNR)[70], establishment of limit of detection (LOD) in multivariate component systems[71-72], pre-processing[58, 73], and more. Many of the potential disturbances co-vary together with the sample matrix, and even more so as the chemical complexity of the sample increases. Some of the spectroscopic challenges that are explored in Part I of this thesis are the issues of nonlinearity, non-selectivity, overlapping and co-dependent signals, and calculation of figures of merit for multivariate systems.

The research presented here is not only concerned with biological matrices but also complex environmental phenomena. The earth's environments, be they on micro or global scale, are intricate systems influenced by a multitude of factors. A challenge that arises then is how such processes can be understood as thoroughly as possible. Multivariate techniques have indeed contributed to better understanding of distribution of polycyclic aromatic hydrocarbons (PAHs)[74] in the environment, but many environmental phenomena still await to be approached as multivariate problems. One such example is the production of halogenated volatiles by marine algae, which contribute to the degradation of tropospheric and stratospheric ozone. The decades of research dedicated to understanding the biogenic production of these volatile halogenated organic carbons (VHOCs) showed that the production is influenced by several factors, some of which are directly linked to climate change. However, the research has so far considered the influence of factors in a univariate fashion, creating thus a knowledge gap with regards to multifactor influences, synergies, and interactions. For the research focused on analysis of biogenic VHOC production, the author evaluates multivariate machine learning for understanding of influence of abiotic factors on said production, contributing thus to the lack of knowledge with regards to how several environmental parameters may influence VHOC production.

A detailed description of these challenges, as well as how they were addressed, can be found in the introductions to Part I and Part II.

PART I

APPLICATION OF MACHINE LEARNING IN RAMAN SPECTROSCOPY FOR THE PURPOSE OF STUDYING BIOLOGICAL MATRICES

INTRODUCTION

MACHINE LEARNING APPLIED TO SPECTROSCOPIC DATA GENERATED FROM BIOLOGICAL MATRICES

Næs et al.[32] listed the following problems that may arise in multivariate calibration and data exploration algorithms:

- **The non-selectivity problem** – the difficulty of finding all analytical responses in the sample.
- **The collinearity problem** – when variables have near or exact linear correlations.
- **The nonlinearity problem** – when data behaves nonlinearly and linear models cannot resolve this issue.
- **Calibration data selection** – the quality of the calibration model depends strongly on the number of samples and how they are selected.
- **The outlier problem** – variables or observations that deviate from the bulk of the data.

Complex biological matrices studied with Raman spectroscopy are in focus in Part I. Here, the **non-selectivity** problem and the **nonlinearity** problem were the most relevant to study. For example, to address selectivity, how should the experiments be designed; which factors should be considered? Can we represent and understand the matrix and the matrix effect by using DoE and projection-based algorithms (PCA, OPLS, T-OPLS)? Also, in the case of nonlinearity, DoE is central. Moreover, if the system studied is nonlinear, what can linear projection-based algorithms offer in terms of interpreting such systems?

THE NON-SELECTIVITY PROBLEM

International Union of Pure and Applied Chemistry (IUPAC) defines selectivity as the extent to which a method or an instrument can be used to determine analytes in matrices or mixtures without interference of other components in the matrix or mixture that display similar behaviour[75-76]. The non-selectivity problem is, in essence, the issue of finding the exact response of all constituents in the sample, including the analyte itself and its matrix.

One of the traditional ways of analysing Raman spectroscopy is the selection of a single wavelength and subsequent modelling based on univariate approaches. For a univariate analysis to work at its best, an analytical utopia should be established where detection of the analyte in a complex matrix should follow a set of criteria in order to be easily detectable[69]. First, if the aim is qualitative enquiry, the spectrum of the analyte should be a distinct and reproducible fingerprint that is entirely unique so that unambiguous analysis is as simple as possible. This minimizes the overlap with a sample matrix. In other words, the challenge here is the requirement of high selectivity, i.e. the input data must be that of pure analyte without any contributions from other species in the matrix. When responses others than that of analyte are present, a univariate model will become biased[35]. Selectivity is much harder to address univariately in case of spectra originating from biological matrices, as matrix effects on the analyte are unavoidable when working with intact cells or tissues. Secondly, if one wishes to perform qualitative and quantitative analysis, the spectrum should be consistent between different replicates and independent of the sample matrix. Lastly, the optical properties of the matrix should be noninterfering. This utopian scenario has been referred to as 'white multicomponent systems'[77], i.e. all spectral information for all chemical species and interferents present is available. This is in contrast to grey and black multicomponent systems, for which, respectively, partial information regarding all chemical species available or no information is available.

If the reader finds this to be a tall order, well, it is. Although white multicomponent systems are not uncommon, the majority of biological matrices are not amenable to separation of each chemical constituent. Hence, the decomposition of spectroscopic signals from such matrices and their subsequent classification poses a challenge as it is arduous to estimate the composition of the matrix in its entirety, let alone detect one or few analytes within it. However, with multivariate techniques one can handle the presence of non-analyte signals, provided that the response is not identical to that of the analyte.

Analyte detection in such mixtures often suffers from overlapping signals, and, in case the of Raman spectroscopy, inherently weak biological signals. Having *a priori* information (e.g. spectra of pure chemical species or computed spectra) about such samples can assist in attempts at classification and calibration, but the availability of such information is not always guaranteed[78].

Methods for the analysis of these grey and black multicomponent systems[77] are therefore are sought after. One such approach is the computational approach, where *ab initio* spectra of pure substances are mathematically modelled based on density functional theory (DFT)[79]. This eliminates the need for standards. Apart from the impracticality and resource-intensiveness associated with the acquisition of *ab initio* information, pure spectra do not account for the interactions between the analyte and its matrix, nor do they account for how such dynamics may affect spectral output. If the goal is to find an approximate correlation to interpreting the spectral data qualitatively, computed spectra can have an advantage. In particular, this can be advantageous in case of T-OPLS, as only the Y spectrum needs to be calculated. Yet such modelling can become challenging to interpret if the matrix effects on the analyte spectrum are large. If, however, quantitative calibration is the goal, it is arguably better to base modelling on the signals generated from the analyte in its original milieu or from as close to that milieu as possible.

An alternative way to analyse Raman spectra – which was utilized within the scope of this thesis - is to use the full spectrum and to apply multivariate algorithms, which can capture covariance between all variables. For instance, comparing two or several models through loading vectors **p** highlights spectral differences. Since PCA is a matrix decomposition method the loadings are, in essence, decomposed spectra. The first loading vector, **p1**, represents the largest spectral variation. When using Raman spectroscopy, such variation often includes variation in intensity and in the nature of the signal. The loadings that follow also often contain valuable information. In such loadings, it is possible to detect minor variation in spectral response, as well as in separate overlapping peaks, without the need for deconvolution

THE NONLINEARITY PROBLEM

The second requirement for an ideal univariate analysis is linearity. E.g. for calibration there ought be a linear relationship between spectral response and analyte concentration[35]. However, there are a number of nonlinearities in Raman spectroscopy, and these can make univariate approaches problematic. Some instrumental sources include stray light, baseline drift, and nonlinear detector response. When liquids and solutions are analysed, the sample itself can introduce nonlinearity through changes in turbidity, solvent composition, and temperature[80] . The issue of baseline inconsistencies is often solved by using pre-processing techniques such as rolling circle filter (RCF)[63], wavelet transformation[81], or Whittaker filter[65]. These methods do not resolve all nonlinearities, yet they can significantly improve SNR and thus subsequent interpretive power.

In addition to 'real' nonlinear sources listed above, there could also be 'apparent' nonlinear sources[80]. These include spectral variations, such as changes in absorption band width and shifts in the position of spectral peaks. For liquids and solutions, even the smallest change in temperature may influence the width and position of a band[80]. Changes in the dielectric strength of the solvent and in the concentration of equilibrium species can also change the features of bands[80]. These factors combined can mask smaller shifts in position and band width, and these errors may pose a problem when univariate calibration techniques are applied.

Nonlinearity of univariate character becomes less of a problem if the data is treated multivariatly. When nonlinear data is put though multivariate linear algorithms such as PL, the inherent univariate nonlinearities are captured better the more PCs are calculated. However, in calibration, if there are nonlinear relationships between the X and the Y matrices, the data cannot strictly speaking be well and consistently modelled with linear functions.

If linear algorithms are nevertheless used, there are a number of ways to address nonlinearity. As mentioned at the beginning of this section, spectral pre-processing can assist in improving some real nonlinear sources[32]. Another approach is to add extra terms to the X matrix that would represent nonlinear terms[32]. For instance, if quadratic interaction is suspected for variable A, one can add an A^2 variable to the experimental matrix thus including quadratic behaviour into the model. Provided that the data input has been designed, such interactions can be easily identified by analysing DoE coefficient plots (see examples of coefficient plots in Chapters 3 and 4). Further, the nonlinear data can be split into subsets in range, so that linearity is found in each of the subsets. Yet another way to avoid nonlinearity in is by deleting the nonlinear regions of the spectrum[32] provided that there is enough linear variation left at the other regions to make a valid model. Alternatively, it is possible to add weighting factors to reduce the significance of the nonlinear region, or approximate the curvature in principal components with nonlinear functions[80].

Last but not least; should linear algorithms be insufficient, nonlinear methods can be applied. Such approaches could prove particularly advantageous since linear projection algorithms may face difficulty in identifying pure spectra of chemical species from loadings generated from a multicomponent matrix[82]. Some examples of such nonlinear methods are locally weighted regression and neural networks[32]. It should however be noted that nonlinear calibrations require larger sample numbers, and, if too few samples are present, issues may arise with both accuracy and precision. With the projection-based methods in focus in this work however it is possible to handle the presence of non-analyte signals, provided that the response is not identical to that of the analyte and that the analyte and the non-analyte spectra vary somewhat independently of each other.

PLS and OPLS can also be successful when an internal standard (IS) is present[83]. However, a lot of noise in principal components or residuals, may render the model ineffective[80]. It is advisable to apply PLS when only small[84] nonlinearities are present if a satisfactory quality of calibration can be achieved due to a need for a lesser number of samples to achieve calibration.

RAMAN SPECTROSCOPY

Raman spectroscopy has seen many applications in analytical chemistry since the discovery of Raman scattering in 1928, and is a highly selective, non-invasive technique used to observe rotational and vibrational modes of molecules. Raman scattering effect occurs when light, typically from a laser, is scattered with different wavelengths than those of incoming light. This inelastic scattering of photons occurs as a result of a change of polarisability of molecules[85]. In other words, the polarisability of the analyte changes as consequence of vibrations[86].

There are six different types of scattering phenomena. Rayleigh scattering occurs when scattered light has the same frequency as the excitation laser, while Raman scattering is when the scattered light has a shifted frequency[87]. The Raman scattering in turn can have two different outcomes, Stokes and anti-Stokes Raman. A Stokes shift emits light of lower energy than that originally absorbed and anti-Stokes does the opposite. Other types of scattering are pre-resonance Raman, resonance Raman, and fluorescence, the latter often being the cause of nonlinearities in the baseline component.

The scattering pattern, or fingerprint, is unique for each compound; therefore it is possible to conduct qualitative analysis of complex sample mixtures or individual analytes in a composite matrix. This in turn means that since the signal is intrinsic to the studied system, using staining and labelling methods becomes obsolete[88].

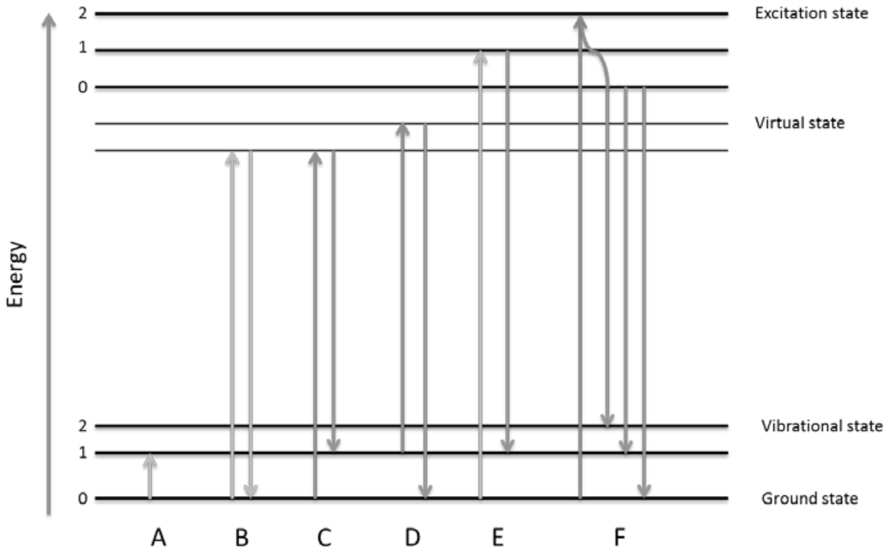


Figure 10. Jablonski diagram. (A) infrared absorption; (B) Reyleigh scattering; (C) Stokes Raman scattering; (D) anti-Stokes Raman scattering; (E) resonance Raman scattering; (F) fluorescence.

Also, in comparison to infrared spectroscopy, Raman enables measurements in aqueous samples without the interference of water, which is a weak Raman scatterer. Because of this Raman spectroscopy facilitates the analysis of biological systems such as living cells.

One of the applications of Raman spectroscopy is the study of biological matrices and processes, as Raman has the distinct advantage of water not being an interferent. This was of particular advantage in Chapter 4, where wood was the examined. There are several analytical techniques for quantification and detection of analytes inside biological matrices, ranging from mass spectrometry to chromatographic methods[57]. However, such methods usually involve tedious sample preparation steps and long measurement times.

Yet, as already stated, biological matrices are often complicated chemical system, making distinguishing between all the chemical species present a challenge. In Raman spectroscopy, there is the additional issue of spectral bands having low intensities and the presence of interfering fluorescence.

As Raman signals are intently weak, complementary techniques that can provide signal enhancement. One such method is resonance Raman – by illuminating the sample with a laser wavelength close to the electronic absorption of a molecule, its signal can increase by a factor of 10^6. This process gives rise not only to inelastically scattered light, but also to fluorescence and luminescence[89]. This attenuation of the laser wavelength to a specific molecule can facilitate analysis of complex matrices, as it provides higher selectivity and sensitivity because of increased SNR.

The issue that may arise when using resonance Raman is the occurrence of interfering fluorescence, as florescent species may use the same resonance transition wavelength[57]. Further difficulties may arise due to thermal decomposition and photoreaction due to the strong absorption of the excitation light[57]. Here, resonance Raman was not utilised as a signal-enhancement technique here.

To further increase the sensitivity of the technique, the use of metal colloid nanoparticles to achieve signal enhancement from the matrix is possible. Moreover, the presence of the metal enhancement surface also quenches fluorescence. This is highly advantageous when working with samples containing biological chromophores, such as. DNA. The utilisation of metal surfaces to enhance signal is called surface enhanced Raman spectroscopy (SERS), which was used in Chapters 1 and 3.

SURFACE-ENHANCED RAMAN SPECTROSCOPY (SERS)

Despite its advantages in the analysis of complex biological matrixes, Raman spectroscopy's insensitivity renders quantitative (and sometimes qualitative) analysis difficult. This insensitivity comes from two factors. Firstly, Raman's low scattering probability, with only 1 in 10^8 photons scattered, produces weak signals, which makes it impracticable to use when analysing sample concentrations below millimolar range[57]. Secondly, Raman signal struggles with competing fluorescence (**Figure 10**), which may be of higher occurrence than the Raman scattering itself. As a consequence of these shortcomings, several other techniques have been developed to address these issues, including SERS. SERS is successful both in enhancing the Raman signal and in quenching fluorescence.

In SERS, the surface of a metal interacts with photons generated by an enhanced electric field in the close proximity of the metal's surface, producing a significantly enhanced Raman signal[90]. The metal surfaces usually are composed of noble metals such as gold and silver[89, 91], but also other substrates such as graphene[92]; hybrid metal substrates, such as SiO_2-encapsulated gold particles; metal alloys; and alkali metals[93]. The enhancement surfaces come in a broad variety of shapes depending on the purpose behind their usage. Some examples are laser-ablated surfaces, electrodes, metal tips, and nanoparticles[91, 94]. Within this thesis, nanoparticle suspensions were the only substrates used to achieve surface enhancement. This is due to the fact that SERS enhancement is strongly distance-dependent, meaning that bulk substrates, such as electrodes or surface structures, are not as suitable for the purpose of *in situ* studies of living cells (see examples in Chapter 1 and 3), because the cell and the analyte of interest must be as close to the metal surface as possible[83]. The desired proximity can therefore be achieved by using colloidal substrates, which can be internalized into cells. Despite the lack of coherent structure and because of a lower reproducibility compared to solid substrates, colloids can disperse throughout the sample giving them an edge in cellular studies. In addition, colloidal substrates are cheaper to produce, which contributes to their popularity in research[95]. So, only the theory with regard to nanoparticle suspensions will be considered in this section.

The mechanism behind the surface enhancement phenomenon in SERS is a combination of two enhancement mechanisms, that is, chemical enhancement and electromagnetic enhancement. Chemical enhancement mechanisms are considered as contributing less to overall enhancement than electromagnetic mechanisms. Chemical enhancement involves molecular interaction with the metal's surface. This kind of enhancement can occur, for instance, by a charge transfer in the creation of a new chemical bond (changing thus the molecular polarizability of the analyte), or by chemisorption-induced resonance Raman scattering[44, 47]. Chemical enhancement can be negatively influenced by parameters such as defects in a metal surface and in oxidation layers.

Electromagnetic enhancement depends instead on a match between the surface plasmon resonance of a metal and laser frequency[85]. The excitation is achieved by free electrons oscillating in resonance with the light[44] which results in an effect called localised surface plasmon resonance (LSPR), provided that the size of the metal nanoparticle is much smaller than the excitation wavelength[96] (**Figure 11**).

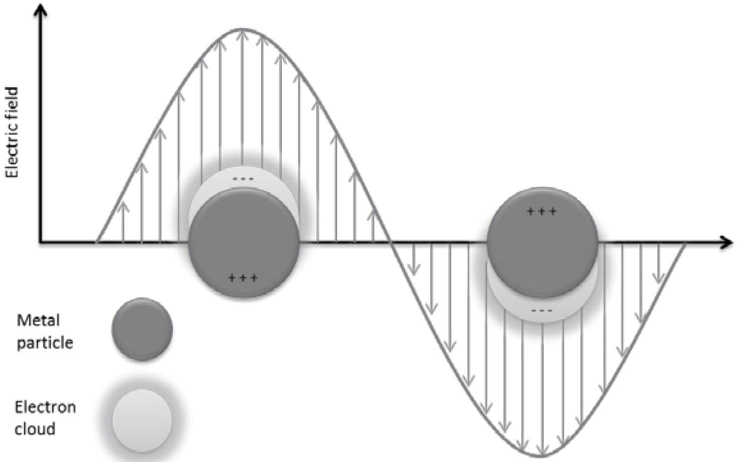


Figure 11. Illustration of the LSPR effect. When incident light interacts with a metal nanoparticle, the oscillating electric field initiates the oscillation of the electron cloud.

LSPR is the frequency resonance between the electron cloud of the metal nanoparticle and the incident electromagnetic field[97]. Interaction with light is determined by the metal surface's roughness and, in case of colloidal nanoparticles, the size, shape, and aggregation state. In addition to the aforementioned parameters, the LSPR is also influenced by the refractive index of the medium[86] and by the wavelength of the incident light. The surface plasmons generate a large electromagnetic field, which decreases by 80% in strength at about 10 nm from the surface of the colloid[98]. Molecules in the vicinity of this field will experience an enhancement of their signal[86].

The greatest drawback of using unmodified colloidal nanoparticles is the issue of the irreproducibility of signal enhancement. Since the enhancement is dictated by the colloidal shape and size, it becomes difficult to reproduce as colloids aggregate. It has previously been noted that metal nanoparticles aggregated into assemblies produce the strongest SERS signals. Although a certain level of aggregation is often desirable in order to achieve the highest possible level of sensitivity[99-101], too extensive aggregation has been previously noted as lowering SERS signals[102]. For instance, a cluster size of up to ca 20 gold nanoparticles has been reported to increase SERS signals, which would then decrease beyond this cluster size until the signal increase reached a stable asymptotic value[102]. Reproducibility can be improved by promoting controlled aggregation, by e.g. controlling nanoparticle aggregation during synthesisis[103], or by evaporating the nanoparticle suspension on a surface[104]. Another approach to control aggregation is by surpassing the aggregation completely. The latter can be achieved by modifying the nanoparticle surface so that they remain monodispersed. In Chapter 1, this was done by modifying the nanoparticles with a thiol self-assembled monolayer (SAM), in order to avoid aggregation[105].

CONFOCAL RAMAN MICROSCOPY

Analysis of inter-cellular contents can be enhanced if another Raman sub-technique is used, namely confocal Raman microscopy. As the name suggests, a Raman spectrometer is combined with a confocal microscope. The confocal properties make the selection of small volumes possible, which makes this combination favourable for measuring for instance, living cells and for creating highly detailed 2D and 3D. Confocal Raman microscopy is therefore widely used for obtaining high-resolution Raman images of cells[90]. These are the subject of study in Chapter 1.

In a confocal Raman instrument (**Figure 12**), the laser beam is focused on the sample through a microscope objective. Measurements performed in small volumes prevent distant scattered light from entering the microscope objective, thus enabling a discriminant illumination of a selected point in the sample. The light from anywhere else in the sample is hence nullified. The light emitted from the sample is then passed through the notch filter, the pinhole, and a slit. After passing the slit, a holographic grating is illuminated, which in turn disperses the incoming light depending on the light's frequency[44]. Finally, the light reaches a detector array, usually a charge-coupled device (CCD), where it is processed electronically[106].

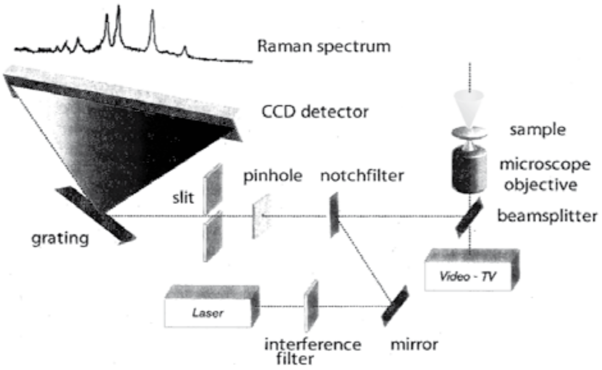


Figure 12. The set-up of a confocal Raman spectrometer[57]. This particular image illustrates an instrument with inverted objective; it is not a common set up, rather, the sample is more commonly illuminated from above.

However, since only a selected point is measured, it becomes necessary to map everything larger than a pixel. Fortunately, the confocal Raman set-up used here also allowed for measurements in x-y axis thanks to a motorised microscope stage. It is also possible to measure in z-axis if a piezoelectric actuator is installed in the microscope[57]. The optical resolution is in the order of 1 μm in all directions, whereas the mechanical resolution of xy-stage and actuator is 0.1 μm.

Confocal Raman microscopy was used in Chapter 1 for imaging of living cells and for quantification of intracellular drug concentrations. In Chapter 2, a confocal Raman setup was used to quantify an analyte in archaeological wood. However, the measurement volume was deemed too small, and hence confocal Raman setup was not suitable for this particular enquiry. The method could have been more suitable for creation of distribution images of the analyte.

CHAPTER 1

T-OPLS METHODOLOGY TO COMPENSATE FOR LOW REPRODUCIBILITY OF INTRACELLULAR SERS FOR SUBSEQUENT QUANTIFICATION OF DOXORUBICIN

INTRODUCTION

Modern intensive treatment with anti-cancer drugs is the major cause of improved cure rates in childhood cancer. The effect of the drugs is determined by the blood and intracellular drug concentrations attained during treatment and to the inherent sensitivity of the cancer cells to the different agents[107-108]. The drugs have a steep dose/efficacy curve, with a narrow therapeutic index, which requires careful dosing to balance therapeutic effects and side effects. There are large differences between individuals in uptake, metabolism, and in the excretion of the drugs, which, therefore, in some patients, either can lead to a low drug exposure with reduced treatment efficacy, or to a dangerously high exposure resulting in unacceptable, sometimes life-threatening side effects. To a large extent, these individual differences in drug disposition are caused by polymorphisms in the genes involved in drug metabolism, which result in variable drug exposure that may have a significant impact on both chance of cure and severity of side effects[109].

The cancer types in focus in this chapter are leukaemias, the cancers of the blood. One of the leukaemias – acute lymphoblastic leukaemia (ALL) – is the most prevalent leukaemia type in Swedish children[110]. In ALL paediatric treatments, it is common practice to dosage cancer drugs in proportion to the estimated body surface of the patient. The body surface area (BSA) is calculated as follows[47] (**Equation 1.1**).

$$BSA = \sqrt{\frac{(weight \cdot height)}{3600}} \qquad \text{(Eq. 1.1)}$$

Equation 1.1. The body surface area (BSA) formula, where weight is given in kilograms, height in centimetres, and the BSA in square meters.

A widely used cytostatic drug doxorubicin (DOX) is one of the alternatives for treating paediatric ALL. Despite its efficiency, DOX has a small therapeutic index, which means that there is a narrow dosage range between the drug being therapeutically effective and it becoming toxic. For instance, DOX poses a risk of cardiotoxicity[57]. BSA is used mainly due to the fact that conventional methods for measuring drug concentrations in blood, as well as in malignant and normal cells, are laborious and time-consuming and are not applicable in a clinical setting as a guide to drug dosing. Yet, DOX administration based on BSA has its problems. First and foremost, the BSA approach does not account for the inter-individual variability of metabolic genotypes in patients. For example, it can be said that the median plasma concentration of DOX varies significantly (22.6-334 ng/mL) between children depending on, among other factors, their age[111]. A second issue lies in the inter-variability of a cancer phenotype, meaning that the cancerous cells may develop a resistance towards DOX by upregulation of their efflux pumps. Due to these individual variations there is always a risk of treatment with DOX becoming either inefficient or toxic if BSA is used a guide to dosing.

The goal of the research presented in this chapter was to develop a reliable and rapid method for measuring drug concentrations inside the blood cells of paediatric cancer patients to circumvent the shortcomings of BSA. This would allow for individual dose titration, with the potential of increasing cure rates with fewer side effects, and would also provide an invaluable tool for drug development, pharmacogenomic studies, and for examining the mechanisms of drug action and resistance.

The author used a linear projection algorithm T-OPLS in order to arrive at a method capable of quantifying anti-cancer drugs intracellularly, which would also assist in illustrating the drug's distribution within the cell. The primary aim of this enquiry was to make a contribution to the established clinically viable method for anti-cancer drug quantification in patients suffering from ALL. The cells under scrutiny were lymphocytes, monocytes, granulocytes, and PC12 cells, where the first three were white blood cells sampled from human blood of a healthy individual. PC12 cells, on other hand, belong to a cell line derived from pheochromocytoma in rats, the cancer of adrenal medulla. More data was gathered from other types of blood cells, neutrophils, to evaluate whether the method could be used on other types of white blood cells unaffected by the common blood cancers. This was done in part to estimate how healthy cells not involved in ALL are affected. Lastly, PC12 cells, which produce various catecholamine neurotransmitters, were used to see if it was possible to separate the DOX signal from the dopamine (DOP) signal, thus accumulating further information about the viability of the developed method. The cells were mapped with surface-enhanced Raman spectroscopy (SERS), and spectra were collected from each pixel, thus producing hyperspectral images. The acquired data for these types of measurements are a mixture of analyte signals, matrix effects, and noise. To fully utilize all spectral information, and, at the same time, separate the essential information from the noise, multivariate analysis is commonly used[112] and successful calibrations for low amounts of chemotherapeutic drugs in plasma using SERS have been made[113]. To address the issue of reproducibility inherent to SERS, gold colloids coated with 4-mercapto-benzonitrile (MBN) were used, where MBN also functioned as an internal standard (IS)[114].

The secondary aim of this research was to examine T-OPLS as a tool for quantification of an analyte in a complex biological matrix, as well as to provide further insight into the advantages and disadvantages of T-OPLS in spectral analyses. T-OPLS is still a poorly explored machine learning method, both with regards to the theory and to its practical application in analytical research. So far, T-OPLS has been applied in the context of hyperspectral imaging. Some examples include *in situ* analysis of carotenoids[50, 115], paracetamol in tablets[55], and halogenated non-volatile compounds in macroalgae[52]. None of these studies provide a comprehensive investigation and validation of how the algorithm functions when applied to complex systems. The exception to this is the comparative study by Dumarey et al[54], where multivariate curve resolution-alternating least squares (MCR-ALS), weighted OPLS, and T-OPLS were evaluated as tools for hyperspectral imaging . The insights provided in this chapter contribute therefore towards deeper understanding of T-OPLS.

HYPERSPECTRAL IMAGING

When used for purposes of imaging, confocal Raman spectroscopy allows focusing the laser point on a small area of the sample and then moving on to the next point, until the entire area of interest is mapped. Each point becomes a pixel in an image and contains a Raman spectrum representing the fingerprint of that particular point in the sample. These spectra can then be reconstructed into hyperspectral images by putting the intensities of score vectors **t** at each separate frequency as images. This can be achieved by, for instance, plotting the intensity of one peak of interest or the whole spectrum[116]. It is done by first stacking the two-dimensional images into a three-dimensional matrix, also called a hyperspectral data cube, where the third axis makes up the spectral axis[82].

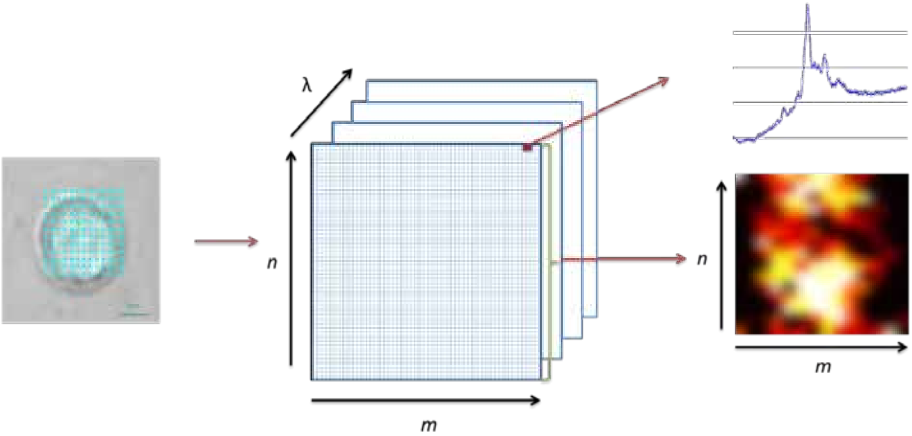


Figure 1.1. Hyperspectral image generated with point mapping. Images in *n* and *m* dimensions may be unfolded to a table of spectra (observations), similar to an experimental matrix X used in PCA. If multivariate calculations are made on the observations, the results from the multivariate score vectors can be reconstructed as images in the same pixel coordinates as the original image.

The hyperspectral data cube is generated in this chapter with point mapping, i.e. a grid is defined above the sample surface. A spectrum is then measured at each point of the grid; the points then function as pixels in an image[82]. The size of each pixel and the distance between pixels can be set in the instrument. Although point mapping is slow compared to modern alternatives, such as line imaging and global imaging, the limitations of the instrument used in this work did not make the use of such alternatives possible. On the other hand, point mapping offers a better SNR[82] and offers better spectral resolution[117]. Therefore, point mapping can be better suited than line or global imaging to the a study of the distribution of compounds with low concentrations[118].

As evident from the way the hyperspectral cube is constructed, it contains abundant and complex information. The classical way of analysing the hyperspectral cube begins with spectral pre-processing (see the section above for more information), which aims to reduce the contributions of baseline variations and other undesirable effects, such as cosmic rays. Distribution images such as the one seen on the right side in **Figure 1.1** may then be generated, often with the aid of multivariate algorithms, such as PCA, PLS, 3-way multivariate methods, or artificial neural networks (ANN). This chemometric analysis allows for the spatial location of the compounds in the sample. Lastly, quantitative parameters can be extracted from the images[82].

The choice of algorithms is dictated by the data available, by the types of noise and nonlinear effects, as well as by which compounds are to be detected. The simplest and most straight-forward method used is univariate analysis. A single wavelength of a compound is selected, say, a peak at 1000 cm^{-1}, and then the distribution of this peak is mapped, thus localising the compound to which this peak corresponds. The highest intensity in the image represents the highest abundance of that peak in the sample. Alternatively, a peak can be integrated and its area can be used in a similar way to the univariate approach above[82]. The greatest disadvantage of these methods is that they necessitate *a priori* data about the studied compounds, i.e. they require white multicomponent systems[77]. In the scope of this thesis, however, the systems studied are predominantly grey, i.e. only partial *a priori* information is available about the sample. This is not an issue if, as is the case of drug imaging in a sample, the analyst has the spectrum of the drug. Unfortunately, if the analyte matrix is complex, such as a living cell, identification can be hampered by overlaps from the matrix or can be hampered if the analyte has a low SNR. For instance, the spectra of the analyte and the IS in Chapter 1 are known, but it is not known how their spectral behaviour is affected by matrix effects and to what extent the matrix signals overlap with the analyte. Put differently, the univariate approach in this particular case is considered to be not sufficiently selective.

This **non-selectivity** issue can be addressed if the images are analysed multivariatly. PCA, for instance, is particularly suited in this case as it reduces the dimensionality of the data, thus making it possible to sort the noisier parts of the spectra into less relevant components. The scores in the PCA are then folded back to form an image which represents pixel variability along the corresponding loading vectors. PLS-based algorithms, on their own or combined with DA, are also favourably applied when comparing several hyperspectral images. In the case of the PLS-based algorithms, Y can be set as a continuous variable, e.g. a concentration variation of the sought analyte. If combined with DA, a discreet Y is instead created. OPLS, which is a cousin of PLS that is an interpretable cousin, has also been used for imaging purposes[119-124].

The greatest challenge in applying linear projection algorithms, whether it is imaging or for other enquiries, is that the loadings can be difficult to relate to chemical species present in the sample[82], i.e. it can be challenging from loadings alone to verify that we have a model that is valid for the target compound. This problem can become even greater if the data used for imaging suffers from **nonlinear issues**. There are several ways this issue can be addressed; nonlinear algorithms can be applied, but, as discussed previously, such approaches require larger amounts of data, which was not possible to acquire within this work. Other ways include the application of the DoE and, in case of this chapter, the T-OPLS algorithm. T-OPLS is an attractive method for imaging, as its Y matrix can be set to the spectrum of the sought substance – the variation irrelevant to the sought spectrum should be sorted into orthogonal components, making imaging easier to achieve. Examples of this imaging approach can be seen in the works of Abbas et al.[50, 52] and Josefson et al.[53]. Once a hyperspectral cube is reduced and then extracted into images, the next step is the interpretation of the distribution seen in the images and subsequent quantification. Examples of such quantifications can be found in the literature[125-126].

CELL IMAGING WITH SERS AND CONFOCAL RAMAN SPECTROSCOPY

Cellular imaging can be roughly divided into targeted and label-free methods. A common targeted method in imaging is based on fluorescent probes, which can be engineered to target a specific analyte within the cell. This allows visualisation of the distribution of the analyte of interest intracellularly. Unfortunately, florescence based imaging does not provide specific molecular information, which in turn limits the amount of target probes one can introduce into the cell[116]. Imaging based solely on vibrational spectroscopies, such as Raman and SERS, offers the advantage of distinct molecular information for each molecule and thus allows simultaneous identification of several intracellular species. In addition, Raman is sensitive to any structural and molecular changes, making it advantageous for the study of biological processes[127]. However, since fluorescence interferes with Raman signals, combining fluorescent probes and Raman spectroscopy is a challenge. These seemingly incompatible techniques have nonetheless been combined by van Manen and Otto[128] by utilising semiconductor fluorescent quantum dots. However, their work has not been evaluated for the research in this chapter.

As mentioned earlier, normal Raman has a weak signal. If the aim is to detect an intracellular component with a label-free method, this might prove to be a challenge due to limit of detection (LOD) limitations. Imaging with standard Raman spectroscopy is therefore limited to the study of chromophores, where resonance Raman provides the needed enhancement. To avoid the demerit of low signal intensity and interfering fluorescence, Raman microspectroscopy has become an interesting alternative for imaging. This in turn means that complex samples such as cells can be studied directly, with no or little sample preparation. In addition, SERS can provide stronger signals and still allow for cellular imaging if used in combination with a confocal microscopy set-up. However, for intracellular imaging to occur, SERS requires the internalisation of metallic enhancement surfaces by the cells, which is a disadvantage compared to fluorescence imaging. Moreover, since SERS enhances signals of all adjacent molecules, be they matrix or analyte, it means that the analyte signals can become lost in the plethora of matrix signals[116]. One way to resolve the analyte from matrix signals is the use of chemometric methods. Such approach is particularly advantageous if the sought analyte is sufficiently distinct from the cellular matrix, or is not intrinsic to the matrix. In this work, the analytes of interest were DOX and the catecholamine DOP, both of which have spectral features distinct from the cell.

GOLD NANOPARTICLES FOR SERS

SERS enhancement is strongly distance dependent meaning that bulk substrates, e.g. electrodes or surface structures, are not as readily applicable to the purpose of imaging intracellular environment of living cells. This is due to the fact that the cell and the analyte of interest must have as optimal proximity to the metal surface as possible[83]. The desired proximity can be achieved by using colloidal substrates. Despite the lack of coherent structure and low reproducibility compared to solid substrates, colloids can disperse throughout the sample giving them an edge in cellular imaging.

In general, gold colloidal nanoparticles (AuNPs) are considered to be the most advantageous choice for the imaging of living cells[90]. Since gold is inert, the background fluorescence from gold nanoparticles and species present in the sample is reduced. The inertness of gold also reduces the heat damage to cells[97]. For this thesis, AuNPs were also advantageous due to their negative surface net charge. For spherical colloidal substrates generated LSPR is isotropic, i.e. uniform in all directions from the colloidal particle[86].

The surface plasmon band can be adjusted for spherical particles by changing their size – this causes the active spectral range to shift towards higher wavenumbers with increasing particle size[96]. To achieve optimal enhancement, the size of nanoparticles should be matched to an appropriate excitation wavelength. In addition, the choice of laser for live cell imaging is important to consider due to factors such as photo-damage to the cells, auto-fluorescence of the matrix components. In the case of gold specifically, the plasmon resonance condition is readily met both by using near-infrared (NIR) and by using visible spectral ranges. However, when using AuNPs, wavelengths within the range of 600 to 800 nm are considered best. This is due to the fact at shorter excitation wavelengths gold produces photoluminescence that potentially can interfere with the SERS signal[90]. Longer wavelengths on the other hand would produce lower signal intensity. Based on these considerations, 633 nm wavelength was used throughout this chapter.

INTRACELLULAR UPTAKE OF NANOPARTICLES

The most common way to introduce nanoparticles into the cell is through endocytosis from an external solution. The uptake is dependent on the size of the nanoparticles, toxicity, and by which endocytic mechanism uptake occurs[116, 129]. Factors that influence the uptake also include the particles' surface charge, and the cell type that internalises the nanoparticles[130]. In general, smaller nanoparticles mean lower toxicity and a higher likelihood of internalisation, and if they are spherical, the internalisation is further favoured[129, 131]. It has also been shown that nanoparticles tend to be dispersed through the cytoplasm or congregate near the nuclear membrane[132-133]. The particles used in this chapter were spherical and 60 nm in diameter, which suggests cellular entry through clathrin-mediated endocytosis and/or caveolae-mediated endocytosis[134]. These uptake mechanisms are referred to as receptor-mediated endocytosis, and that occurs when nanoparticles exposed to cell media become coated with proteins, which then interact with the receptors on the cell membranes[129]. The process of the protein coating of nanoparticles is referred to as opsonisation.

THIOL SELF-ASSEMBLED MONOLAYERS AS AN INTERNAL STANDARD

If SERS is used for quantitative imaging, it is crucial that the enhancement is consistent[135]. A major problem in the use of colloidal nanoparticle suspensions for surface enhancement is the reproducibility of the enhancement factor, which is determined by the aggregation, roughness, size and shape of the metal colloids, as well as by the physiosorption of the analyte to the enhancement surface[83, 136-137]. There have been numerous advances in research where the issue of SERS reproducibility has been examined for both qualitative and quantitative investigations[138-142] However, many of the reproducible SERS methods published cannot be applied if the goal is the internalisation of SERS substrates into living cells.

Provided that internalisation is successful, the application of colloidal nanoparticles for imaging of living cells involves uncontrollable distribution within the cell and the irreversible uptake may cause problems, especially in time-dependent studies[97]. A further challenge, which was mentioned in the beginning of this chapter, is **non-selectivity**. All molecules in the vicinity of SERS substrate are enhanced, generating complex spectra of matrix species and the analyte.

Both issues can be addressed by coating the colloidal nanoparticles with molecules that would saturate the enhancement surface[116], but would not interfere with the analyte. The signal of the analyte will not be optimally enhanced with such a coating; however, this trade-off allows for correction of the enhancement factor. One example of such functionalisation is self-assembled monolayers (SAMs)[83]. SAMs provide high coating percentage, minimized random interactions with non-analyte molecules, and minimized interaction between the nanoparticles and the matrix[83,105]. Using the SAM principle, we developed a method where the colloids are coated with a monolayer of a Raman active substance[143]. The choice of the SAM molecule was pivotal; not only was it crucial to in ensuring cellular uptake, but also in making sure that the molecule could be used to compensate for the irreproducibility of the sought SERS signals. The Raman signal from the monolayer was therefore used as an internal standard (IS) to allow for normalisation of the enhancement. This method was first applied by Lorén et al[83, 143] – the research here draws upon those insights. In principle, the IS will be enhanced in a similar way to the analyte so that the signals are enhanced simultaneously, and the ratio between them will be constant[114]. Thiols are a common choice for SAM when working with AuNPs[144]. These sulphur-containing compounds are favoured when coating AuNPs because the Au-S bond is quite strong (40 kcal mol^{-1}) [145]. Adsorption of thiols to AuNPs is a chemically irreversible process that can be described by following equilibrium:

$$\mathrm{RSH} + \mathrm{Au} \leftrightarrow \mathrm{RS\text{-}Au} + \tfrac{1}{2}\,\mathrm{H_2} \qquad \text{(Eq. 1.2)}$$

Equation 1.2. Adsorption equilibrium of thiols to gold surfaces, where RSH is the general formula for a thiol and Au is the gold adsorption site[145].

The adsorption of the thiols to gold is very rapid, but is followed by a reorganisation of thiol groups on the gold surface, a process that can take days[146]. The adsorption process can be described by the **Figure 1.2**:

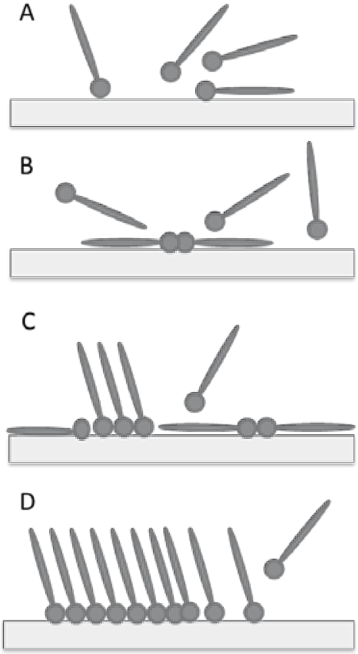


Figure 1.2. The stages of the self-assembly of a thiol on gold surface: (A) physisorption/chemisorption, (B) ' lying-down' phase formation, (C) initiation of 'standing-up' phase, (D) completion of 'standing-up' phase[146].

The first stage of the self-assembly of thiols involves both physisorption through van der Waals interactions between each molecule, and chemisorption through covalent binding of the sulphur to the gold surface (**Figure 1.2A**). Both mechanisms determine the structure of the SAM. After these initial sorption processes, the SAM starts to grow. The growth is most favourable in defective sites of the AuNPs, in which the thiol molecules 'lie down' (**Figure 1.2B**). Through this nucleation of lying down thiols, the SAM reaches saturation and thiols start to 'stand up' (**Figure 1.2C**), a process that can take up several days until it finalises (**Figure 1.2D**). Finally, the surface coverage (θ), which is a measure of how many absorption sites are occupied divided by how many sites are available[76], reaches a maximum at ca $\theta = 1/3$[146].

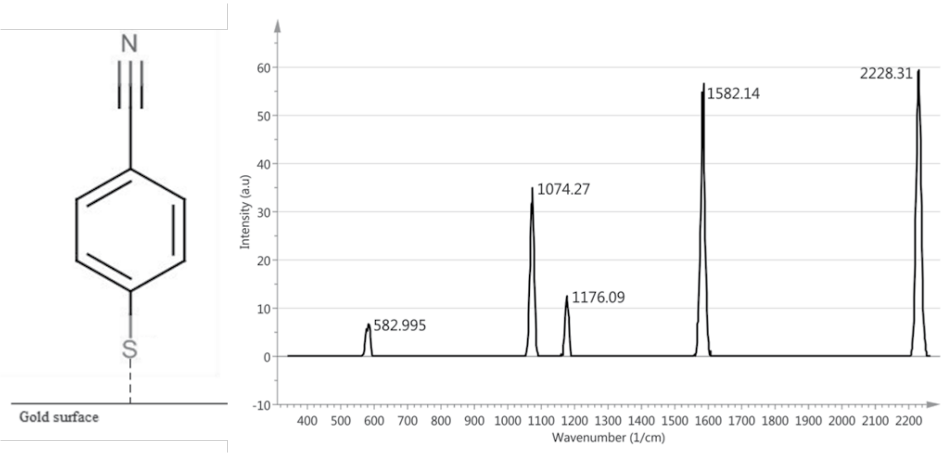


Figure 1.3. (Left): Structure of 4-mercapto benzonitrile (MBN) attached to gold surface. (Right): spectrum of MBN coated AuNPs. The spectrum was pre-processed with RCF. Baseline between peaks was detected by using Savitzky-Golay derivative filters.

In this chapter, AuNPs have been coated with an arenethiol, 4-mercapto benzonitrile (MBN) (**Figure 1.3**), a substance previously used as an IS[83, 143]. This particular thiol was chosen due to its small size, which is in angstrom (Å) range. The small size of the thiol would ensure that the enhancement of nearby molecules would not be reduced (remember the 80 % percent enhancement loss at the distance of 10 nm[98]). In terms of coverage, arenethiols have in general lower surface coverage compared to alkanethiols, with some reported values of $\theta = 1/10$[147]. No value could be found in literature specifically for MBN.

In our measurements, MBN displayed distinctive peaks at 365, 583, 1074 (C-H in-plane deformation[148-149]), 1176 (C-H in-plane deformation[148-149]; C-CN stretching[149]), 1582 (aromatic stretching[148-149]), and 2228 (-C≡N stretching[149-150]) cm^{-1}. The peak at 583 cm^{-1} has been previously observed in DFT calculated spectra[149], but has not been assigned. At neutral pH, this SAM was expected to have neutral net charge. This could be disadvantageous since it has been shown that positively or negatively charged functionalised nanoparticles penetrate the cells faster than non-functionalised or neutral nanoparticles[151].

MBN is notable for its strong Raman signal and peaks that fall outside the fingerprint region. Our pre-studies have shown that a 1:1 blend of 1 mM MBN and AuNPs produced a clear and strong signal. The extent of the MBN coverage on AuNPs was examined with time-of-flight secondary ion mass spectrometry (TOF-SIMS) (results courtesy of Dr. John Fletcher; results were not examined further in this thesis). It was shown that the AuNP surface reaches saturation when approaching 10 mM MBN at a volume mixture of 1:1 with AuNPs.

HUMAN WHITE BLOOD CELLS

Figure 1.4. illustrates the blood cells found in humans. The white blood cells studied here were monocytes, neutrophils, and lymphocytes (red circles in **Figure 1.4**). The first two cell types stem from the precursor myeloblast, while the lymphocytes stem from lymphoblast. Myeloblast and lymphoblast neoplasticity are behind acute myeloblastic leukaemia (AML) and ALL, respectively, and are two of the most common leukemia types. Leukaemia, a cancer affecting the bone marrow, is the most common childhood cancer in Sweden, constituting 30 % of all cases, and ALL is particularly prevalent in younger children[110]. Overproduction of lymphoblasts occurs in the bone marrow and has adverse effects on the normal production and function of other blood cells. In time, the cancerous cells overpopulate the bone marrow and start to migrate to other tissues.

Monocytes are phagocytic cells, meaning that they can take up solid matter throgh their endocytotic mechanisms[152]. There have been a number of studies on nanoparticle uptake mechanisms in cancerous monocytic cell lines, such as THP-1[153-154] and U937[155], and there have also been studies on primary (i.e. cells from normal donor tissue) monocytes[156-157]. Neutrophils, the main constituents of the granulocyte fraction in human blood, are also phagocytic and follow similar endocytic pathways as monocytes.

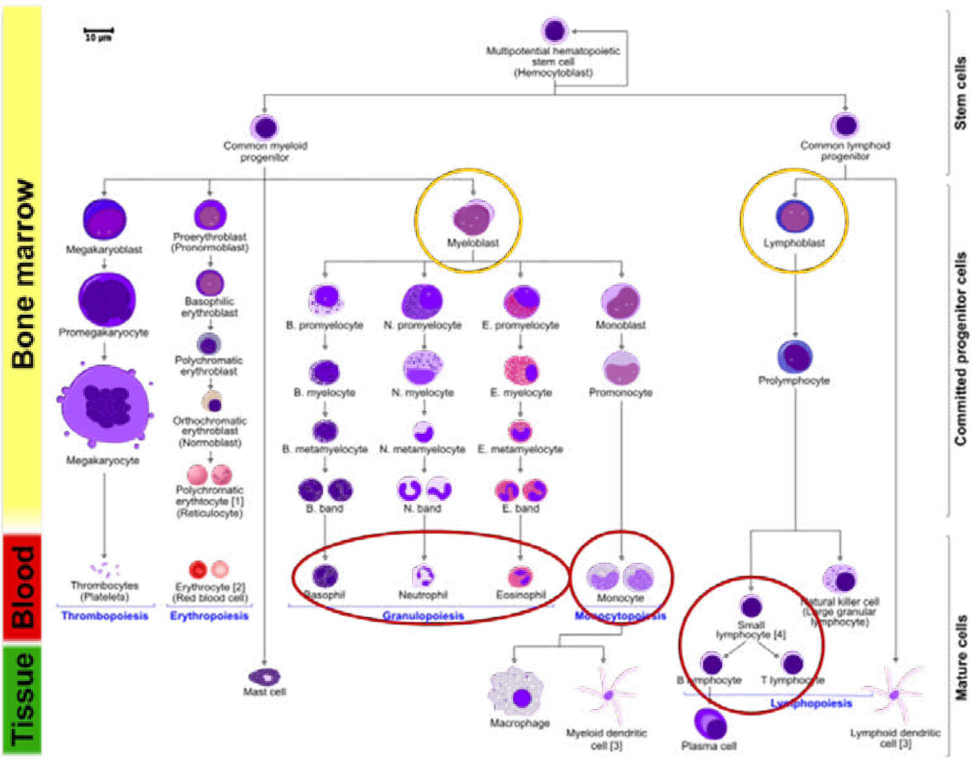


Figure 1.4. Differentiation of blood cells in humans (image license: CC BY-SA 3.0)[158]. Cell types studied in this chapter are marked with red circles. No differentiation was made between T and B lymphocytes. The granulocyte fraction (basophils, neutrophils, and eosinophils) mainly consists of neutrophils, and so no differentiation was made between the granulocytes. The orange circles mark the lymphoblast and myeloblast, the cell types affected by ALL and AML, respectively.

In the research presented below, all of the mentioned white blood cells were primary, acquired from a healthy male donor. Due to ethical concerns, we were hesitant to request blood from ALL patients before a reliable method was established. Also, blood from a healthy individual was readily available at the time. ALL cell lines from human paedeatric patients are commercially available, and would have been a suitable alternative for this research enquiry. Examples of ALL cell lines include variations of CEM and CCRF-CEM cell lines, MOLT-3, and TALL-104. Using such cell lines would make the developed method more comparable to the literature, as many intracellular drug quantification studies have been performed on cancerous cell lines. However, this was not done because of the need of having a robust method before an attempt was made to measure cancer cells, which may have had a high biological intervariability in terms of drug uptake and efflux. An additional rationale behind using healthy cells instead of equivavelent cancerous cell lines was that cancerous cells have been shown to accumulate nanoparticles more readily compared to their healthy counterparts[159-162]. As the toxicity of cytostatic drugs also affects healthy tissue, we also wanted to first establish a method for healthy cells in order to examine the issue of the side effects associated with chemotherapy.

DOXORUBICIN (DOX)

DOX belongs to a group of naturally occurring antibiotics called anthracyclines. DOX was first isolated in 1967 from bacteria, and is still successfully used for treating a wide range of neoplastic abnormalities such as different solid tumours, lymphomas, leukaemias, and paediatric cancers[163]. DOX enters cells through passive diffusion and accumulates predominantly in cells exhibiting high proliferation rates. Due to its planar structure, DOX intercalates with DNA, which in turn inhibits macromolecular synthesis by topoisomerase II[163]. In addition to acting as a topoisomerase poison, DOX also affects DNA through the hydroxyquinone moiety to form DOX-iron complex. This generates reactive oxygen species, which cause a breakage of the DNA chain[163]. Together, these mechanisms ensure that the replication of the DNA double-helix is halted – in turn, various metabolic and growth processes are disturbed, and, in time, apoptosis is induced.

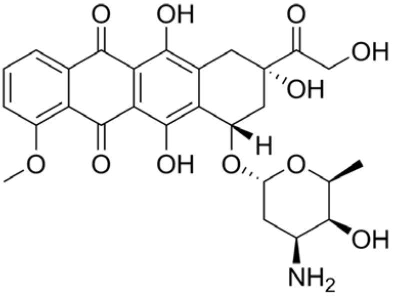


Figure 1.5. Structure of doxorubicin (DOX).

DOX possesses a positive net charge at the neutral pH[164] of the cellular environment. The drug is a chromophore and has a distinctive Raman spectrum, suggesting that label-free detection of DOX should be possible. Eliasson et al.[165] and Yan et al.[166] have reported spectral features of DOX measured over silver colloids. These bands are listed in **Table 1.1**.

Band	Bond
352 cm^{-1}	C-C-O in-plane deformation
450 cm^{-1}	C=O in-plane deformation
1200-1300 cm^{-1}	In-plane bending of C-O, C-O-H, C-H, respectively
1406 cm^{-1}	Ring vibrations
1433 cm^{-1}	Ring vibrations
1517 cm^{-1}	Ring vibrations
1640 cm^{-1}	H bonded C=O stretch

Table 1.1. Spectral features of DOX measured with silver colloids[165-166].

QUANTIFICATION AND DETECTION OF CYTOSTATIC DRUGS WITH RAMAN IMAGING

Lin et al.[167] have attempted to track the cytotoxic effect of DOX in single cells with Raman spectroscopy. Their results have suggested that no significant spectral changes have occurred in the proteins and the DNA of the cell. They concluded that on the whole Raman spectroscopic imaging does not offer any advantages compared to conventional cytotoxicity assays, unless only the cell nucleus is subjected to analysis. However, their study, as well as other studies[168-171], did not have as its aim the quantification of the drug itself; instead, the focus lay on the detection of the effects of the drug. A common observation in such studies is that the signal of DNA and proteins changes in cells treated with cytostatic drugs compared to control.

In contrast, the research dedicated to the *in situ* quantification of a drug in living cells is limited, although the interest in developing such methods exists[172]. Some examples include the research performed by Eliasson et al.[165, 173], Abbas[47], and Lorén et al.[83, 113, 143], which suggests that quantification of sub-millimolar concentrations of DOX is possible, both in blood plasma and intracellularly in lymphocytes. Another study has reported the successful intracellular quantification of the drug theraphthal in A549 cells[174]. A promising alternative is to extract the chemical components of the cells such as proteins, DNA, and RNA and quantify them separately, as was done by Draux et al.[127]. In other words, if it is not possible to detect the drug itself, it is possible to quantify its effect on the cellular distribution of biomolecules. Despite that possibility, the ability to quantify intracellular drugs for purposes of improved patient dosage would not be achieved with such a method, unless a correlation between dosage and DNA deterioration is firmly established. Furthermore, methods based on cellular extractions are too laborious and resource intensive to be readily applicable in a clinical setting.

PC12 CELLS

The PC12 cell line was first isolated from rats' adrenal medulla in 1976, and is distinguished by its synthesis and storage of the neurotransmitter DOP. The cells contain large dense-core vesicles containing DOP and other catecholamines[175]. PC12 cells are primarily used in neuroscience research. Here, the PC12 cells were used as a model for cells different from the blood cells to see if novel information would be revealed when applying machine learning methods for hyperspectral imaging and spectral interpretation.

DOPAMINE

DOP has two ionisable functional groups resulting in two pKa values. However, if pH is greater than 12, the molecule can exist as a zwitterion. This means that, depending on the pH of the DOP's environment, it may exist as a mixture of anionic, cationic, and zwitterionic forms[176]. At a physiological pH of 7, the pH of cellular environment, DOP exists primarily in its cationic form[177]. The reference spectra for cationic DOP measured at pH 5 can be seen in **Figure 1.6**.

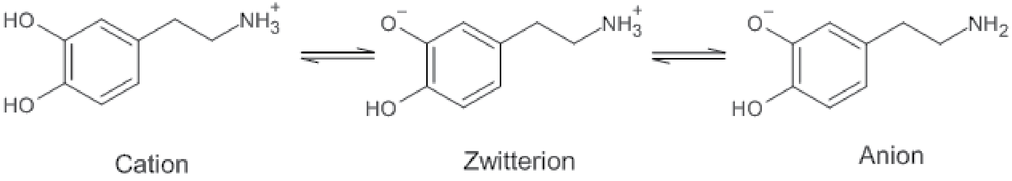


Figure 1.6. The ionic of dopamine (DOP)[178].

METHODS

WHITE BLOOD CELLS

MATERIALS AND STOCK SOLUTIONS

A stock solution of DOX 1 mM was prepared by the dilution of a 2mg/mL doxorubicin hydrochloride solution (Sigma Aldrich, Sweden) with MilliQ water. The stock solution was then diluted in order to achieve desired concentrations in each individual cell fraction.

Spherical gold colloids, AuNPs, 60 nm in diameter, were obtained from BBI Solutions (Sittingbourne, Kent, UK), with the following properties: a mean size of 57.0 - 63.0 nm, a mean diameter 59.6 nm, a gold chloride concentration of 1.26, the number of particles per mL was $2.60*10^{10}$, number of moles particle per mL $4.3175*10\text{-}^{14}$, the molar particle concentration (moles/L) was $4.3175*10\text{-}^{11}$,and the mass of gold per mL was $5.68*10^{-5}$.The colloidal coating, 4-mercapto benzonitrile (MBN), was purchased from SynChem OHG (Felsberg-Attenburg, Germany) and was dissolved in EtOH (90 %) to a concentration of 10 mM. The MBN stock solution was then further diluted with EtOH to concentration of 1 mM in order to be used for the coating of AuNPs. Dulbecco's phosphate buffer saline (PBS) with Ca^{2+} and Mg^{2+} (Fisher Scientific, Sweden) and 1 M HEPES (Fisher Scientific, Sweden) were used to prepare a washing buffer for by mixing 400 μL 1 M HEPES (Fisher Scientific, Sweden) with 39.6 mL PBS. Dulbecco's modified Eagle's medium, DMEM (Fisher Scientific, Sweden) combined with iron-fortified calf serum (Sigma Aldrich, Sweden) was used as cell incubation medium.

8-chambered Nunc™ LabTek™ coverglass (Thermo Scientific, Sweden) was used for cell cultures. A 10 mM solution of poly-L-lysine hydrobromide (PLL), 30-70 kDa (Sigma Aldrich, Sweden) dissolved in milliQ was used to coat LabTek™ chambers.

Boric acid (Sigma Aldrich, Sweden) and sodiumtetraborate-10-hydrate (Labassco, Mölndal, Sweden) were used to prepare a borate buffer with final pH of 8.2. This was done by first preparing a 0.2 M solution of boric acid and a 0.05 M solution of sodiumtetraborate-10-hydrate. Then, 50 mL of boric acid solution was mixed with 7.3 mL of sodiumtetraborate-10-hydrate, and finally diluted with MilliQ to a total volume of 200 mL. The resulting borate buffer was used to dissolve γ-irradiated lysed powder of poly-D-lysine hydrobromide (PDL), >300 kDa (Sigma Aldrich, Sweden), to give a final coating solution of 10 mM. The solution was subsequently used to coat LabTek™ chambers.

Trypan blue, TrB (Sigma Aldrich, Sweden), was used to prepare an aqueous (0.4%) cell stain for evaluation of cellular viability. The differentiation between living and dead cells is based on the fact that dead cells become permeable and hence absorb colour, becoming blue.

COATING OF GOLD COLLOIDS

AuNPs were mixed in individual Eppendorf tubes with 1 mM MBN in EtOH solution at a ratio of 1:1. The number of tubes corresponded to the number of chambers to be examined. The Eppendorf tubes were stored at 6 °C for ~48 h.

Before use, the Eppendorf tubes with the colloid solution were centrifuged for 5 minutes at 5000 rpm. The supernatant was aspirated from each tube and pellets were resuspended with 200 μL milliQ. The process of centrifuging and washing with MilliQ was repeated in total of 3 times.

Non-coated colloids were centrifuged for 5 min at 5000 rpm without any washing required.

ADHESIVE COATING OF THE COVERGLASS

LabTek™ coverglass with 8-chamberes was used. Each chamber was filled with 200 μL 10 mM PLL dissolved in MilliQ or 10 mM PDL dissolved in borate buffer and left to coat at RT for 30 minutes. Then, all wells were aspirated and washed two times with 200 μL MilliQ. The coverglass was then placed in a fume hood to air dry for at least 10 minutes or until cell seeding.

CELL PREPARATION AND INCUBATION

Different protocols were used depending on whether the cells were crypreserved or fresh.

Protocol 1: Cryopreserved cells

The cells were donated, prepared, and cryopreserved by Dr. Mats Bemark.

The cells were treated with one of the DOX concentrations (0, 11 or 40 mM) and were thawed in a water bath (~37.5 °C). The cells were thereafter diluted dropwise with 4 mL DMEM in a Falcon tube. The contents of the Falcon tube were then divided into Eppendorf tubes and centrifuged at 1900 rpm for 10 min. After centrifugation the supernatant was aspirated and each cell pellet was resuspended using 250 μL of calf serum and DMEM (1:9) mixture.

200 μL of the resuspended cell pellet was transferred to an individual chamber on a LabTek™ coverglass (**Figure 1.7**).

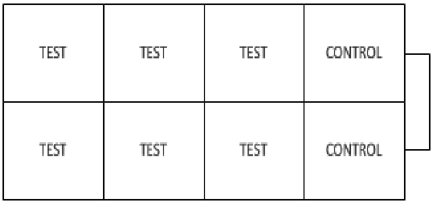


Figure 1.7. Coverglass distribution of one cell batch. All chambers contained cells treated with the same DOX concentration, 0, 11 or 40 mM.

A washed colloid pellet was added into each test chamber and the coverglass was put in a CO_2 incubator for ~16-24 h.

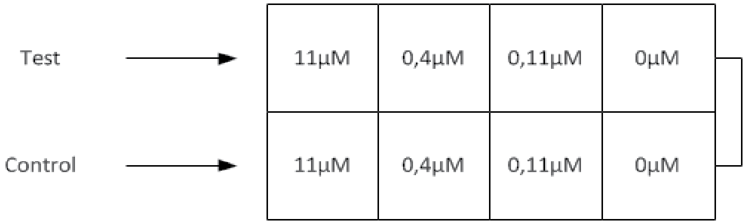
After the incubation, all chambers were aspirated and washed two times with a PBS+HEPES buffert mix. One of the control chambers was stained with 100 μL of aqueous solution of TrB (0.4 %), for 4 min to assess post-incubatory cell viability. The cells remained in the PBS+HEPES buffer throughout the measurements.

One test chamber was examined per day. Cells were selected visually under the microscope and as many cells as possible were measured during the day. After measurements, the test chamber was stained with TrB to assess of cell viability as described in the paragraph above. Test chambers that were not examined were aspirated and each were refilled with 200 μL of fresh calf serum and DMEM (1:9) mixture, and were put back into the incubator until the following day. The post-incubation steps were repeated for each chamber until all test chambers has been examined. On the final measurement day, the second control well was stained with TrB for a viability assessment.

Protocol 2: Fresh cells

Cells were donated and prepared by Dr. Mats Bemark.

Each cell fraction (V_T = 400 μL) was transferred to a Falcon tube and dropwise diluted with DMEM to a total volyme of 3 mL. The contents were shaken gently. 200 μL of cell suspension was transfered to coated coverglass chambers as follows (**Figure 1.8**):



Test	11µM	0,4µM	0,11µM	0µM
Control	11µM	0,4µM	0,11µM	0µM

Figure 1.8. Coverglass distribution for each fresh cell batch. Each chamber contained a cell fraction treated with one of the four DOX concentrations. The test row contained cells that were subjected to Raman measurements, while a control row functioned solely as a post-incubation viability test control.

A washed colloid pellet was added to each chamber. The coverglass was then put into CO_2 incubator. Incubation time varied from short-term (1-5 h) to long-term (16-24 h) depending on the purpose of the experiment.

All coverglass chambers were aspirated and washed twice with 200 µL PBS+HEPES after incubation. Prior to measurement, each chamber in the control row was stained with 100 µL TrB (4 %) for 4 min. The control chambers were then washed repeatedly with PBS+HEPES, and the cells were counted under a microscope.

The cells remained in the PBS+HEPES buffer throughout the measurements. Each chamber in the test row was examined under microscope in order to select suitable cells, for measurement with respect to their viability and type. Each measurement was done over a total of 10 cells per chamber, starting with the chamber with highest concentration of DOX. All test chambers were measured during the same day. After all measurements were completed the test row was stained with TrB as described earlier.

INSTRUMENTATION

Cell analysis was performed using a confocal Raman spectrometer (Dilor Labram INV) equipped with a confocal microscope (Olympus), holographic grating (900 grooves/mm), and a charged coupled device (CCD). The laser (He/Ne) excitation wavelength was 632.8 nm and the cells were examined under 100x magnification. The cells were mapped by the xy-stage of the microscope in 1 µm or 1.5 µm steps, and covered an area from 36 to 121 µm. Some of the measurements were mapped in several layers along the z-axis, with 0.5 to 2 µm between each layer. Optimal instrument parameters were selected based on pre-studies, when the maximum signal was sought. The optimal pinhole size was selected to be 900 µm and the optimal measuring time 2x1secs. No preference was given to either a 1000 or a 500 µm slit size.

DATA ANALYSIS

The spectral data was recorded with LabSpec 4.18. The spectral files were then converted through in-house software into .mat datasets and were subsequently examined in Matlab R2013a. Plots for each observation were examined visually to detect and remove observations containing fluorescence and/or interferences produced by cosmic rays. The base line correction, iteratively re-weighted penalized least squares, was done based on a script described by Zhang et al.[66], followed by ratio calculations between the DOX and MBN signals.

The datasets were imported into SIMCA 13.0.3 (Umetrics, Umeå, Sweden), where they were modelled by various multivariate techniques, including T-OPLS[50, 179] and OPLS-DA[180]. Generation, unfolding, and post-modelling reassembly of hyperspectral images was done by in-house software (DeterminatorIV 0.540) and routines from the Python Matplotlib. In the case of the T-OPLS models, the datasets were mean centred and was unit variance scaled where necessary. In the OPLS-DA models, the datasets were mean centred only.

PC12 CELLS

A DOX (Sigma Aldrich) standard stock solution of 1 mM was prepared in MilliQ water and stored at -20°C. An isotonic solution, pH 7.4, was prepared according to an earlier published protocol[181]. Series of dopamine concentrations (Sigma Aldrich) in isotonic solution were prepared by consecutive dilution of the stock solution of dopamine (10 mM). Aliquots of the dopamine stock were made, with pH 7.4 (in isotonic solution) and pH 5 (adjusted with 0.1M HCl), and stored at -20°C.

A stock solution for adenine (Sigma Aldrich), 10 mM, was prepared in 3 M HCl. Guanine (Sigma Aldrich) stock solution, 10 mM, was prepared in 5 M HCl. Cytosine (Sigma Aldrich) was dissolved in 1 M HCl to a final concentration of 10 mM, and uracil (Sigma Aldrich) was dissolved in 1 M NaOH to a final concentration of 10 mM. 60 nm gold nanoparticles, AuNPs, were purchased from BBI solutions (Cardiff, UK).

The internal standard 4-mercapto-benzonitril (MBN) (Synchem UG & Co. KG, Felsberg, Germany) was dissolved in 95% ethanol to a final concentration of 10 mM. Cell washing was prepared by mixing $CaCl_2$ and $MgCl_2$ free phosphare buffer saline, PBS (Sigma Aldrich), with a 1 M solution of HEPES (Sigma Aldrich), to a final HEPES concentration of 10 mM.

Preparation of AuNPs

AuNPs were mixed with 0.5 mM of MBN solution to a final MBN concentration of 100 μM. The mixture was stored for 24 hours at 4 °C, in order to allow the MBN monolayer to stabilize. Prior to usage, the colloid mixture was washed according to procedures described in Josefson et al.[182].

Preparation and Incubation of Cells

PC12 cells from the adrenal glands of *Rattus norvegicus* were purchased from the American Type Culture Collection (Manassas, VA). Cells were grown on mouse collagen IV coated flasks (25 cm^2) in a phenol red-free RPMI-1640 culture medium (Fischer Scientific, Sweden) supplemented with 10% horse serum (Sigma Aldrich, Sweden) and 5% fetal bovine serum (Sigma Aldrich, Sweden) in an incubator at 37°C in 7% CO_2 and at 100% humidity. The growth medium was replaced every 1-2 days during the lifetime of all cultures. Cells were split every 7 days. The cell suspension was mixed in a vortex shaker and 1 mL of the cell solution was mixed with 7 mL of cell growth media and transferred to a clean 25 cm^2 culture flask. The culture was subcultured 10 times before the cells were used for measurements.

Prior to AuNPs incubation, an 8-well sticky slide (LRI AB/Ibidi, Lund, Sweden) was attached to a Raman grade CaF_2 microscope slide (Crystran Ltd., Poole, UK). The cell suspension was then seeded into five wells. For one well, uncoated AuNPs were added to the cell suspension in a 1:1 volume ratio. Four more wells were prepared by adding washed MBN functionalized AuNPs in a 1:1 volume ratio. The slide was then put into an incubator with 5% CO_2 and at a temperature of 37.1 °C for 1 h. Thereafter, DOX was added to wells containing MBN-colloids to reach final concentrations of 100, 50, and 1 μM of DOX. To one well no DOX was added. Cells were incubated again for 30 minutes.

All wells were aspirated and washed twice with a PBS+HEPES solution, then refilled with 400 μL PBS+HEPES solution prior to measurement.

PC12 Measurements

Spectra were collected with a Dilor Labram IV Raman spectrometer equipped with an inverted confocal microscope (Olympus) and a piezoelectric stage. The excitation wavelength was 632.8 nm and the cells were examined under 100x magnification. Spectra were collected within the spectral range 370-2270 cm^{-1}. A charge coupled device (CCD) was utilized as detector. Measurements were performed by mapping a 12x12 μm pixel grid in xy-plane over one cell at a time, measuring a total of 8 cells per well. To ensure measurement inside the cells, the laser was first focused on the surface of the cell and then the focus was adjusted to a 2 μm depth. In each pixel, the spectrum was measured twice and the spectral acquisition time was 2 s. The morphology of the cells was evaluated visually before and after in order to assess their viability.

MEASUREMENT OF REFERENCE SOLUTIONS

Reference solutions of DOX (100, 50, and 1 μM), dopamine (0.1, 1.5, 5, 7.5 and 10 mM), adenine (5 mM), cytosine (5 mM), guanine (5 mM), and uracil (5 mM) were obtained by mixing stock solutions with non-coated AuNPs on a separate CaF_2 slide with an attached 8-well sticky slide. MBN (0.5 mM) was measured as is, without any additional non-coated colloids. Reference spectra were obtained by using an average of 10 spectra, accumulated for 30 s each.

DATA ANALYSIS

SIMCA 14.1 (MKS Data Analytical Solutions, Malmö, Sweden) was used for the construction of PCA and T-OPLS models. Generation, unfolding, and post-modelling reassembly of hyperspectral images was performed with in-house software (DeterminatorIV 0.54) and with routines from the Python Matplotlib[183]. Interpolation of spectra was performed with an in-house Python script. After baseline correction with RCF[63], an in-house algorithm for peak finding[184] based on finite state automaton with inputs from Savitzky-Golay first and second derivative filters were applied. The baseline between the peaks was then subtracted. These corrections were done to remove the influence of noise at the baseline and thus to remove its influence on the subsequent multivariate models. In addition, the interpretability of spectra was improved.

RESULTS AND DISCUSSION

A crucial objective of this chapter was to find a way of ensuring colloidal entrance into living cells, which is a prerequisite to the project's main goal – the detection and quantification of intracellular DOX. Being able to develop methods capable of detecting and quantifying anti-neoplastic drugs inside living blood cells is of analytical interest, because they would enable easy tracking of concentration variations of a drug at the location where it is therapeutically active in the patient. This would in turn reveal potential under- or over dosages of the patient and further individualise the treatment strategy used for them.

Lymphocytes were the first choice selected to represent ALL. Lymphocytes were chosen because the precursors of lymphocytes, lymphoblasts, are the central cell type in the occurrence of ALL. However, as it will be shown later in this chapter, lymphocytes had to be replaced with other cell models. The cell models studied apart from lymphocytes were neutrophils and monocytes. The cells used were either fresh or cryopreserved. The rationale for using cryopreservation is to halt biological processes, a common practice in clinical applications, where cells and tissues sampled for analysis are often cryopreserved[185]. Furthermore, in a clinical setting it would be impractical to need to analyse fresh cells – it would mean that the peripheral blood sample had to be taken from the patient and then analysed within a frame of 24 hours. Therefore, there is a need of protocols for the development of cryopreservation that would still maintain good cell viability, and allow prolonged storage. That process would allow for long-term storage and would facilitate sampling strategies that do not require fresh blood, making the method more clinically suitable.

Measuring living cells was considered to be crucial since dead cells become permeable, which means, for instance, that colloids and residual DOX could accumulate in a dead cell. Therefore, dead cells were not considered representative for the purposes of this chapter, which deals with identification and quantification of DOX. The differentiation between dead and living cells was done through q visual assessment of their vitality, and through staining with trypan blue (TrB).

In earlier experiments, the cells were treated with DOX dosages of 0, 11 or 40 mM. However, in the later experiments with granulocytes and monocytes, lower concentrations were chosen (0, 0.11, 0.4 and 11 μM) in order to imitate plasma concentrations of DOX in ALL paediatric patients. The concentrations were calculated based on the median plasma concentration found in children treated for ALL, 62.8 ng/mL[111]. As the experiments with neutrophils and monocytes did not reveal the presence of DOX, the dosage range was increased to 1, 5, and 100 μM for PC12 tests.

Proof of colloidal internalisation and of the presence of DOX was achieved by microscope experimental observation and by subsequent analysis of acquired spectral data by multivariate linear projection algorithms. During the analysis, the T-OPLS models were constructed in such way as to detect spectral features of the analytes, DOX and DOP, and the IS, MBN (**Figure 1.9**). The presence of spectral features of the latter inside the cells served as evidence of AuNPs' presence inside the cellular environment.

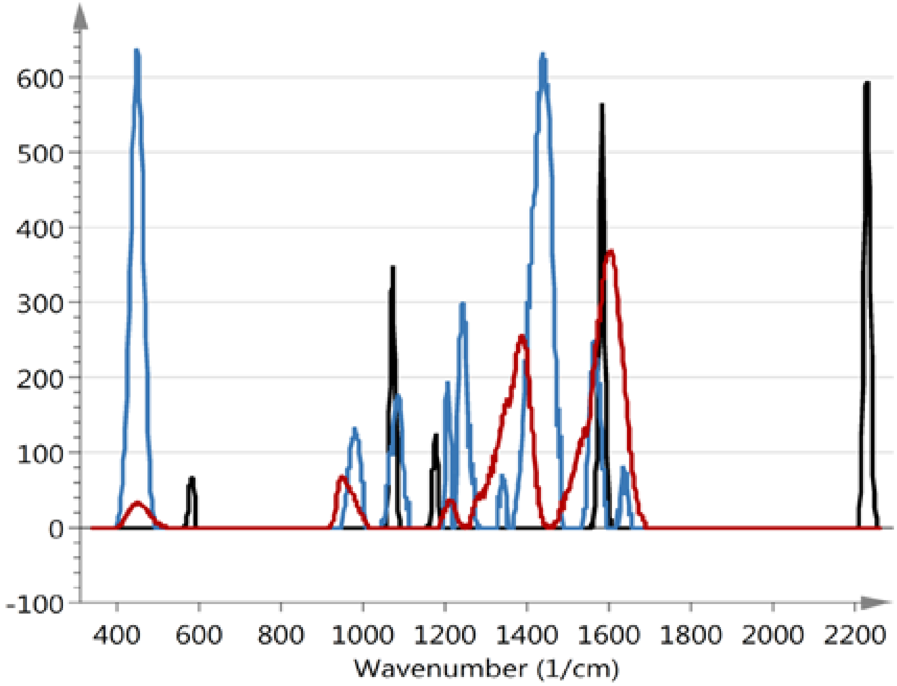


Figure 1.9. The reference spectra of MBN-coated AuNPs (black), 100 μM DOX and uncoated AuNPs (blue), and 10 mM cationic DOP and uncoated AuNPs (red). The spectra were pre-processed with RCF. The baseline between peaks was subtracted by using a peak finding finite automaton based on Savitzky-Golay derivative filters, leaving medium to strong peaks.

During the course of the project several challenges arose that had to be solved in order to detect DOX. In addition to difficulties in the selection of cells capable of colloidal internalisation, we also faced difficulties when adhering cells to the coverglass surface and in the cryopreservation of cells.

CELL ADHESION

One of the first issues to solve was poor cellular adhesion to the surface of the coverglass. The evaluation of cell adhesion was performed by visual examination of cell motility under microscope magnification. Cells need to be as immobile as possible for SERS measurement, otherwise they would move out of the laser focal point thus rendering the measurement useless. This was a particular problem when studying lymphocytes, which lack the desired capabilities of adhesion to surfaces and tissues.

Initially, a MilliQ water solution of poly-L-lysine, PLL (30-70 kDa), was used but proved to be unsuccessful as an adhesive coating for lymphocyte immobilisation. To promote adhesion PLL was switched to poly-D-lysine, PDL (> 300 kDa), based on the assumption that the *dextro* variant of the polymer would not be subjected to degradation by cellular metabolism. The higher molecular weight of the protein was supposed to increase adhesion further[186]. Adhesion was still failing.

It was later realised that the pH of the PDL solution can influence the adhesion quality of the cells. The solutions were previously made in MilliQ water and had acidic pH of ~3-4. The pH of the PDL solution was therefore increased towards the alkaline range, ~8, by dissolving PDL in borate buffer instead of in MilliQ. This improved the adhesion of lymphocytes. However, when coated or uncoated AuNPs were added to lymphocytes the adhesion was failing again, suggesting that the presence of colloids somehow disrupted the fastening of the cells to the PDL treated glass surface. Interestingly, when trypan blue (TrB) was added during viability tests, adhesion was again successful. The mechanism for this phenomenon is unknown and no description of it was found in the literature. Unfortunately, TrB is a chromophore with a high fluorescence and therefore adhesion promoted by it was not feasible for the purposes of measuring cellular spectra and presence of DOX by SERS.

Other cell types that were used, neutrophils and monocytes, did not display the same adhesion problems as lymphocytes. The only factor that increased motility was prolonged incubation times. Therefore short-term incubation times (< 5 h) were deemed optimal for neutrophils and monocytes.

ANALYTE DETECTION IN CRYOPRESERVED LYMPHOCYTES

The blood samples used in the experiments with cryopreserved lymphocytes contained both monocytes and lymphocytes. Lymphocytes were selected by their visual characteristics during each measurement, differentiating them thus from monocytes. The selection was made through an evaluation of motility and size. Lymphocytes, which did not adhere well, displayed mainly Brownian-type motion. Lymphocytes are also smaller with a size range of roughly 5-8 μm in diameter. Monocytes on the other hand displayed an amoeba-like motility and were larger, 8-10 μm.

To build an OPLS regression model, the common way is to have more than one Raman spectrum with known concentrations to span a calibration range. However, when analysing images of biological samples introducing an analyte with known concentrations into cells is difficult. To overcome this, using T-OPLS similarly to a one-point standard is possible. The experimental observation and T-OPLS models have shown that lymphocytes do not internalise colloids (**Figure 1.10**)

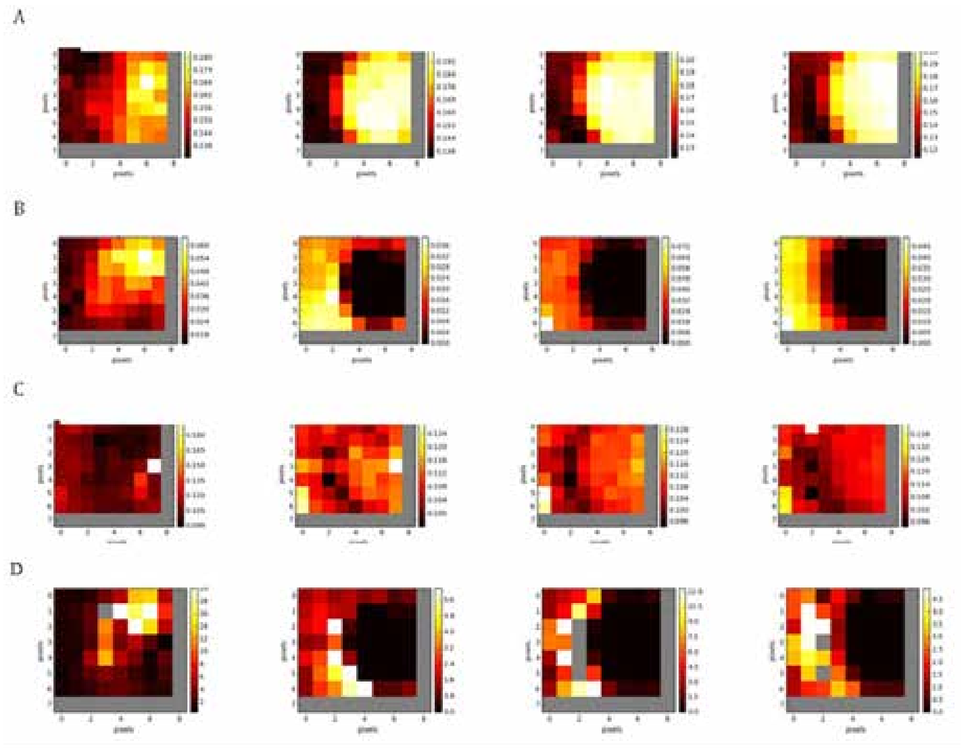


Figure 1.10. Hyperspectral image of one lymphocyte treated with 40 mM DOX and in MBN-coated AuNPs. The cell was mapped in 4 layers, 0.8 μm apart, with the top layer shown to far the right in each row. Image a) is the total intensity of the signal for the lymphocyte, where the brightest parts represent the cell derived from **po3** in a T-OPLS model. Image b) is the distribution of DOX, c) the distribution of MBN signal from AuNPs, and d) the ratio between DOX and MBN signals. The grey pixels in the images are pixels that were removed because they were outside the measured area. In case of row D, the grey pixels mark locations with a negative concentration when the ratio between DOX and MBN was calculated, and these were therefore removed. The images were based on spectra T-OPLS model calculated on full spectra.

Figures 1.10 shows a lymphocyte mapped in four layers with signal intensities. The left-most images in each row represent the bottom of the cell (where the cell touches the coverglass) and the right-most images represent the top of the cell. A T-OPLS model was created using MBN and DOX reference spectra as Y matrices. Separate T-OPLS models were created for each Y. The model was calculated on four cells (only one cell is shown here) and the diagnostics for the model gave cumulative values of $R^2X = 0.996$, $R^2Y=0.959$, and $Q^2=0.955$.

The first set of images in **Figure 1.10A** shows the total signal intensity from a lymphocyte. The brightest parts of the image represent the cellular outline, showing the location of the cell in the mapped window frame. The image furthest to the left in **Figure 1.10A** is the layer measured closest to the coverglass surface, -1.2 μm in z-axis, and is representative of the cell surface.

The second set of images (**Figure 1.10B**) represents the distribution of DOX detected during measurements, with brighter pixels indicating a higher presence of DOX, and darker or black pixels point to either very low DOX enhancement or to its absence. As it becomes evident from looking at **Figure 1.10B**, there was no DOX related signal issuing from the cell. When compared to **Figure 1.10A**; the comparison reveals that despite the presence of a cell within the image frame, there was no overlap of DOX signal in the same pixels as in **Figure 1.10A**. The exception was the layer to the far left, the top surface of the cell, where there seemed to be some DOX present. This can be explained either by smaller amounts of DOX leaking into the cell medium from dead cells or by left over DOX which was not washed away. This extracellular DOX possibly got enhanced and detected because of the presence of AuNPs on the cell's exterior.

A similar situation can be seen in **Figure 1.10C**: although the absence of a MBN signal is not as apparent, the images are inconclusive on whether a MBN signal is present within the cell. Experimental data has, however, led to the assumption that the colloids have not been internalised. In the course of experiments, no MBN signal was detected in the pixels corresponding to the mapped cell. The last four images in **Figure 1.10D** represent the ratio of the DOX and MBN signals. The ratio of DOX and MBN was calculated in order to compensate for signal enhancement from colloidal aggregates, which produce much stronger signals compared to isolated AuNPs. If DOX enhancement was present, it would mean that the colloids had internalised and were able to produce that enhancement. With colloidal presence inside a cell, it should therefore have been possible to detect the MBN signal, since colloids are coated with that substance. In other words, where there was a DOX signal, there also must have been a MBN signal present. However, by calculating the relative enhancement ratio of DOX and MBN signals gave hyperspectral images with no signal present. The exceptions are the image to the far right (bottom of the cell) and the image to the far left (top of the cell) in **Figure 1.10D**, where there seems to be a signal enhancement, and the presence of DOX outside the cell. The behaviour observed in **Figure 1.10** is representative of all lymphocytes that were modelled in this particular data set, as well as in other data sets containing lymphocyte measurements[•].

At first, the result was puzzling, since Eliasson et al.[173] have reported a successful acquisition of rhodamine 6G spectra in 30-40 % of all lymphocytes measured: rhodamine 6G has a similar entry pathway to DOX[57]. Similarly, Abbas[47] has reported the detection of 11 μM DOX inside lymphocytes. Both cases suggest successful colloidal internalisation. Upon closer scrutiny of their work, however, it was understood that, just as in results here, the measurements were done on a mixed sample of lymphocytes and monocytes, but no differentiation was made between the two cell types. Hence, it is assumed that their tests were performed indiscriminately on both cell types.

DOX could thus not be measured in lymphocytes due to an absence of surface enhancement provided by internalised colloids. In turn, it was concluded that lymphocytes are not capable of internalising colloids. This can be attributed to the inherently low phagocytic capacity of lymphocytes[187], although the phagocytic activity is not entirely absent. The detection of DOX achieved by Eliasson et al.[173] and Abbas[47] can be explained by successful AuNP internalisation by monocytes, which possess more active endocytosis.

ANALYTE DETECTION IN CRYOPRESERVED GRANULOCYTES

The granulocyte fraction consists of neutrophils, basophils and eosinophils. Neutrophils comprised the largest percentage (ca 90-95 %) in samples, and hence no visual differentiation was made between the cell types. Neutrophils were chosen over previously used lymphocyte/monocyte fractions due to their higher phagocytic capacity. Despite that, cells that qualified as viable produced puzzling spectra, and just like lymphocytes, did not reveal any spectral features of either cell content or DOX and MBN.

This led to unreliable measurements and produced inconsistent results upon determination of the presence of colloids and DOX inside the cellular matrix. T-OPLS models were constructed to analyse the data further.

[•] Over the course of this project, a total of 193 lymphocytes, 201 neutrophils, and 500 monocytes were measured and analysed with linear projection algorithms.

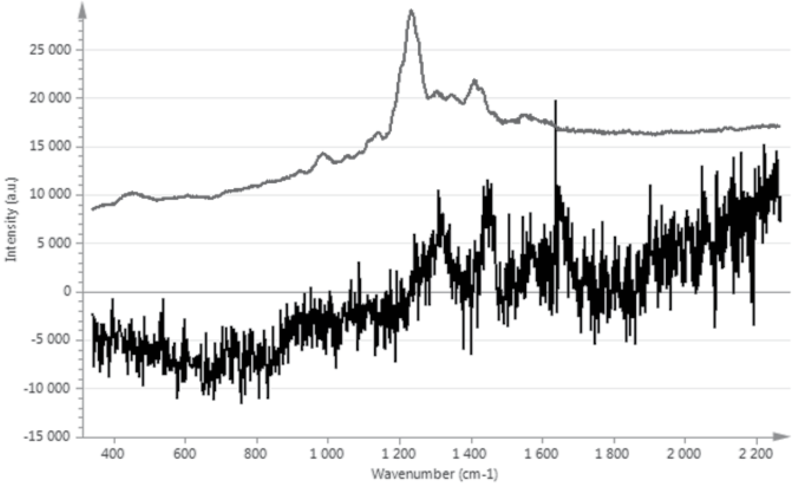


Figure 1.11. An example of a living granulocyte treated with 11 μM DOX with presence of MBN-coated AuNPs. The reference spectra of DOX (blue, top spectrum) was plotted together with T-OPLS score vector **t1** (black, lower spectrum).

Figure 1.11 shows that there was no or weak correspondence between the score vector **t** and the set Y variable, i.e. the DOX reference. A cell that had internalised AuNPs was expected to display a signal enhancement of the cell interior, including DOX. In such a case, the score vector **t** would have greater spectral resemblance to the DOX reference spectra. In other words, the cells that were deemed to be alive by visual assessment of their motility and morphology did not display signal enhancement of either DOX or the cellular matrix. This led to the conclusion that the few successful measurements were done on dead cells; because when cells die their membranes become permeable, thus letting colloids in.

These results were suspected to be due to the failing phagocytic activity of neutrophils as a consequence of cryopreservation. The protocol was therefore changed and the cryopreservation step removed. Protocol 1 therefore needed extensive revision in order for it to fit the handling of fresh cells. The greatest change lay in addressing the inconvenience of the centrifuging step. The generated cell pellets from fresh fractions were too small resulting in too few cells per test chamber. Another change was based on the assumption that starving phagocytic cells through the omission of calf serum would improve the colloid internalisation[154, 187]. Lastly, the incubation time was significantly reduced, to 5 hours or shorter, to minimize potential cytotoxicity induced by AuNPs; it has been shown that the internalisation for this cell type occurs already after 30 minutes[157]. The revisions resulted in Protocol 2.

ANALYTE DETECTION IN FRESH NEUTROPHILS

The viability of granulocytes was improved by the omission of cryopreservation and by omitting thawing steps, however, the internalisation of the colloids was still failing and similar inconsistencies were displayed in models and spectra as has had occurred with cryopreserved neutrophils.

It was shown by Bratneck et al.[157] that neutrophils display phagocytosis only towards particles coated with opsonins, molecules which promote phagocytosis by cells. Non-opsionised particles are not internalised and are instead trapped outside the cell by neutrophil extracellular networks (NETs). This would explain to the inconsistent results with both cryopreserved and fresh granulocyte fractions. Neutrophiles were therefore abandoned as a cell model for the detection of DOX.

ANALYTE DETECTION IN FRESH MONOCYTES

Bratneck et al.[157] reported that gold nanoparticles, coated and uncoated, were internalised by monocytes. Hence, lymphocyte/monocyte fractions were reinstated as cell models for DOX detection experiments. The cell preparation Protocol 2 was used as described earlier.

Monocytes were selected for measurements by visual characteristics as described above. Through experimental observations it was concluded that the cells had successfully internalised colloids. The success can be in part be linked to work reported by Eliasson et al.[173] and Abbas[47] – the successful internalisation experiments in their work could have been performed on monocytes, not lymphocytes. Monocytes were subjected to both long-term and short-term incubation. Both proved successful, but, in case of the former, the adhesion to the coverglass was reduced. Short-term incubation showed lower motility and also showed internalisation after as soon as 3 hours. The incubation time was further decreased in later experiments to first 2 h and then 1 h. These results are in line with those reported by Bratneck et al.[157], who observed colloidal internalisation after 30 minutes.

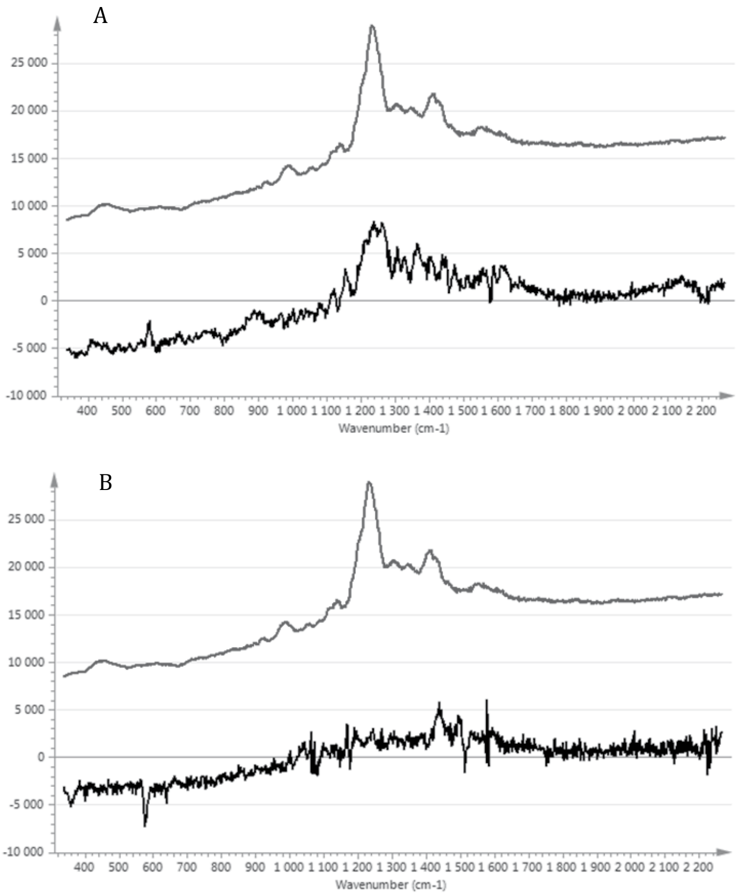


Figure 1.12. DOX reference spectrum (blue, top spectrum) compared to score vectors **t** (black, bottom spectrum)from T-OPLS models for (A) a single monocyte treated with 11 μM DOX and incubated with uncoated AuNPs, and (B) a single monocyte treated with 11 μM DOX and incubated with MBN-coated AuNPs.

Unfortunately, multivariate analysis proved tricky in terms of DOX quantification. The T-OPLS models that were done with monocytes incubated with uncoated AuNPs pointed to a detectable presence of DOX, as it is evident from **Figure 1.12A**.

Although far from optimal, some of the component's peaks could be correlated to the DOX reference spectrum (**Figure 1.12A**). The situation was however worse when it came to the T-OPLS modeling of monocytes incubated with MBN-coated AuNPs (**Figure 1.12B**), where there was no observable correlation between the principal component and DOX reference spectra. These results suggest that it was not possible to detect, and therefore quantify, DOX in the presence of MBN-coated AuNPs with this particular protocol.

There was therefore a discrepancy between experimentally acquired information and the T-OPLS models. Although models that were calculated suggested that both DOX and colloids were internalised by monocytes, it was impossible to discern a clear DOX component from the rest of the cellular noise if the colloids were coated with MBN. There could be several explanations for this.

Firstly, the difference in signal intensity between MBN and DOX signals could be too high, making it hard to detect in images when calculating the DOX/MBN ratio. Another scenario could be that DOX did not reach sufficient proximity to the surface of AuNPs to be able to produce signal enhancement. This could be attributed to the fact that AuNPs coated with MBN have a neutral net charge, and hence no attraction between DOX and the surface of AuNPs occurs or is limited. It is more plausible that attraction is limited rather than non-existent, since SAMs usually do not coat a colloidal surface fully[146]. It is also a possibility that the reference spectrum for DOX used to calculate T-OPLS models was not of sufficient quality. On the other hand, a T-OPLS analysis was performed with a DOX reference spectrum generated by Abbas[47] (results not shown) without any improvement. Further experimental work would be required to address the issues of DOX detection. Yet another possibility would be the failed uptake of the nanoparticles; like neutrophils, monocytes prefer the internalisation of opsonized matter, a process that may have been disrupted by MBN coating and/or absence of calf serum in the cell medium. Lastly, the acquired results could indicate a flaw in the experimental procedure, which was identified in literature that was published three years after the experiments performed here. It was reported there that DOX (50 μM) was not observable in the cytosol until after 48 h[188]. Our incubations with the drug lasted between 1 h and 24 h in the experiments discussed above.

DETECTION OF DOX AND DOPAMINE IN PC12 CELLS

As DOP is produced in abundance by PC12 cells and has distinctive peaks, PC12 cells were chosen in order to evaluate T-OPLS performance further. The T-OPLS models were built on a data set containing 24 cells treated with DOX at different concentrations (1, 50, and 100 μM) and MBN-AuNPs, 8 cell treated only with MBN-AuNPs, and 7 blank cells treated with uncoated AuNPs. Therefore, DOP signals were expected to be in both blanks and samples, whilst MBN and DOX were expected to be modelled in samples only. Each sample cell was mapped with 144 pixels, and each blank cell was mapped with ca 149 pixels. In case of an optimal model, the predicted score must resemble the reference spectra of the analyte or IS of interest.

Hyperspectral images revealed that T-OPLS was prone to false positives. The score vectors isolated MBN peaks in blanks, despite the blanks not having any MBN present. Thus, the pattern in Y was 'coaxed' out of spectra whether they contained the substance represented by Y or not. To illustrate this, images of the DOX/MBN and DOP/MBN ratios were made using the full spectral range (**Figure 1.13**). By dividing with MBN reference spectrum, the differences in enhancement should have been compensated, showing the distribution of DOX and DOP.

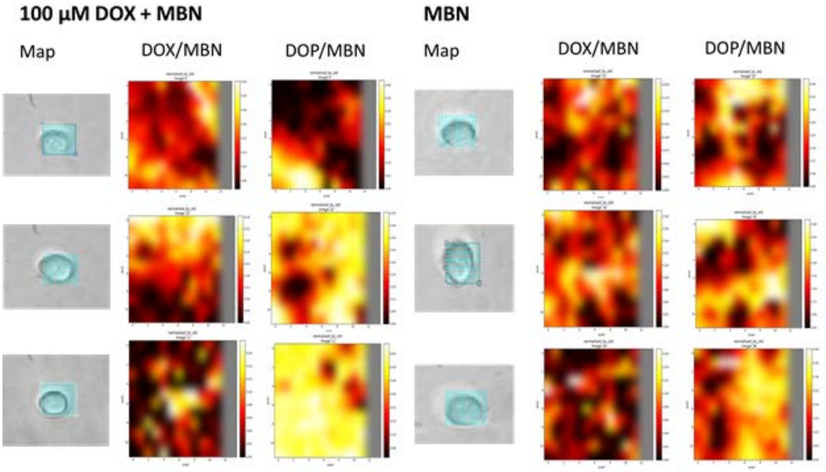


Figure 1.13. MBN corrected hyperspectral images showing the distribution of DOX and DOP. To the left are images of three PC12 cells treated with 100 µM DOX and MBN-AuNPs. The first column shows the bright-field image of the cell along with the mapping grid. The second column is the distribution of DOX after MBN correction, and the third column is the distribution of DOP after MBN correction. To the right are three PC12 cells treated only with MBN-AuNPs. The first column shows the bright-field image of the cell along with the mapping grid. Second column is the distribution of DOX after MBN correction, and the third column is the distribution of DOP after MBN correction. All images were mapped using the full spectra.

Both sets of images in **Figure 1.13** were calculated based on loadings from three T-OPLS models, where Y = MBN, Y = DOX, and Y = DOP. The loadings from those models were then used to arrive at the distribution images above. All the images are of comparable intensity range (ca 0.2-0.6), with the exception of the top-most cell treated with 100 M DOX and MBN-AuNPs, which had a maxima at 2.64 (DOX/MBN) and 0.4 (DOP/MBN). Where MBN was present, the images have no or low intensity (black areas) – the presence of black areas within an image frame shows that colloidal internalisation was indeed successful in that cell.

Comparing distribution of DOP and DOX in DOX treated cells revealed that both substances have been modelled and that they occupy overlapping, but not identical, regions of the cells. These results suggested that DOX and DOP could indeed be separated and that enhancement correction with IS worked. We expected that the distribution of DOX would have been more uniform in one region of the cell, as DOX accumulates predominantly in cell nucleus, while DOP would have been more evenly distributed throughout the cell. However, there was no consistent behaviour between cells in terms of DOX and DOP distribution. On examination of the images of PC12 cells, which were treated only with MBN-AuNPs, it was realised that the DOX signal was modelled, although there ought to have not been a DOX signal in those cells. Since the distribution of the alleged DOX follows similar, although not identical pattern to that of DOP (compare the images DOX/MBN and DOP/MBN), an issue of false positives was suspected. False positives could also partially explain what was observed in DOX-treated cells, where there was no consistent distribution pattern between

DOX and DOP. The occurrence of false positives could have been caused by overlapping spectral regions with broad peaks, since DOX, DOP, and cellular matrix have broad and/or weak vibrations in the same region of the spectra. Further, it was observed that MBN tended to be modelled in cells that were blank, and should not have had any MBN present (results not shown).

To verify the source of the false positives, the author performed a spectral interpretation of the scores generated by nine T-OPLS models. The first shortcoming became apparent when the background correction methods used were scrutinised. To achieve better DOX resolution with background corrections, RCF was used as it offered on the whole better peak resolution, especially for the small peaks originating from the cell. However, some of the DOX signals could not be observed. The corrections worked well for peaks over a certain SNR, but for peaks with low SNR the chosen background corrections failed to distinguish between low intensity peaks and noise. As such, better cell response peaks were gained, but DOX signal, which was potentially low, cannot be seen as clearly when hyperspectral images were constructed. These finding suggest that, below a certain SNR, performing RCF in combination with T-OPLS and other baseline correction methods used here (see **Methods** section) was infeasible.

The second part of the interpretation of T-OPLS scores looked at which peaks could have caused the occurrence of false positives. The results of this examination were compiled in **Table 1.2**: note that the list of peaks is not exhaustive. The table shows cases of **overlapping signals** and of **covariation**, two terms that ought not to be confused with each other. The former means that the signals of two or more substances appear at the same wavelength, but do not covary with each other in any way. On the other hand, covariation refers to two or more species that have signals that are dependent on each other. If the signal of one compound changes, so does the signal of the covarying compound. For instance, covariation can be caused by molecular interactions while overlaps can stem from molecules having the same or similar vibrations and thus they give rise to signals at the same wavenumber. Næs et al.[32] mentioned covariation as a potential pitfall encountered in multivariate modelling under the term **collinearity** (see **Introduction** to Part I). The reader ought to be aware that the terms covariation[58] and collinearity[32] are not entirely appropriate in the context of this research as they imply linear correlation. With T-OPLS, it was not possible to ascertain whether the dependency between signals was indeed linear. Hence, the author proposes to use the terms **linear** and **nonlinear co-dependencies**, instead, the presence of which, along with overlaps, contributed to the false positive behaviour observed.

The presence of co-dependencies creates certain implications for PLS-based modelling, as PLS is capable of discriminating between overlapping signals, but performs worse when separating co-dependent signals, especially if they are caused by nonlinearities. This was clearly true for T-OPLS as well shedding light on the appearance of false positives. Using a co-dependent signal as an example, consider, for instance, the peak at ~1074 cm^{-1} in **Table 1.2**. It was one of peaks present in the MBN reference spectrum that was used as Y for T-OPLS modelling. This peak appeared in the predictive score vectors **t** for all models where Y = MBN spectrum. However, the peak also appeared in the first orthogonal score vectors **to1**, indicating that this peak varied as a consequence of a substance other than MBN. When models where Y = DOP were examined, the peak appeared only in **to1** vectors, suggesting therefore that DOP did not cause the covariation observed in Y = MBN models. For models where Y = DOX on the other hand the peak can be seen in both **t** and **to1** vectors. However, since the model where Y=DOX calculated on blank cells (DOX b in **Table 1.2**) did have this peak in **t**, it could not have originated from DOX, since blank cells did not have DOX. Therefore, the covariation with the MBN signal at 1074 cm^{-1} stems from the cell matrix. Similar co-dependent behaviour was observed for numerous peaks (**Table 1.2**).

Nonetheless, there were several peaks that T-OPLS had identified as dependent on one substance only. For a peak to be attributed to a single chemical species with certainty, certain criteria had to be met. The MBN peak at 365 cm^{-1}, for instance, was present in **t** vectors where Y = MBN, meaning that the response at that wavenumber was correlated to the spectrum of MBN. The absence of that peak in **to1** shows that at 365 cm^{-1} there was no other correlation apart from MBN. In all other models where Y was not MBN, the peak was consistently present in **to1** vectors in models containing samples. If the variation of that peak was dependent on DOX or DOP, then it would also be present in the **t** vectors of models where Y was either DOX or DOP. The peak could also have originated from the cell matrix; if it was the case, the peak would show up in **to1** of MBN models and be present in the **t** and/or **to1** vectors of models performed on blanks. It is noteworthy that this peak was absent from the reference spectrum used for MBN as a Y matrix, although earlier experimentation and literature[148-150] indicate that this peak should have been present in an MBN spectrum. Its absence was likely to be due to an accidental removal by the baseline correction method used here. This indicates that although the Y did not contain this wavenumber, the remaining peaks of MBN present in the Y could have a linear co-dependency with the absent peak. T-OPLS was, in other words, able to model peaks that were missing, but which ought to have been there.

For MBN, the peaks that were identified as unique were 365, 549, 583, 706-708, 773-779, 1178, 1199, and 2228 cm^{-1}. Observe that some of those peaks were not present in the reference spectrum of MBN, and were likely lost in a background correction similar to 365 cm^{-1}. Similarly, DOP also had independent peaks at 947-951, 1211-1217, and 1590-1595 cm^{-1}. The reader has undoubtedly noted that many of the peaks were given as ranges; this variation in peak position was concluded to be due to spectral shifts caused by matrix effects. For DOX, however, no peaks could be identified as free from co-dependency or as an overlap with other species present in the matrix. This provided another reason as to why DOX was not clearly seen in hyperspectral images, and hence not possible to quantify. Naturally, this could also have been caused by the fact that no DOX was accumulated in the cells due to too short incubation times (see earlier discussion in **Analyte detection in fresh monocytes**). If that was indeed the cause behind the lack of a clear DOX signal, the images presented in **Figure 1.13** can only be interpreted as representing false positive DOX signal.

Table 1.2. Comparison between the reference spectra of MBN, DOP, DOX, and principal (**t**) and first orthogonal (**to1**) score vectors from T-OPLS models. A total of nine T-OPLS models were calculated where Y = MBN reference spectrum (MBN s+b, MBN s, and MBN b), Y = DOP reference spectrum (DOP s+b, DOP s, DOP b), and Y = DOX reference spectrum (DOX s+b, DOX s, DOX b), where s+b stands for models containing both samples (s) and blanks (b), only samples (s), or only blanks (b). Blanks were cells incubated with uncoated AuNPs, while samples were cells treated with DOX and MBN-AuNPs. 'Cell' refers to signals from the cellular matrix. Abbreviations used in the table: n = negative score peak; n. a. = not assigned; + = cation; ± = zwitterion; δ = in-plane deformation; ν = stretching; φ= bending; γ = out-of-plane deformation. If a peak was negative (n) in **to1** vector while also present in **t** vector, it is assumed to be the same as if it was absent. The table continues on the next page.

Reference			**MBN s+b**		**MBN s**	**MBN b**		**DOP s+b**		**DOP s**		**DOP b**		**DOX s+b**		**DOX s**		**DOX b**		**Tentative assignment**
MBN	**DOP**	**DOX**	**t**	**to1**	**t**	**t**	**to1**	**t**	**to1**	**t**	**to1**	**t**	**to1**	**t**	**to1**	**t**	**to1**	**t**	**to1**	
			365		365				365		365				365		365			MBN (δC-N/γ C-N)[149]
	450	449	447		447				447		447				445	447	445			MBN and DOX[165, 189]
														458				458		Possible DOP+ [176]
			549		549				549		549						549			MBN (n.a.)[149]
583			583	583 n	583				583		583				583		583			MBN (n.a.)[149]
			708		706				706		706				708		706			MBN (γ aromatic)[149]
			778		778				779		773				779		778			MBN (γ C-H)[149]
			821	825	821		821				823		824		822		820		824	MBN (γ C-H)[149], cell covariance
	947			951			950	949		951		950								DOP+ [176]
		981												987		984		987		Possible DOX (ν aromatic)[190], covariance with cell
					999		999			999	997				999					Cell, co-varies with MBN (n.a.) and DOP (n.a.)
				1001					1001 n				1003						1001	Cell
			1047		1039				1047		1044				1047		1045			MBN (n.a)
1074			1074	1074	1074	1074	1076		1074		1074			1078	1074		1074	1076		MBN (δ C-H)[149], co-varies with cell
				1164			1162						1159						1164	Cell, possible DOP[191]
1178			1178		1178				1178		1178				1178		1178			MBN (δ C-H/ν C-CN)[149]
			119		1199				1199		1199				1199		1199			MBN (n.a.)[149]
								1213		1217									1211	DOP+[176]
		1244	1244		1244				1248		1244							1244		MBN (n.a.), co-varies with cell

Reference			MBN s+b		MBN s	MBN b		DOP s+b		DOP s		DOP b		DOX s+b		DOX s		DOX b		Tentative assignment
MBN	DOP	DOX	t	to1	t	t	to1	t	to1	t	to1	t	to1	t	to1	t	to1	t	to1	
	1277		1273				1272								1273					(n.a.)
				1282				1286		1284			1280						1282	DOP^{+176} and cell
	1347			1343			1343						1345						1347	(n.a)
								1372				1370								Possible DOP (n.a.)
									1398		1398	1398			1398		1398			Possible DOP (n.a.), likely co-varies with cell
		1436												1436		1438		1439		Possible DOX[165] overlap with cell
			1475	1471	1477		1471		1475		1477				1477		1477		1479	MBN (ν aromatic), possible covariance with cell
				1529			1528	1525							1528				1529	Possible DOP^{-176}
		1564												1564		1560		1566		DOX (n.a.) overlap with cell
1584			1584	1586 n		1584	1582 n		1584		1584	1588	1589		1584		1584			MBN (ν aromatic), co-varies with cell
				1593				1595		1591									1590	DOP[176]
	1606						1604					1606								DOP^{+176} overlap with cell
				1751			1752		1758 n										1751	Cell
				1871		1871													1871	Cell
			2129	2123	2126	2127	2122			2127			2123		2128		2126		2123	MBN (n.a.), co-varies with DOP and cell
			2173	2170	2177				2173						2172			2172		MBN (n.a,.), co-varies with cell
2228			2228	2228 n		2228			2228		2228				2228					MBN (ν C≡N)[149]

In order to clarify the T-OPLS models further, the majority of the regions of the spectra that were overlapping and co-dependent were removed. Peaks that did not shift their position were left in. After the exclusion of regions where overlap or co-dependency were suspected, a comparison was made between those models as in **Table 1.2**. It was observed that some peaks separated from earlier observed co-dependent signals. This was clear in particular in the case of MBN peaks. For instance, the peak at 1074 cm^{-1} has been identified as stemming from MBN alone once the spectra were trimmed, rather than from MBN and as stemming from a co-dependency between MBN and cell signals. Other MBN peaks that were resolved from other sources were 547, 1584, and 2178 cm^{-1}. Unfortunately, the removal of spectral regions did not have the same positive effect on the resolution of DOP and DOX signals, with the exception of the DOP peak ~1394-1398 cm^{-1}.

Lastly, we plotted hyperspectral images univarietly, that is, by selecting only one wavenumber instead of a full spectrum. The idea was that a single wavenumber clear of co-dependency and overlaps may have produced clearer distribution images.

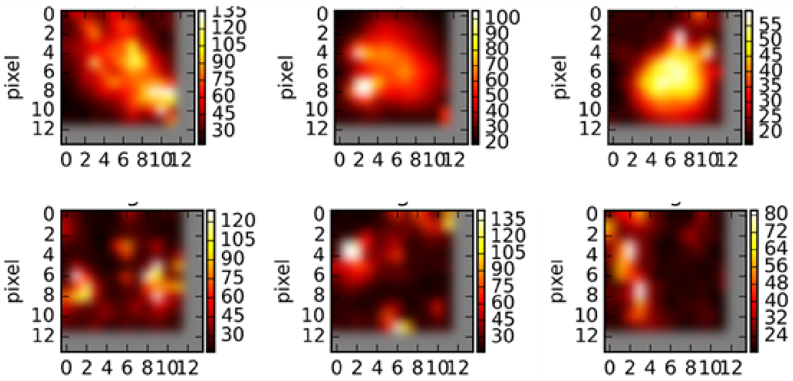


Figure 1.14. Hyperspectral images of PC12 cells plotted univariatly with reference to 365 cm^{-1} peak of MBN. The upper row contains three cells treated with 100 M DOX and MBN-AuNPs, and bottom row contains three cells treated only with MBN-AuNPs.

When using the univariate approach, the false positive issue mentioned earlier remained as illustrated clearly in **Figure 1.14** – a MBN signal was found where there should have been none (blanks). Similar images were done for a peak of DOP at 1590 cm^{-1}, but the images looked exactly the same as images for 365 cm^{-1}, as well as for other MBN peaks, e.g. 1074 cm^{-1}. Put differently, the univariate application made things worse, implying that there were spectral overlaps between 365 cm^{-1} and the cell and analytes. This was in contrast to the multivariate modelling performed here, which clearly indicated the 365 cm^{-1} peak as independent of any influence that may have caused false positives.

CONCLUSIONS

Several issues concerning method development have been addressed. The cell adhesion problems were solved for granulocytes and monocytes by switching from PLL to PDL. Longer incubation times promote cell detachments, thus shorter incubation times were recommended. The cell adhesion protocol devised within this thesis did not work for lymphocytes.

It has been demonstrated through the course of this work that healthy, living lymphocytes were not capable of colloid internalisation in light of their inherently low phagocytic activity. This claim was supported both by experimental observation, multivariate analysis, and the literature. Therefore, the quantification of this type of cells based on the utilised methods, i.e. SERS and Raman spectroscopy, was not appropriate. Lymphocytes as a model would have been ideal due to the fact that the developed method aimed to improve treatment strategies for ALL patients. Internalisation also failed when neutrophils were used as model cells, since they did not display phagocytic activity towards particles without opsonins. To validate whether internalisation has occurred or not, the method in this chapter could have been complemented with other techniques, such as transmission electron microscopy (TEM). Such techniques would have enabled tracking of the AuNPs internalisation in the cellular compartments[116]. Internalisation could also have been improved if cancerous cell lines had been used, as they are more permeable to nanoparticulate matter. But as we had issues in detecting DOX in healthy cells, we did not validate the method in cancerous cells.

Internalisation did succeed in monocytes – this conclusion was based on the experimental observations, where a signal enhancement was observed in cellular interior and from hyperspectral images. If this evidence of the colloid internalisation of monocytes was further supported by for instance TEM, it would be feasible to use monocytes as a model for ALL in the future, since both lymphocytes and monocytes are: a) morphologically similar and; b) to the author's knowledge do not display differences in the accumulation of DOX. Also, monocytes could provide a good model in similar projects where the focus lies on acute myeloid leukaemia (AML) instead of on ALL. Yet, with current machine learning methods it was not possible to determine the presence of DOX inside monocytes at the levels of DOX investigated in this work. The reason for this was suspected to be a failed accumulation of DOX inside the cytosol within the used incubation time frames and/or that the used DOX concentrations were below the instrument's and the method's LOD. Establishing a LOD would therefore be of benefit if DOX calibration were to be established in the future. Yet, as the methods applied here rely on multivariate data analysis, the LOD must also be calculated multivariatly. Although LOD does not increase sensitivity per se, its establishment can be argued to be of good analytical practice. Unfortunately, there are currently no standardised methods available that would allow for multivariate LOD calculations. Initial enquiries have been made by the author into this matter, as it is significant for both for this chapter and for Chapter 3. For a further discussion on multivariate LOD, see **Appendix 1.1** at the end of this chapter.

Moreover, it is still unclear why earlier investigators were able to detect DOX inside the blood cells, while attempts to do so in monocytes and PC12 cells within this thesis had low levels. Apart from not sufficient accumulation times during incubation with DOX, another reason for failing to detect it in the presence of MBN-coated AuNPs can be attributed to the lack of attraction between the positively charged DOX and the neutral MBN-modified gold surface. If that is the case, it would be interesting to attempt to replace MBN with another IS that would possess a negative net charge. Alternatively, it would be possible to further investigate how the extent of the MBN coating influences the DOX signal enhancement. If this detection issue were clarified, it would be possible to impose further improvements on the method by developing a DOX calibration curve based on OPLS. Such a curve would make it possible to find intercellular DOX concentrations.

After peak analyses had been performed on the score vectors from T-OPLS modelling done on PC12 cells, it was concluded that T-OPLS could indeed separate signals of IS from analytes and the matrix. MBN did not suffer the same bleed-over effects between its principal and orthogonal vectors as had occurred in models performed on monocytes. This was likely to be due to the use of a different pre-processing strategy compared to the one used on blood cell spectra. This different strategy resulted in distinct spectral features of MBN with many strong, clear peaks, several of which were outside the fingerprint region. Moreover, T-OPLS was able to model peaks for IS that were absent from its reference spectrum, but which ought to have been there. Unfortunately, DOX and DOP fared worse in T-OPLS as their spectra had spectral regions that overlapped with each other and with the cell signal. Hence, the origin of many peaks unrelated to MBN remained ambiguous, both with respect to which substance a peak corresponded and whether it was a matter of spectral overlap or signal co-dependency.

Co-dependency as the source of ambiguity was not anticipated during the conception of this enquiry, but was highlighted during the T-OPLS analysis of PC12 cells. These co-dependencies became evident in **Table 1.2**. and in hyperspectral images, where signals between the IS and the analytes (DOX or DOP) – sometimes even between analytes – were not separated by T-OPLS into appropriate components. To establish whether the observed co-dependencies were linear or nonlinear, it would have been beneficial to apply DoE, since such correlating signals when observed in the application of linear projection algorithms can be an indication of the presence of quadratic and cross-terms. Indeed, the observed **non-selectivity** between MBN and DOX was in itself an indication of a poorly-designed experimental space. Unfortunately, considering the matrix in which analytes were measured, it would not have been possible to create a design where the factors chosen for the analysis (AuNP entry, cytosol analyte concentration, etc.) could be controlled. For instance, by varying the AuNP coating, the uptake could be affected, thus introducing another factor into a design space which would be out of control.

In summary, the research presented in this chapter did contribute towards understanding the advantages and the shortcomings of T-OPLS as a tool for intracellular imaging, quantification, and qualitative spectral interpretation. Because of the PLS-based nature of T-OPLS we have arrived at the conclusion that although it can give an advantage when separating of overlaps and modelling absent peaks that have linear co-dependencies, T-OPLS performs poorly when faced with the task of separating numerous and versatile co-dependencies. As a linear algorithm, T-OPLS is strongly dependent on the spectra to have linear co-dependencies only for optimal functionality. The good separation of the MBN signal indicates also that the algorithm needs the presence of strong to medium, clearly resolved peaks, which are not attributes that can be ascribed to biological matrices. To improve upon the algorithm, 2-way cross-validation (i.e. cross-validation in both columns and rows) as it is used in PCA, would have been beneficial. At present, T-OPLS only has cross-validation in rows, meaning that until 2-way cross-validation is available, there is not enough information to go on to provide the necessary improvements.

APPENDICES FOR CHAPTER 1

APPENDIX 1.1 – THE QUEST FOR MULTIVARIATE LOD

INTRODUCTION

Among the traditional ways of analysing Raman spectroscopy are the selection of a single wavelength and subsequent modelling and/or imaging based on univariate approaches[82]. However, to achieve a robust univariate model two basic requirements need to be fulfilled. The first is **selectivity**, i.e. the input data must be that of pure analyte without any contributions from other species in the matrix. When responses others than that of analyte are present, a univariate model will become biased. The second requirement is **linearity**, i.e. there should be a linear relationship between the spectral response and the analyte concentration[35]. While the linearity requirement can be amended with nonlinear solutions, selectivity is much harder to address when spectra originate from for example. biological matrices. With multivariate techniques, however, one can handle the presence of non-analyte signals, provided that their response profiles are not identical to that of the analyte and that the non-analyte signal is varying somewhat independently of the analyte variation.

All analytical methods' development benefits from validation. The analytical chemist delving in Raman spectroscopy of multicomponent matrices must therefore concern him/herself with appreciating any shortcomings by performing an evaluation of basic analytical parameters. In analytical chemistry, the most common way to validate a method is to use figures of merit, which are measures describing the quality of a method. The afore-mentioned real and apparent nonlinear sources in Raman spectroscopy (see section **The nonlinearity issue**) provide a challenge not only in multivariate analyses described in this thesis, but also in estimating multivariate analytical figures of merit.

One way to make low Raman signals more interpretable is by applying multivariate tools instead of the traditional univariate ones[35]. Chemometric tools for qualitative and quantitative analysis have been gaining steadily in popularity, particularly in the evaluation of spectroscopic responses[35]. One prominent use of multivariate tools is constructing predictive calibration models in order to quantify an analyte. When it comes to calibration, multivariate tools can be defined as implicit modelling methods, the opposite of explicit methods where a theoretical model is imposed on experimental data (see difference between EDA and other statistical approaches in section **1.1 Assumptions behind the application of machine learning**). In other words, implicit methods incorporate variance in the data set that is does not necessarily stem from the analyte of interest[192]. This enables the use of a better statistical starting point, which in turn provides improved accuracy. For quantitative enquiries, a multivariate calibration model aims to find a predictive model that relates the instrument responses to concentrations[193], i.e. that attempts to predict unknown concentrations of the analyte.

To evaluate the precision and sensitivity of Raman response for purposes of quantification and detection, limit of detection (LOD) is one of the fundamental analytical parameters that can be calculated. If there is a poor SNR, and if the studied concentrations are near the detection limit, LOD estimation can be challenging. Nonetheless, for purposes of multivariate calibration, it is advisable to find a LOD value for the analyte of interest before constructing a calibration model. There are several methods for estimating LOD in spectroscopy, but due to the heterogenic nature of data studied, each data set requires dissimilar calculations. Moreover, LOD calculations proposed in literature often do not take into account all of the uncertainties present[72]. There is therefore a need to develop LOD calculations along with multivariate statistics that can be applied to a wide variety of data.

Bearing various sources of nonlinearity in mind (for instance spectroscopic nonlinearities and matrix effects), the need to quantify and predict analyte concentration becomes clear. Further, finding a unified estimation of analytical parameters becomes challenging in the sense that a plethora of nonlinearities require dissimilar mathematical methods to create a fitting model[84]. LOD is one of the basic analytical tools for method evaluation in a Raman spectroscopic measurement; yet, there is to date no well-developed standard for multivariate LOD estimation within Raman spectroscopy[72].

The main challenge in multivariate LOD estimation is that Raman signals are not constrained to a specific analyte, i.e. analyte signals may often be confounded with other responses. There are strategies that can be used to address this relying on standard deviation (SD) of the blank based on residuals or on the net analyte signal (NAS). However, these strategies consider only the uncertainty in the signal measurements, and disregard sources of uncertainty in calibration concentrations[72].

In multivariate calibration the leverage *h* for each observation assumes different values depending on the composition of the sample. As the leverage is different for different samples, when approaching LOD calculation in a multivariate calibration, it is plausible to consider a LOD interval rather than a single LOD value. In this way, the LOD values within the interval will depend on the variability of the matrix composition, thus minimizing the propagation of concentration errors into the standard error in predicted concentrations[72]. Therefore, it is necessary to estimate the calibration concentration uncertainties in a reliable manner (either from replicate reference measurements or from error propagation considerations).

The aim of this research was to elucidate whether it was possible to establish LOD calculations inspired by Allegrini et al.[72] – as is or modified – for full-spectrum responses. Before exploring multivariate LOD calculations, the author wished to gain a deeper insight into how Raman spectra behave when used for multivariate calibration with linear projection algorithms. With regards to this aim, a series of simulations were performed containing spectra of three substances; two interferents and one analyte, following the simulation described in Allegrini et al.[72]. Each substance had one single peak.. By gradually adding interfering features and nonlinear behaviours (noise, quadratic function, co-dependencies, baseline variations, overlaps, etc.) to the simulated spectra, the aim was to see where standard OPLS start to fail in quantification and prediction. This investigation was highly relevant for the research in this thesis where T-OPLS was applied; because little information exists on how the algorithm functions, so modelling with T-OPLS based on simulated spectra would provide the much needed insights into its capabilities and shortcomings. Successful simulated data would also allow the application of the method to real data taken both from previous projects[25, 47, 57, 194] and from new experiments.

RESULTS

Using Python scripts with numpy library v1.19.0, data sets containing 100-5000 randomly generated spectra with Gaussian peak shapes were generated. The data sets differed in the type of noise that was added; spectra with noise only in spectral signal, and spectra with noise in signal + concentration. All noise was additive and no nonlinearities were added. In each data set, one of the two interfering components had a slight overlap between the spectral features of the analyte, but each substance had a concentration variation independent of each other. Thus, the simulation was aimed at examining the effect of the overlap and noise on the behaviour of OPLS, when co-dependency was avoided using randomised concentration combination for the analyte and the interfering compounds. The response for all substances was linearly proportionate to their concentrations.

In each of the simulations, OPLS models were calculated with varying amounts of noise added to the Y vector. The Y vector described the analyte concentration variation. With spectral noise on the 30% standard deviation in relation to the maximum spectral magnitude and no noise in concentrations on the Y side, there was a complete separation between the analyte peak in the **p** loadings and the interfering peaks in the **po** loadings irrespective of the training set size that was varied in the range of 100, 500, and 5000 samples.

For all models, at least two OPLS-components were always used even when the second component caused a lower Q^2. For the noise levels on the concentration side, the relative standard deviation (yRSD) was set to 5 %, 10 %, 15 %, and 30 % relative to the maximum concentration.

At 5 % yRSD the **p** loadings for the 100 sample model showed some hints of interfering peaks while the 500 and 5000 sample models were without any interference. At 10 % yRSD **p** loadings for the 100 and 500 sample models show similar results. At 15 % yRSD there were interfering peaks for the 100 and 5000 sample models but less so for the 500 sample model. At 30 % yRSD, there were clear interfering peaks for the 500 and 5000 sample models but only weak ones for the 100 sample model, where they were possibly hidden by the larger noise level in the **p** loadings due to few samples (**Figure 1-A**). On the orthogonal side of the model there were clear analyte peaks for all three models in the 30% yRSD **po** loadings, while the analyte peak was not expected due to the normal function of OPLS. Weak hints in the 15 % yRSD **po** loadings, weak hints for 100 and 500 sample models at 10 % yRSD po-loadings, and no hints for any sample number at the 5 % yRSD level.

The Q^2-values for the 100 sample models were: (0.90, 0.87 (3 comp), 0.84, 0.68, 0.26) for the yRSD % levels: (0, 5, 10, 15, 30), respectively.

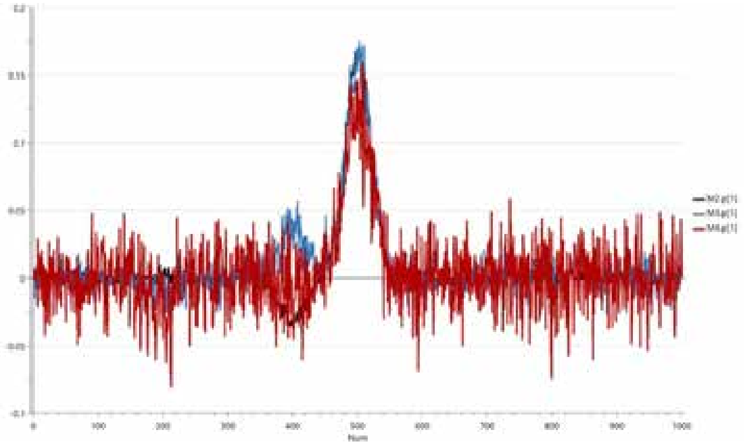


Figure 1-A. Predictive loading vectors **p** from an OPLS model where Y = analyte concentration. The predictive loadings show variation dependent on the selected Y from model with 100, 500, and 5000 samples, (red, blue, black) from models with 30 % yRSD noise on the X and Y sides. The peaks of the analyte signal was present at x = 500. Interfering component peaks 1 and 2 were located at x = 200 and x = 400.

Thus it can be seen that when the noise in spectra is high at the same time that the noise in concentration levels increase, there exists a limit for OPLS where the y-predictive vs orthogonal part of the model blend due to the noise level but not due to the relative independence of interferences and analyte spectral contributions. This effect appears to be alleviated somewhat by using more spectra per model. The Q^2 value will, as usual, be helpful here. For the 30 % yRSD the Q^2 is clearly low at (0.26, 0.39, 0.38) for the 100, 500, 5000 model sample sizes while at 10 % yRSD level, the Q^2 values are at (0.84, 0.85, 0.86). As regular OPLS is a linear method, non-linear effects will also appear as higher noise levels to the standard OPLS algorithm as this cannot be modelled directly. This may be one reason why there is also a blend effect for non-linear models. However, these results also mean that the blend effect cannot be taken as a pure signal for the presence of non-linear effects when measurements are close to the noise level.

SUMMARY

It was observed that OPLS becomes less suitable at certain levels of noise, regardless of sample size. The bleed-over effects could therefore have been caused by a mixture of smaller nonlinearities and low SNR. It should be noted that the enquiry reported here was a slight expansion of the LOD concept in the sense that it covered the interpretability of OPLS models at low SNR levels. This was in contrast to traditional multivariate LOD calculations attempts, which deal more with the predictive ability at low SNR levels.

The original aim which instigated this enquiry, the development of multivariate LOD calculations, would necessitate the deconstruction of typically observed linear and nonlinear effects in Raman spectroscopy. This was done through the simulation of spectra that resemble real Raman spectra as closely as possible, but that would provide an opportunity to study separately the influence of each possible linear and nonlinear effect on multivariate modelling separately. An additional enquiry was to see how different response functions (e.g. linear or quadratic response) influence modelling when present within the same spectrum. This in particular was intended to relate to the quadratic response and bleed-over effects observed in the literature[48] and in the author's own research[194], since, in case of the latter, the reasons behind the majority of the reported behaviour could not be explained. Unfortunately, the author was not able to make further enquiries into this matter. Future investigations are recommended which include nonlinear effects at low SNR and which extend the simulated spectral peak combinations with peak shapes common in Raman in addition to the already simulated Gaussian peaks, such as Lorentzian, Gauss-Lorentzian, and Voigt peaks.

CHAPTER 2

RAMAN SPECTROSCOPIC METHOD FOR *IN SITU* QUANTIFICATION OF PEG IN ARCHAEOLOGICAL WATERLOGGED WOOD

INTRODUCTION

When artefacts of wood are excavated, they must undergo a conservation process that preserves the structural integrity of the artefacts before they are incorporated into humanity's cultural heritage. Although initially appearing structurally sound upon excavation, waterlogged archaeological wood is usually severely degraded. In waterlogged archaeological wood such degradation is often caused by bacteria inhabiting the anaerobic waterlogged environments of archaeological finds of this kind (bogs, sediments, ocean floor, etc.)[195]. These erosion bacteria are capable of degrading many wood constituents, thus compromising the integrity of the cell walls[195-197].

Should such wood be dried without the necessary conservation steps, the material will suffer anisotropic shrinkage. To prevent this, waterlogged archaeological wood is often impregnated with a consolidator, e.g. the water soluble polymer polyethylene glycol (PEG).

In Sweden, one of the major conservations projects with PEG is the warship *Vasa*[198]. PEG enters the wood by slow passive diffusion, a process spanning decades. The wood is spray treated or submerged in aqueous solutions of PEG. The process is slow in part due to the fact that the PEG concentration has to be kept low in order to avoid large osmotic pressures that can damage the artefact further[199-200]. Nevertheless, the impregnation of *Vasa* with PEG was an apparent success and became thus widely used for other object of waterlogged archaeological wood. Since the impregnation with PEG became popular, it has been criticised due to PEG's hygroscopicity, which has contributed to a weakening of the wood matrix, and to chemical instability[199, 201].

The process of impregnation requires continuous and reliable monitoring in order to determine whether PEG impregnation is complete. In the case of *Vasa*, the quantification of PEG levels has been done by sampling cores of the ship's wood, from which PEG was extracted and then quantified by chromatography or spectrophotometry[202]. Clearly, such quantification attempts are both resource-intensive and destructive to the cultural heritage. The development of methods that would permit PEG quantification without the destruction of archaeological artefacts is therefore of analytical interest.

There have been attempts to develop *in situ*, non-destructive analytical techniques. Jeremic et al.[203] and Gierlinger et al.[204] used Raman spectroscopy to estimate the ratio between wood constituents and PEG spectral bands. These bands were found to be proportional to PEG concentration.

Although simple to implement, these methods are lacking in the sense that they were developed for microanalysis and do not produce absolute values. This makes them unreliable to base conservation strategies upon. Another study used Raman band ratios in PEG impregnated archaeological wood. In it, a calibration was created based on known mixtures of PEG and wood[205]. Unfortunately, few details of the validation procedure were reported. The common drawback of these studies is that it is assumed that the levels of wood constituents are constant and are used therefore as internal standards. However, this approximation is not true for the most archaeological wood as some of these constituents (e.g. cellulose) are also subjected to microbial decay[197, 206].

As wood is a complex biological matrix, it contains several chemical species that may vary and potentially complicate quantification. In Norway spruce (*Picea abies*), the species used in this study, the lignin and hemicellulose content are at about 27 % each and cellulose at 40 %. The remaining 6 % are extractives (waxes, alcohols, terpenoids etc.) and inorganics[207].

In this chapter, the first steps towards an *in situ*, non-destructive Raman spectroscopic method for quantification of PEG in waterlogged archaeological wood are reported. A method of this kind would enable rapid measurements directly on the sample without any sample preparation or destruction[208].

As the chemistry of archaeological wood is highly heterogeneous and amorphous, a simplex mixture design was used to reflect variation in PEG and to include the two major constituents of wood – cellulose and lignin. Since the matrix of PEG is complex, undesired effects inherit in Raman spectroscopy such as photoluminescence might obscure PEG signals. Alternatively, the signals of the various chemical species present may overlap[209-210]. To address this, the accumulated spectra were analysed with projection-based multivariate algorithms of principal component analysis (PCA) and orthogonal partial least squares (OPLS). The latter was used to build calibration models aimed at predicting PEG concentrations by seeking the maximum covariance[211] between the Raman spectra and the known concentrations of the mixture design components.

WATERLOGGED ARCHAEOLOGICAL WOOD

Waterlogged archaeological wood is a delicate matrix, both with regards to its physical properties and also considering the challenges in conservation of such cultural heritage. One of the main challenges in preservation of waterlogged archaeological wood is its weakening upon drying. To prevent the collapse of such objects upon drying, archaeological artefacts of this nature are often impregnated with polyethylene glycol (PEG). However, since the impregnation process relies on passive diffusion once the artefact has been submerged in PEG, it is a lengthy process. Therefore, there is an interest in having a continuous monitoring of the impregnation process, as this ought to be non-destructive towards the analysed cultural heritage. The objective in Chapter 4 was to develop an *in situ* method for the quantification of PEG in waterlogged archaeological wood.

Wood is a heterogeneous matrix, with common constituents varying greatly depending on the extent of degradation[197, 206] and wood species. The main chemical species present in wood are hemicellulose, lignin, and cellulose. The celluloses are polysaccharides and lignin is made up of phenolic units forming a polymer chain. Both lignin and hemicellulose are present in wood at about 20-30 % of the dry weight, while cellulose constitutes about 50 % of the dry weight in most wood types. The cellulose polymers are mainly ordered into microfibrils, held together by both intermolecular hydrogen bonding and van der Waals interactions (**Figure 2.1**). These aggregations exhibit varying degrees of ordering, with regions of both crystalline and amorphous structures present in wood[206-207].

Lignin on other hand is a crosslinked polymer - commonly described as amorphous - of three precursor alcohols; coniferyl, sinapyl, and p-coumaryl (**Figure 2.1**). Unlike cellulose, lignin is highly disordered and lacks the constant structure that may be ascribed to certain species of wood[212].

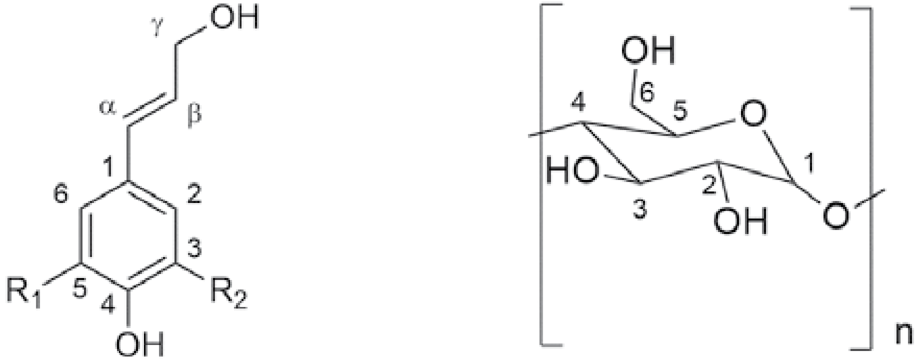


Figure 2.1. Left: lignin where p-coumaryl has no branching, coniferyl has one methoxy group (R2) and sinapyl has two methoxy groups (R1 and R2). Right: glucopyranose. Cellulose is a series of polymers of glucopyranose (1-O-4) with alternating stereochemistry of the monomers.

Hemicellulose consists mainly of xylose, galactose, and xylose arranged into a branched heteropolysaccharide. The order of polymerisation is in places random and in places amorphous[207]

The relationship between the aforementioned three main constituents of wood is not fully understood. Currently, how hemicellulose, lignin, and cellulose relate is described roughly as follows: hemicellulose and lignin form a network in which cellulose microfibrils are entwined, offering rigidity, but also resistance to bacterial degradation[213-215]

Here, the method is developed with Norway spruce (*Picea abies*), a softwood, which compared to hardwoods has a simpler morphology[206].

METHODS

THE CALIBRATION SET: PREPARATION OF MILLED WOOD LIGNIN (MLW)

In the simplex mixture design, the lignin was represented by milled wood lignin (MLW), which was prepared in house according to the procedure described by Björkman[216]. The steps of wood meal drying with P_4O_{10} and milling it in toluene were omitted. The drying was instead performed over silica gel, which resulted in a somewhat coarser and oxidized product.

For the extraction of lignin, a fine wood meal was prepared, and extractives were removed per established procedure[217]. The wood meal was grinded with a Retsch rotor mill to mesh 40 from recent sapwood of Norway spruce (*Picea abies*). Then, 8 g of the wood meal was place in a Whatman 603 cellulose thimble (VWR, Sweden) and was subsequently extracted in a Soxhlet set-up with 200 mL acetone (GC-grade, Fisher Scientific, Sweden) for a minimum of 24 circulations. The Soxhlet set-up was heated by a water bath at temperature 65-70 °C and was only partially covered with tin foil to avoid thermal decomposition of lignin[218].

After three extractions, the average amount of extractives was 0.8 ± 0.2 wt%, which corresponds to the range reported in literature for Norway spurce[217, 219]. The product was then dissolved in AcOH with 10 % water, and then precipitated by a dropwise addition to water. The precipitate was collected by centrifugation and then washed until it reached neutral pH. The yield was recorded to be 30 %, which is consistent with earlier reports[216]. Raman spectroscopic analysis of MWL showed a clean lignin spectrum (see Figure C in Appendix 4.1 at the end of this chapter). The peaks of MWL are similar in shape to the lignin found in wood[220].

THE CALIBRATION SET: THE MIXTURE DESIGN

The statistical design of experiments is used to model the response-factor space with a mathematical expression. One type of design of experiments is called a mixture design in which the sum of components present must always add up to 100 %. The response then depends then on the relative proportions in the mixture, not their absolute concentrations. There are a number of mixture designs that accommodate different needs. If the aim of the experiment is to test the responses of an untested mixture, then a simplex design is suitable[37].

PLS may perform better in relation to multiple linear regression (MLR)[221] when applied to a mixture design since the factors in the design are correlated. In the best of cases, a PLS models should be able to separate the three different components from the three factors varied in the design[222]; MLW, cellulose, and PEG. A similar behaviour was expected for OPLS models performed in this chapter.

To represent the holocellulose component of wood (cellulose + hemicellulose), cellulose from cotton linters were used (Sigma Aldrich, Sweden). Both cellulose and hemicellulose have been reported to have similar Raman spectra[220, 223], which were confirmed by spectral analysis of cellulose from cotton linters (**Figure 2-B** in **Appendix 2.1**).

Calibration standards were mixed as per simplex lattice mixture design[222] (**Figure 2-E** and **Table 2-E** in **Appendix 2.2**) using cellulose powder, MWL and PEG-4000 (3500-4500 Da, synthesis grade, Merck, Sweden) in concentrations ranging from 0 to 100 %. Mixtures were prepared through grinding cellulose powder and MWL in a porcelain mortar, transferring the powder into a vial and then adding the aqueous PEG solution. The suspensions were composed of 2/3 of water to emulate the water ratios in waterlogged softwoods and moderately degraded waterlogged archaeological wood[224]. After thorough stirring, the suspensions were transferred onto microscope slides for spectroscopic analysis.

The design contained seven levels with 22 points, calculated using JMP® software (version 13, SAS Institute Inc.). Centre and end points of the design were triplicated to evaluate the precision of the method. An additional set of six measurements were prepared to create a test set, three of which were located at the design centre (25:25:50) and three at the design exterior (10:10:80).

THE VALIDATION SET

The validation set consisted of samples of archaeological waterlogged wood (AW) and recent wood (RW), both Norway spruce. The RW was acquired from the same sapwood from which MWL was prepared. The subsamples of RW with dimensions 1x3x5 cm were submerged in aqueous 20 wt% solution of PEG-4000 (3500-4500 Da, synthesis grade, Merck, Sweden) for 3 months.

The AW was taken from a circular pole excavated in Motala Ström (a river in Sweden) in 2010. The pole was found at the site along with other poles driven into the river bed, and it is believed that they were a part of a fishing station. No dendrochronological examination was performed on this particular sample, but the majority of the poles found at the site have been dated to the 11th century. After recovery, the pole was stored in a waterlogged state at 8 °C. In the spring of 2017, the sample was submerged in an aqueous PEG-4000 solution of 20 wt%.

Subsamples were taken from RW and AW by cutting cross-sections by hand with a double-edged razor blade. Each subsample was ca 5x5 mm with a thickness of 100 μm. The cuts were made along the longitudinal axis, as it is the main transport route in wood. The samples were placed on Raman grade CaF_2 microscopy slides (Crystran LTD., UK) fitted with a 8-well sticky slide (ibidi GbmH, Germany). Wells not containing samples were filled with water to ensure a humid environment around the samples. Calibration standards described in the section above were analysed in the same fashion.

INSTRUMENTATION AND MEASUREMENTS

The Raman spectrometer (Dilor Labram IV, Horiba, France) was equipped with a coupled charge device (CCD) detector and an inverted confocal microscope (Olympus IX70, Japan). The excitation wavelength used was 632.8 nm He/Ne laser with intensity of 2 mW at the sample. A 950 groves/mm grating was used to record between 200 and 3200 cm^{-1}. Each acquisition was measured with an x10 objective and was triplicated for 20 seconds and then averaged into a single spectrum. The focal spot was calculated to 5 μm. All samples were photobleached for 6 minutes before spectral acquisition in order to reduce photoluminescence (PL). The photobleaching was not detrimental to the Raman signals from RW or AW (see results of the analysis in Results and Discussion).

The RW and AW samples were analysed in five and six different spots, respectively (**Figure 2.2**)

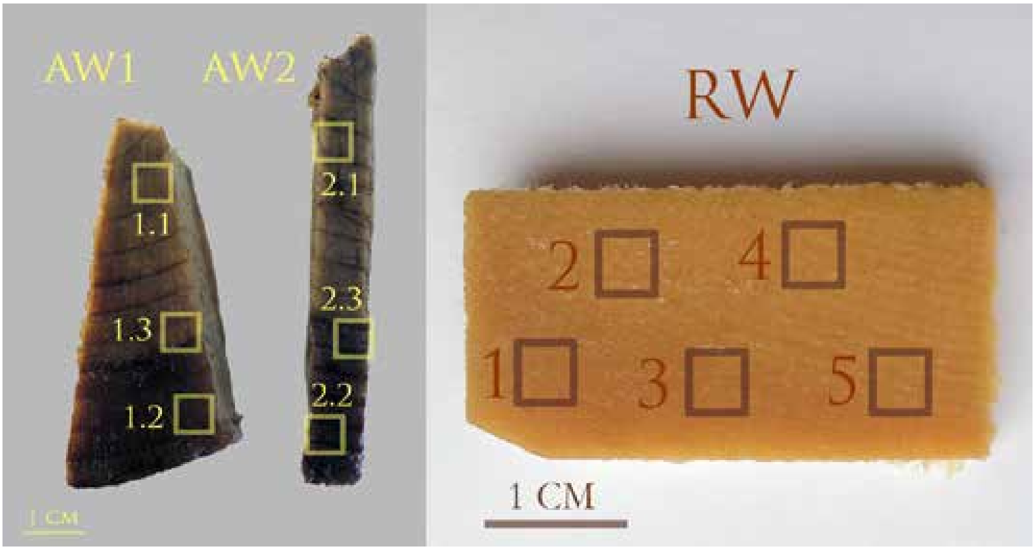


Figure 2.2. Photos of the AW (left) and RW (right) test material. The marked areas represent sampling sites. AW1 and AW2 come from the same pole found in Motala Ström river. Areas 1.2 and 2.2 were the utmost layer of the pole.

The sites were selected at different depths of the wood sample to accommodate for the spread of PEG concentrations. In the case of AW samples, the selection of sites was also governed by the distance to the exterior of the original pole, since surfaces exposed to water were more decayed. This would reflect the variation in lignin and cellulose content, as the latter were being subject to microbial degradation. Three sections were sampled in sequence from each subsampling site in attempt to account for the heterogeneous nature of the material. Since the laser spot (5 μm) was smaller than the wood micromorphology (**Figure 2.3**), nine randomized acquisitions were performed in each section, thus resulting in 27 acquisitions per subsampling site.

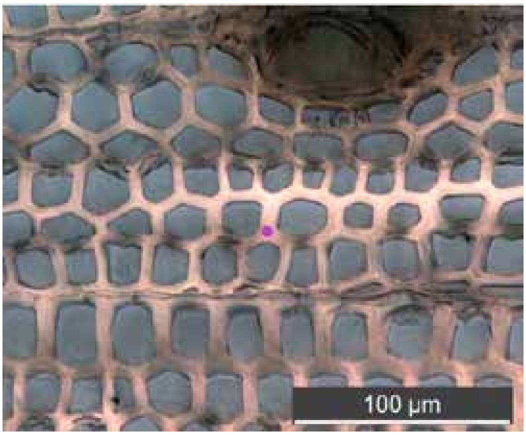


Figure 2.3. Micrograph of AW. In the centre, a purple dot marks the diameter of the laser spot: 5 µm with X10 magnification of the confocal microscope.

The standard mixtures were assumed to be more homogenous, and so only 5 acquisitions were performed for each standard. These acquisitions were regarded as separate observations and were used individually to construct the calibration model.

PEG EXTRACTION

After the Raman analysis was complete, the PEG content in AW and RW was estimated by extracting samples with MeOH. PEG content was estimated gravimetrically. The triplicate sections were dried in a desiccator for a week and then weighed. The dried samples were then placed in 2 mL MeOH (GC-MS grade, Merck, Sweden). The MeOH was changed daily until no further weight change was noted after subsequent 24 h drying in the desiccator. There was no change in weight after 72 h of extraction.

As the weight loss was caused by the loss of extractives, and, in case of AW, by degradation products, the error was estimated. To estimate the error originating from other compounds than PEG, sample material which was not subjected to PEG impregnation was treated according to the same extraction scheme that was described above. AW and RW extracted 1.5 and 0.8 wt%, respectively with a relative standard deviation (SD) of ca 50 % (n=6). This corresponds to the extractive contents in spruce[219] and the error found between PEG extraction with MeOH in archaeological wood for HPLC[202]. The error of gravimetrical determination was assumed to be equivalent to the average extracted +2 SD, which gives an error of 1.6 and 3 wt% for RW and AW.

SPECTRAL PRE-PROCESSING

To reduce the contribution of nonlinear sources (read more in the Methodology section) originating from the samples and the instrument[80], the data was pre-processed using rolling circle filter (RCF)[63] with the circle radius set to 98 cm^{-1}. RCF was performed using an in-house Python 3.5 script.

As Raman spectroscopy is sensitive to molecular conformations, orientation, and ordering[225], spectral discrepancies were expected in the samples in the calibration set and the validation set, and cannot be considered originating from representative concentrations. All data was scaled to unit variance 1 (α=1) (UV).

MULTIVARIATE CALIBRATION AND MODEL VALIDATION

OPLS models were constructed using the calibration set. The models were validated using cross-validation and root mean square error (RMSE) of estimation[37], the latter calculated for the validation set. The quality of OPLS predictions was evaluated using a linear regression of the observed vs. predicted values. The coefficient, which is equivalent to R2Y, the slope, and the intercept were determined. The significance was tested with Student's t-test to highlight any systematic errors.

All PCA and OPLS models were constructed in a SIMCA 15.0.1 (Sartorius Stedim Biotech, Umeå, Sweden).

RESULTS AND DISCUSSION

In this study, the goal was to achieve the prediction of PEG content based on Raman response. This is commonly achieved by regression analysis, where, in this case, the relationship between uncorrelated variables (PEG concentration) and correlated variables (Raman spectra) is evaluated. Revealing covariance between the correlated and uncorrelated variables allows prediction. The uncorrelated variables, i.e. PEG concentrations, were held in matrix Y, and correlated variables represented by Raman spectra of samples in the mixture design were held in matrix X.

The mixture design herein was used under the assumption that PEG's response depends on the relative proportions of other wood components, represented here by cellulose and MWL, rather than by their absolute amounts. The predicted wt% values calculated by the models below give the ratio of PEG to MWL and cellulose. Therefore, any other species present in the matrix not based on saccharide or lignin chemistry were not accounted for in the design.

PHOTOBLEACHING EXPERIMENTS

Raman signals can be optimized through photobleaching, which aims to bleach chemical species causing the interfering PL. The benefit of such a procedure is that the contribution of PL can be reduced, thus improving noise which could otherwise mask relevant Raman signals. A prolonged exposure to the laser could also change the properties of the studied analytes, thus changing the spectral response[226].

The material in this study produced a significant amount of PL. The Raman spectra generated from both AW and RW showed a clear trend in which the background noise diminished upon longer exposure. However, it has been reported that prolonged exposure to a 514 nm laser reduces lignin signals[227] and that lignin absorbs radiation at 633 nm[228]. Because of these exposure effects, the photobleaching had to be evaluated.

AW and RW were exposed to a laser for 30 minutes with spectral acquisition carried out at the 1, 2, 3, 4, 5, 6, 7, 8, 9, 15, 19, 20, 25, and 30 minute points. The integral of the 1566-1637 cm^{-1} lignin band was calculated for each spectral acquisition. The integral of the area of PL was also calculated. Measurements were performed on both early and late wood. Integrals were calculated in the LabSpec 5 (Horiba) software, which cuts a straight line between the start and end point of the integral.

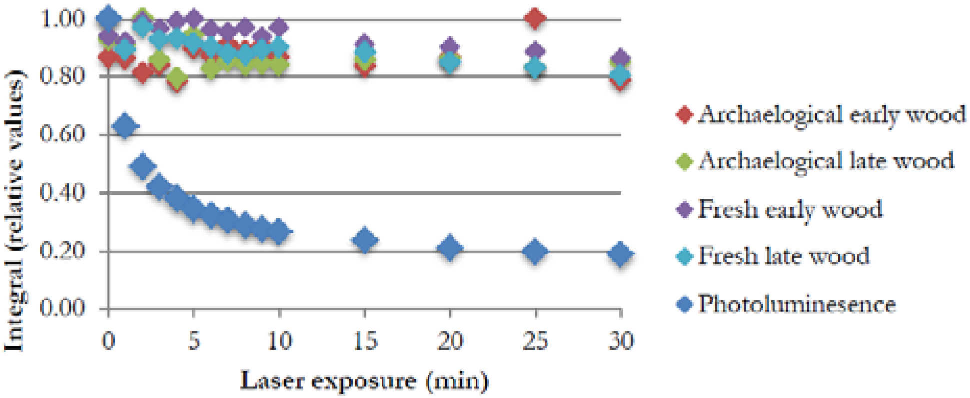


Figure 2.4. Integrals of Raman band of lignin at 1566-1637 cm^{-1} in early AW (red), late AW (green), early RW (purple), and late RW (light blue). Blue is the average integral underneath the Raman band for all four samples, that is a measure of the PL.

Integrals of Raman bands acquire from AW and RW samples were calculated and subjected to a Student's t-test (Table 2.1). A decrease of Raman signals was observed for RW, but not for AW. This could be due to the absence of some of the extractives in AW, such as stilbene[219]. It is therefore likely that it is not the signal of lignin that has been deteriorated, but rather other aromatics that absorbed the light at the incident wavelength of 633 nm.

The photoluminescence decayed exponentially. After 2 minutes, the intensity of the Raman signal was halved, and after 15 minutes the curve flattened at ca 1/5 of the original signal intensity. However, since 633 nm is at the end of the absorption curve for lignin, it is possible that some photochemical degradation of lignin may have occurred. Yet, the photobleaching performed here did not affect the Raman intensity for lignin in a way that would have undermined this study.

Sample	**p-value**	**Slope (intensity%/min)**	**R^2**
AW early	0.579	-	-
AW late	0.229	-	-
RW early	$1.81 \cdot 10^{-4}$	-0.399	0.703
RW late	$2.74 \cdot 10^{-5}$	-0.416	0.781

Table 2.1. Regression statistics for the change in the lignin band (1566-1637 cm^{-1}) over time. The p-value represents the Student's t-test at 95 % confidence interval.

PCA OF THE CALIBRATION SET

To assess the spread in the spectral responses, a PCA model was calculated using the calibration set. An examination with PCA prior to PLS-based modelling can be used as a qualitative validation for the PLS models to come[229]. The PCA model resulted in 8 components ($R^2X = 0.995$ and $Q^2 = 0.993$). The individual components present in the mixture design separated into separate loading vectors (**Figure 2.5**). The first two loading vectors **p1** and **p2** contained peaks of PEG (1141, 1280, and 1478 cm^{-1})[230], **p3** represented MLW (1273, 1600, and 1658 cm^{-1})[220], and **p4** showed the spectral features of cellulose (1094, 1122, and 1376 cm^{-1})[220].

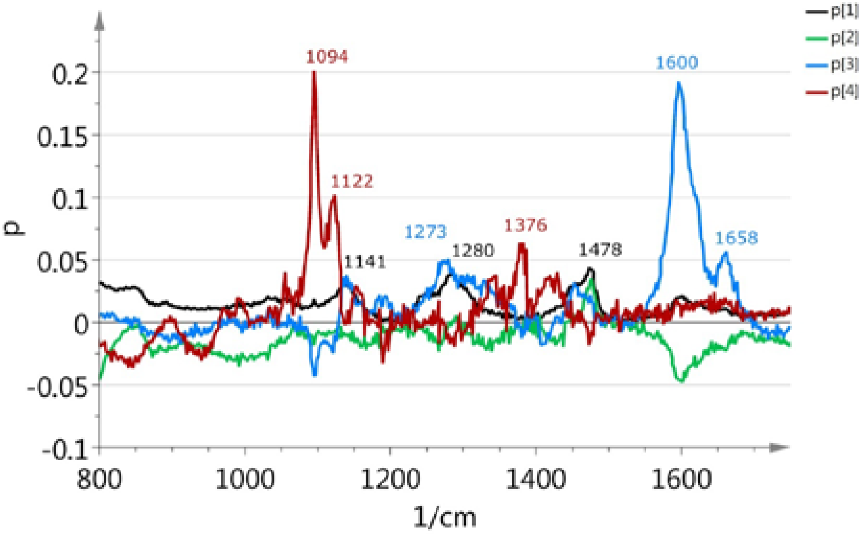


Figure 2.5. PCA loadings generated from the calibration set. The loadings **p1** and **p2** (black and green, respectively) show spectral features of PEG, while **p3** (blue) describes MWL, and **p4** (red) describes cellulose.

The remaining loading vectors **p5**-**p8** feature some peaks of the factors in the design, but were considerably noisier (results not shown). Thus, loadings under a certain R^2 value were not considered.

There was a large spread between replicate measurements, but a smaller spread between replicated mixtures, suggesting that there was a higher degree of similarity between replicated mixtures but that individual measurements varied more. It is likely that 5 replicates were too few to accurately represent the content variation in the standard mixtures. The largest spread was caused by three replicates of the 100 wt% PEG standards. Upon inspection of the original spectra it was revealed that part of the variation stemmed from the variation of the CaF_2 (the microscope slide) band at 320 cm^{-1}. It was concluded that the placement of the focal plane and consequently the focal volume was the origin of the observed variation. Since it was a variation unrelated to PEG content, it was expected that subsequent OPLS modelling should sort it into orthogonal components.

OPLS OF THE CALIBRATION SET

OPLS is applied favourably to designed experimental spaces such as mixture designs[37], since the general variations can be separated into variance correlated to the Y matrix, while any variation that is not correlated to Y matrix is instead described in the model's orthogonal components[48]. As neither

MWL nor cellulose was of interest when quantifying in RW and AW, only PEG concentrations were included into the Y matrix.

The OPLS model with best predictability was achieved by applying unit variance (UV) scaling in the spectral range of 350-3200 cm^{-1}, not including the band of CaF_2 at 320 cm^{-1}. Spectral artefacts were detected at 479-510, 610-670, and 700-790 cm^{-1}. These were ascribed to the variation caused by the RCF background correction, and as they varied systematically with PL, they were considered to be noise, and thus were removed from the model.

The OPLS model consisted of 1+4+0 components (predictive + orthogonal in X + orthogonal in Y) with $R^2X = 0.962$, $R^2Y = 0.902$, and $Q^2 = 0.758$. The predictive loading vector p and the first orthogonal loading vector po1 are plotted in **Figure 2.6** for the spectral region 800-1700 cm^{-1}.

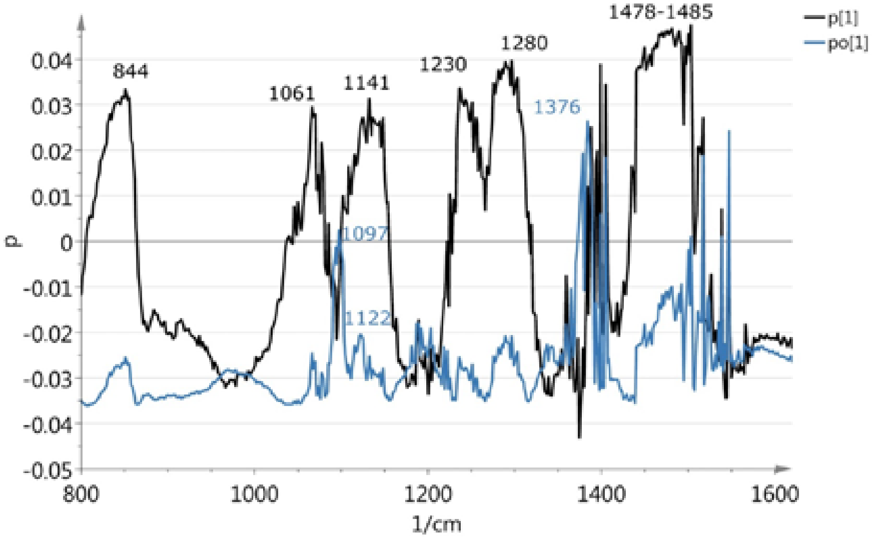


Figure 2.6. OPLS loadings of the calibration set. The predictive loading **p** (black) showed the bands of PEG (844, 1141, 1230, 1280, and 1478-1485 cm^{-1}). The orthogonal loading **po1** (blue) described the bands for cellulose (378, 1097, 1122 and 1376 cm^{-1}), i.e. variation unrelated to PEG.

The reader needs to be reminded that the loading vectors do not represent Raman spectra, but the weight of each wavenumber in relation to a component. An absolute high value means therefore that the wavenumber has higher importance (leverage) in describing variation in the predictive loading vector, while low values represent low correlation. The 'noisy' appearance of the loadings in **Figure 2.6** was caused by UV scaling.

The bands of PEG were found in the predictive loading **p** (844, 1061, 1141, 1230, 1280, and 1478-1485 cm^{-1})[230]. A broad band at 2670-3000 cm^{-1} was also observed, but as that region contains hydrogen vibrations from water, PEG, and cellulose, the signals were too confounded to separate. This region was therefore not included into the model.

The first orthogonal loading vector **po1** included the bands of cellulose (1097, 1122, and 1376 cm^{-1})[220]. The bands of lignin were less apparent, but **po1** has a weak indication of bands 1297 and 1600 cm^{-1}. However, upon the examination of the score plot of the model it became clear that observations separate according to lignin content along **to1** (results not shown), which suggests that the variation caused by lignin must have been separated into **po1**. This separation entails that cellulose and lignin could also be quantified with the method described here, however, that was outside the scope of this study.

The second orthogonal loading **po2** resembled **p**, with many of the same bands present but which were smaller and broader. This component was therefore interpreted as a representative of the intensity variation of PEG, which was uncorrelated to its concentration. **Figure 2.7** provided support for this interpretation; the spread seen along **to2** was uncorrelated to the concentration. The separation between three 100 wt% PEG standards was likely to have occured due to the placement of the focal plane in the sample, which provides further support for the above interpretation and reminds the reader that we a used confocal setup.

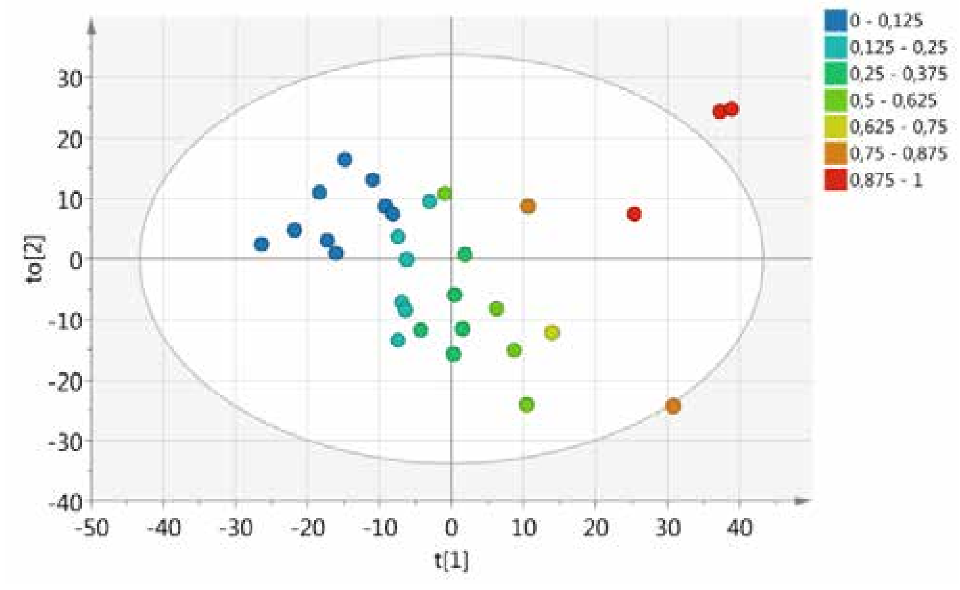


Figure 2.7. Score scatter plot of **t1** vs. **to2** of the OPLS model. Colouring according to PEG content (wt%).

The remaining orthogonal components were found to be noisier variations of the second and first components and were hence not considered further in the analysis.

PREDICTION OF RW AND AW IN THE CALIBRATION SET

The RMSE of cross-validation and RMSE of estimation were calculated to 16 and 12 wt%, respectively. The accuracy of the method was therefore not high. The $R^2Y = 0.902$ indicated however a strong linear model (**Figure 2.8**). The Student's t-test did not show any indication of systematic error.

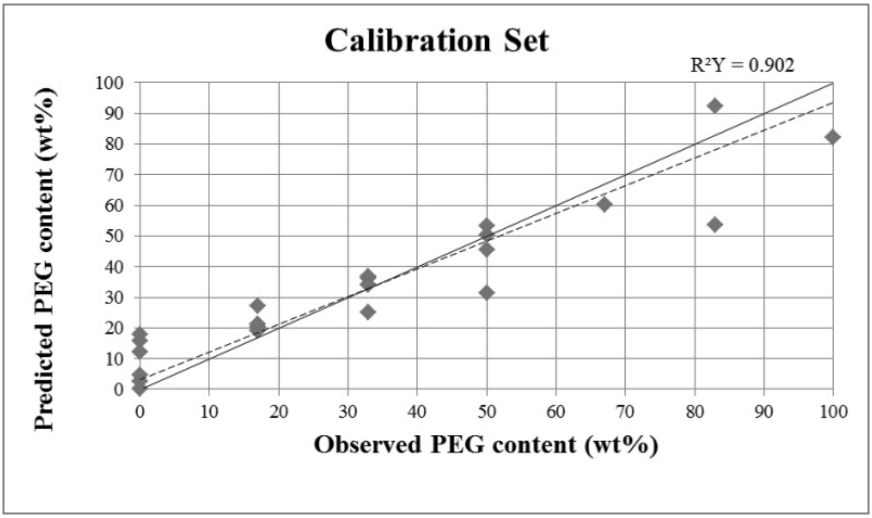


Figure 2.8. Predicted content of PEG of calibration model vs. observed PEG content of calibration standards. The dashed line shows the actual correlation, while the solid line shows the desired (i.e. perfect) correlation. Here, R^2 equals R^2Y.

As the wood constituents and PEG of the calibration set was separated into two components, the predictions performed on the validation set should also differentiate between the two components. As mentioned earlier, because the first orthogonal component predicted the content of lignin and cellulose, there is an expectation of further method development. The poor accuracy of the calibration model could in part be attributed to the afore-mentioned large variation between replicates.

PCA OF THE VALIDATION SET

To gain a comprehensive overview of the validation set, PCA models were calculated. The validation set consisted of responses from AW and RW. The model resulted in 5 components, with $R^2X = 0.986$ and $Q^2 = 0.989$. PEG concentration gradient was observed in the fourth score vector t4 (**Figure 2.9**).

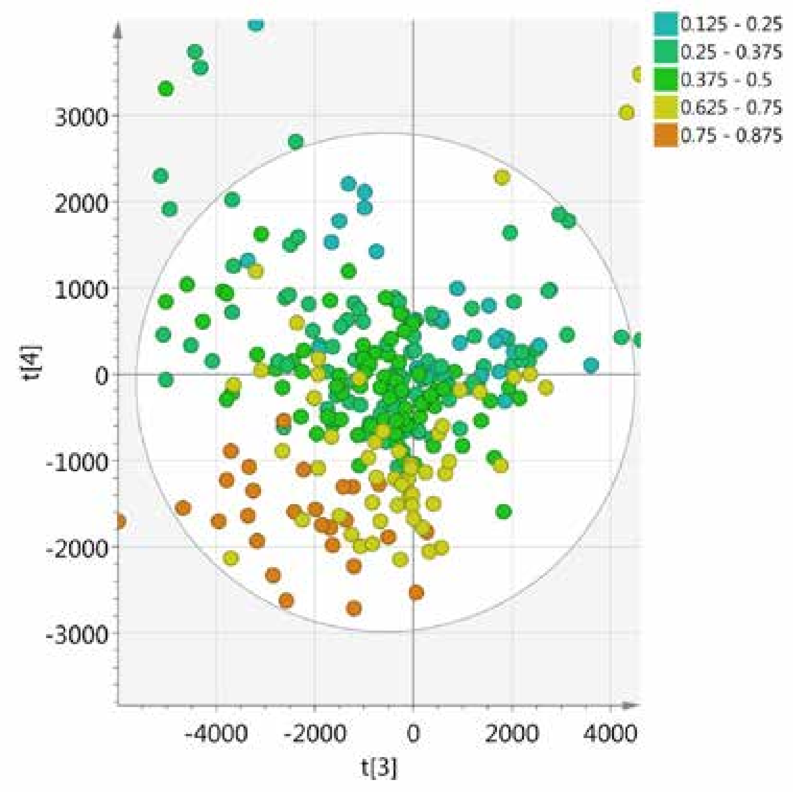


Figure 2.9.PCA score scatter plot of the validation set showing **t3** vs. **t4**. Colouring according to PEG content (wt%).

It is clear from **Figures 2.9** that the validation set showed that observations separated according to a PEG concentration gradient. This in turn indicated that the subsequent OPLS regression would be able to predict the PEG content in the validation set.

PEG content in AW and RW was determined gravimetrically to 33-75 and 21-46 wt%, respectively (**Table 2.2**). To address the narrow ranges, the AW and RW samples were tested separately and grouped. Individually, these ranges were narrow, but when grouped they spanned more than a half the design space (21-75 wt%).

Sample	**PEG ± 0.01 (mg)**	**PEG ± 0.1 (wt%)**	**Y(Pred) for PEG (wt%)**
AW 1.1	5.68	32.5	53.3
AW 1.2	10.7	74.8	81.2
AW 1.3	3.39	43.9	57.6
AW 2.1	5.07	38.4	54.5
AW 2.2	6.12	65.1	66.0
AW 2.3	4.72	66.9	62.4
RW 1	9.52	46.1	45.8
RW 2	4.38	30.3	37.0
RW 3	3.32	29.9	32.6
RW 4	3.44	20.8	32.4
RW 5	4.47	33.3	42.8

Table 2.2. PEG content determined by weight loss during MeOH extraction. Y(pred) is the predicted value for PEG (wt%) calculated by the OPLS model. The error represents the cumulative error of the weighing. The calculated error was 3.0 and 1.6 wt% for AW and RW, respectively.

OPLS PREDICTION OF THE VALIDATION SET

The validation set contained both RW and AW samples. The predicted PEG content acquired from the OPLS model was plotted against the gravimetrically determined PEG content (**Figure 2.10**). The statistics corresponding to the linear regression can be found in **Table 2.3**.

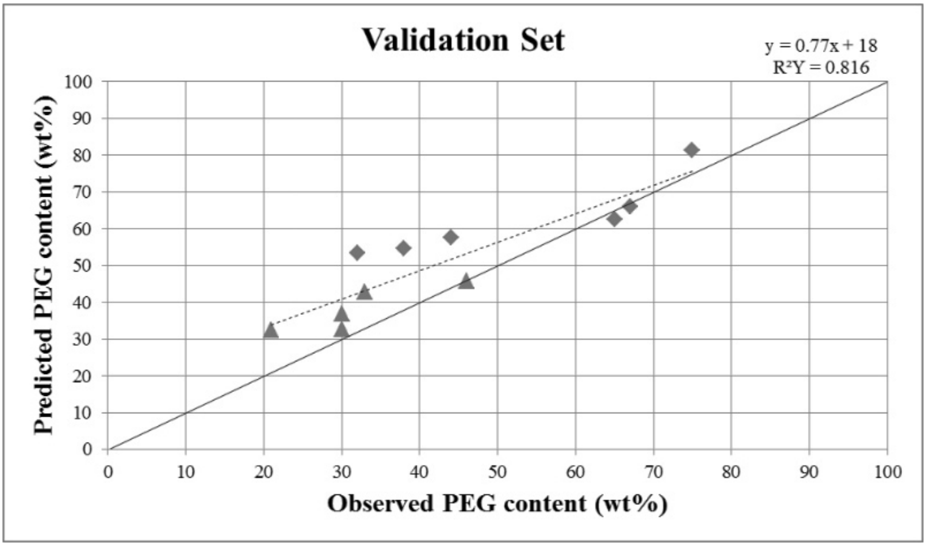


Figure 2.10 Predicted vs observed plot of the validation set. Triangles represent RW subsamples and rhombi represent AW subsamples. The solid line stands for perfect correlation and the dashed line represents a least squares linear regression for both AW and RW.

The RMSE of prediction was high, but not higher than that of the calibration set. The value was higher for AW, suggesting a higher heterogeneity of the decayed wood, as well as higher contribution from PL. The data was linear as evident from R^2Y, however, all linear regressions had a tendency to overestimate the PEG concentration. This was evident from that the intercept was significantly larger than 0, while the slope was smaller than 1. The interpretation of this was that low PEG content was overestimated, but that the systematic error was reduced at higher PEG content.

Validation set **(Number of observations)**	**RMSE of prediction (wt%)**	**R^2Y**	**Intercept (p-value)**	**Slope (p-value)**
AW (6)	13	0.795	35 (0.001)	0.52 (0.017)
RW (5)	8	0.749	20 (0.054)	0.58 (0.058)
AW + RW (11)	11	0.816	18 (0.013)	0.77 (0.000)

Table 2.3. Validation set statistics for the predicted vs. observed values. The data represents AW and RW individually, as well as together. The p-value corresponds to the value of the Student's t-test at 95 % confidence interval.

The discrepancy in PEG prediction observed between wood samples of the validation set and the standard mixtures of the calibration set could have been caused by directionality of the microfibrils in the sampled wood. While the microfibrils in the cellulose standards of the calibration set were probably random (as the cellulose was simply suspended in a liquid), in wood the microfibrils are in contrast highly ordered within the cell walls[206]. Since not all Raman vibration modes are accessible from all directions, the orientation of the microfibrils ought to have influenced the signal[231]. Therefore, the PEG overestimation resulted from the fact that less cellulose was detected in the wood. This was further supported by the fact that as cellulose became less important, the overestimation also declined.

CONCLUSIONS

Quantification of PEG content in AW and RW wood was successfully achieved through OPLS modelling in which a designed calibration set (MLW, PEG, and cellulose) was used. The validation performed suggests a reliable method and the results presented herein show great promise for future development of an *in situ* methodology for PEG quantification in wood. It was also shown that OPLS calibration could be used to predict contents of wood (MLW and cellulose) and that it could discriminate between the wood component and the PEG. However, the current method suffers from low accuracy.

The conclusion was that the overestimation of PEG was caused by the high organisation of directionality of microfibrils in the cell walls of the wood samples compared to the random organisation in the calibration samples of cellulose. Gierlinger et al.[231] suggested modelling the intensity loss of cellulose in OPLS. This would allow for correction of the accuracy of PEG calibration model by using the predictive component from such a model. The issue with microfibril directionality could also be attributed to that confocal Raman setup was used: the measured focal spot was much too small to represent an average signal from the wood matrix. This shortcoming can be ameliorated if a non-confocal Raman setup is used. For instance, using Raman instrumentation equipped with optical fibres would enable measurement of a bigger focal area, thus averaging the signal from a larger sample spot. In addition, using such a setup would also completely eliminate the destructive aspect of the method developed by us.

It is likely that the calibration set model can predict the content of MLW and cellulose apart from its ability to predict PEG content. This was stipulated upon examination of the orthogonal component **po1**, which could predict lignin and cellulose content. If such a model were to be developed it could provide a detailed insight into the extent of PEG impregnation as well as into the state of the decay in archaeological artefacts. This would add a further edge when selecting the of best conservation treatment.

The method presented here could also be utilized when monitoring the impregnation of archaeological wood with other consolidation agents, as long as its spectral features are dissimilar enough to wood constituents. In addition, the method may be adapted to other types of wood provided the source of MWL is changed. In the case of Norway spruce and in a majority of softwoods, the main constituent of lignin is coniferyl. In hardwoods, on other hand, the main lignin constituent is sinapyl. Therefore, the best calibration can arguably be achieved by using MLW extracted from the same species as the one being analysed.

APPENDICES FOR CHAPTER 2

APPENDIX 2.1 – RAMAN SPECTRA OF CALIBRATION AND VALIDATION SETS

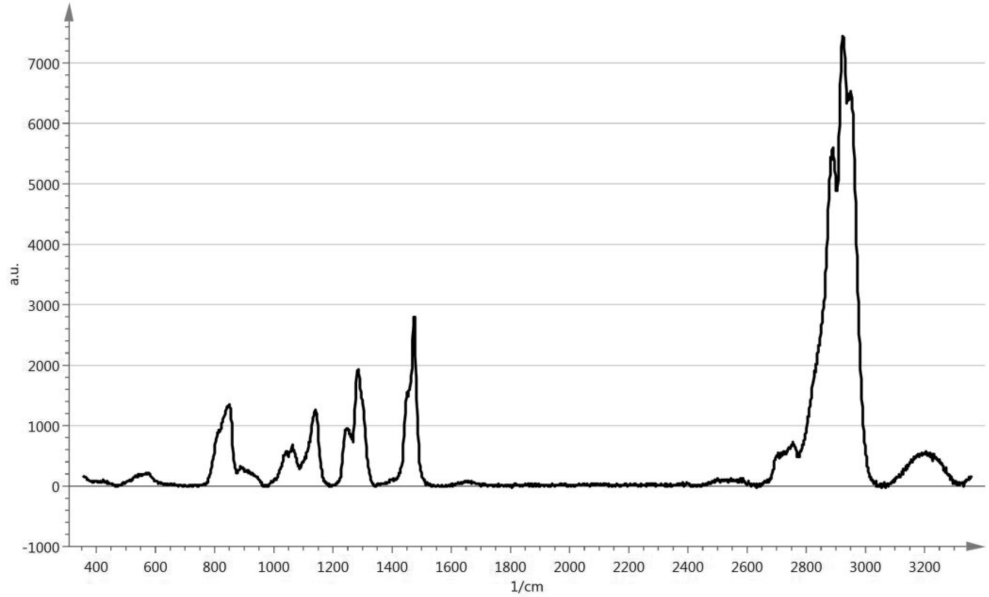


Figure 2-A. Untreated Raman spectrum of aqueous polyethylene glycol (PEG).

Band (cm^{-1})	**Assignment**	**Band (cm^{-1})**	**Assignment**
581	v(C-C)[232]	**1396**	w(CH_2), v(C-C)[230]
844	r(CH_2)[230]	**1478-1485**	δ(CH_2)[230]
860 (shoulder)	r(CH_2), v(CO)[230]	**1445 (shoulder)**	δ(CH_2)[230]
1061	r(CH_2), v(CO)[230]	**1459 (shoulder)**	δ(OCC)[230]
1141	v(CO)[230]	**2890**	v(CH_2)[232]
1125 (shoulder)	v(C-C), w(CH_2)[230]	**2847 (shoulder)**	v(CH_2)[232]
1230	t(CH_2)[230]	**2942 (shoulder)**	v(CH_2)[232]
1280	t(CH_2)[230]	**2783 (shoulder)**	v(CH_2)[232]
1362	w(CH_2), v(C-C)[230]		

Table 2-A. Peak assignment, where v = stretching, t = twisting, r = rocking, δ = bending[230, 232].

Figure 2-B. Untreated Raman spectrum of wetted cellulose powder.

Band (cm^{-1})	**Assignment**	**Band (cm^{-1})**	**Assignment**
353	Cellulose[220]	**1094**	v(CC), v(CO)[233]
381	Cellulose[220]	**1122**	Cellulose[220]
436	Cellulose[220]	**1337**	v(CC), v(CH$_2$)[233]
457	Cellulose[220]	**1376**	δ(HCC), δ(HCO), δ(HOC)[233]
520	Cellulose[220]	**1461**	δ(HCC), δ(HCO)[233]
901	δ(HCC), δ(HCO)[233]	**2740**	Cellulose[220]
1059	v(CC), v(CO)[233]	**2895**	v(CH), v(CH$_2$)[220]

Table 2-B. Peak assignment of cellulose powder from cotton linters suspended in water, where v = stretching, δ = bending[220, 233].

Figure 2-C. Untreated Raman spectrum of wetted milled wood lignin (MWL).

Band (cm^{-1})	**Assignment**
1273	Guaiacyl ring[233]
1336	Syringyl ring[220, 234]
1453	Guaiacyl ring[233]
1598	Aryl-ring[233]
1658	Coniferyl aldehyde[233]
2945	Stretching of OCH_3[233]
3072	Lignin[220]

Table 2-C. Peak assignment of MWL in water[220, 233-234].

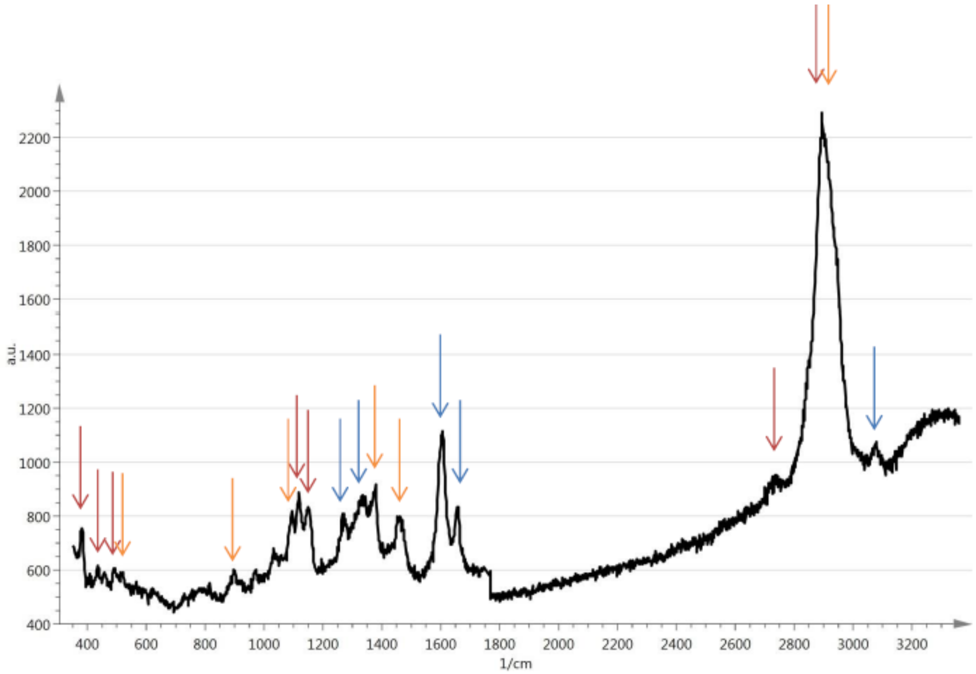


Figure 2-D. Untreated Raman spectrum of archaeological wood (AW). The sample contained signals from lignin (blue arrows), cellulose (red arrows), and PEG (orange arrows).

APPENDIX 2.2 – THE MIXTURE DESIGN

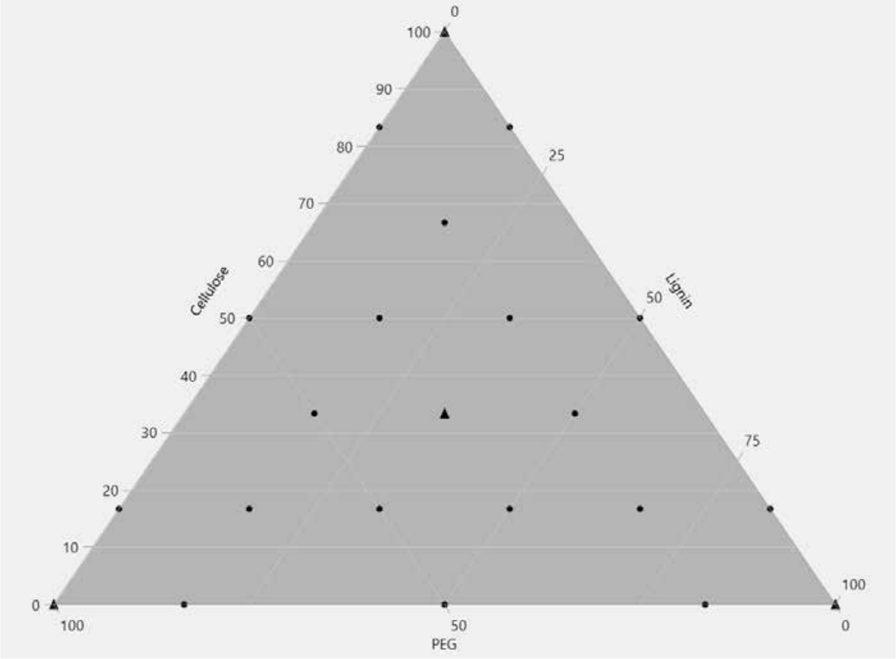


Figure 2-E. Augmented simplex lattice design with 7 levels of the calibration standards: cellulose powder, PEG, and MWL. Triangles and dots represent triplicates and signlicates, respectively. The percentage is presented as wt%.

Run	PEG (mg)	Lignin (mg)	Cellulose (mg)	Run	PEG (mg)	Lignin (mg)	Cellulose (mg)
1	33	50	17	**19**	17	50	33
2	17	0	83	**20**	50	0	50
3	33	33	33	**21**	17	33	50
4	0	0	100	**22**	50	50	0
5	100	0	0	**23**	17	67	17
6	0	100	0	**24**	50	33	17
7	17	17	67	**25**	33	33	33
8	83	0	17	**26**	0	17	83
9	33	17	50	**27**	17	83	0
10	0	50	50	**28**	0	100	0
11	0	83	17	**29**	100	0	0
12	50	17	33	**30**	0	0	100
13	0	0	100	**31**	10	10	80
14	100	0	0	**32**	80	10	10
15	0	100	0	**33**	10	80	10
16	83	17	0	**34**	50	25	25
17	33	33	33	**35**	25	50	25
18	67	17	17	**36**	25	25	50

Table 2-E. Augmented simplex lattice design of calibration standards listed according to the randomized analysis sequence. The SD of the measurements standards was ± 1 m

PART II

APPLICATION OF MACHINE LEARNING METHODS FOR ANALYSIS OF PRODUCTION OF BIOGENIC VOLATILE HALOCARBONS

INTRODUCTION

MACHINE LEARNING IN THE ANALYSIS OF ENVIRONMENTAL PHENOMENA

Machine learning, especially where it focuses on multivariate methods, is not uncommon in environmental sciences[235-236]. The intrinsic multi-variability of environmental data can stem from three main sources: natural sources affected by geographical and temporal influence due to natural phenomena, anthropogenic sources that interfere with natural variability, and experimental errors originating from (for example). sampling and data evaluation[236]. PCA and DA are examples of tools used extensively in environmental analysis[74, 236]. PLS is used as well, although to a lesser extent, while the application of OPLS and statistical DoE appears to be uncommon.

One area of interest in the scope of this thesis was the biogenic production of volatile halogenated organic carbons (VHOCs) by marine algae. Halocarbons have been a subject of scientific scrutiny for several decades. For example, one of the unknowns of early scientific inquiry was the total concentrations of stratospheric bromine. Early investigations into its nature and behaviour revealed that the total amount of inorganic bromine radicals in the stratosphere exceeded the expected concentrations by 3.7 ppt. This was surprising as it had been assumed that the largest contributions of inorganic bromine was mainly due to anthropogenic sources such as biomass burning[237]. However, those early results indicated that there must be other sources of inorganic bromine than had been previously believed. Since then, a better understanding has been achieved with regard to the sources that contribute to VHOC species in the atmosphere. In addition to anthropogenic emission, there are some terrestrial sources such as volcanic eruptions, forest fires, and marshes. However, the anthropogenic contributions of brominated halocarbons to the atmosphere are not significant on a global scale compared to natural sources[238]. Among the natural sources of VHOCs – especially iodinated and brominated ones – are marine algae

Many of the machine learning challenges were highlighted here, such as the analysis of **nonlinear**[236] multivariate systems and the issue of **non-selectivity**. However, the primary aim of the chapters in this part of the thesis was to fill in the existing **knowledge gaps** with regards to VHOC production by drawing on linear machine learning methods. Many knowledge gaps exist in this area of research, many of which were identified in a recent publication by Keng et al.[239]. The most relevant of those knowledge gaps for the scope of this thesis are listed below:

- Too few studies about the diurnal and seasonal variations in environmental factors, and research focused on a longer exposure to environmental change. Such enquiries would provide an insight into biological adaptation in marine algae and how such adaptations influence VHOC emissions.

- Multifactor studies to determine the interactive effects of environmental change on algal VHOC emissions.

- A more standardised approach to VHOC research is necessary to make studies in the field more comparable.

Compared to many other environmental phenomena such as emissions of polycyclic hydrocarbon (PAH), the production of VHOCs using multivariate approaches has largely not been. This may seem odd considering that VHOC emissions, like many other natural phenomena, are influenced by a multitude of factors. One of the exceptions is the study done by Granfors et al.[240], where OPLS-DA was applied to the evaluation of the differences in VHOC production in newly formed ice and older ice. Univariate enquiries have been made into the influence of factors such as exposure to light, pH, and temperature, affirming their impact on biogenic VHOC production. The issue that remains unaddressed, however, is whether environment factors interact and how such interactions influence the biogenic VHOC production[239, 241]. The need for insights into multifactor interaction is reflected in the first two points in the list above. To accommodate this lack of knowledge, DoE was used to create a design space where several factors previously proven to influence algal VHOC production are varied simultaneously and then qualitatively studied by means of PCA, OPLS, and OPLS-DA. The results of this endeavour are reported in Chapter 4. Moreover, the author believes that the methods used in Chapter 4 may provide a starting point towards standardized procedures in research regarding VHOC emissions (see the last point in the list above).

An additional knowledge gap that has been identified in this thesis is the localisation of enzymes involved in VHOC production within their algal matrix – where in the algae are those enzymes located? The lack of information regarding the enzymes' localisation is an issue that influences other, more general enquiries about VHOC production. As will be discussed later, the gene expression for those enzymes, and, possibly, their localisation in the algal tissue, changes in response to environmental stressors. This **localisation** challenge was considered in Chapter 3, where DoE and several multivariate linear algorithms were used in an attempt to acquire a distinct spectroscopic signal of the enzyme responsible for the biogenic VHOC production. Raman spectroscopy, confocal Raman spectroscopy, and SERS were used as analytical methods in that chapter. For theory on these methods, see the introduction of Part I. Chapter 3 can be seen as connecting Part I and Part II, as it focuses on a knowledge gap that exists in VHOC research, but which was examined using methodologies similar to those in Part I.

THE CHEMISTRY OF VHOCS

VHOCs$^{\nabla}$ are volatile and short lived species with atmospheric lifetimes varying from a couple of minutes to several months[244]. Bromocarbons and especially iodocarbons have the shortest lifetimes in the atmosphere. All VHOCs are characterized by having one to four carbon atoms and at least one halogen atom. However, the most common compounds are methane derivatives. In this thesis, compounds examined are listed in **Table 1**.

VHOC	Formula	Atmospheric lifetime (days)
Chloroiodomethane	CH_2ClI	0.1
Dibromomethane	CH_2Br_2	123
Dibromochloromethane	$CHBr_2Cl$	59
Bromoform	$CHBr_3$	24
Diiodomethane	CH_2I_2	0.003
Iodomethane	CH_3I	7
Dichlorobromomethane	$CHCl_2Br$	-
Bromoiodomethane	CH_2BrI	0.04

Table 1. List of studied VHOCs. Atmospheric lifetimes are listed as in Montzka et al.[244]. The atmospheric lifetime for $CHCl_2Br$ was not reported by Montzka et al.

Global distributions of VHOCs are generally the highest at the sites of phytoplankton blooms, coastal zones, oxygen-minimum zones, and upwelling regions[245-246]. These geographical locations have especially high VHOC emissions during the summer months[247].

The halogen constituents influence the halocarbons' reactivity – chlorinated species are the most stable and the iodinated species are the most reactive. Brominated halocarbons have been of particular interest since these species have sufficient stability to reach the atmosphere while at the same time being rather short-lived. In **Table 1** the atmospheric lifetimes of some the investigated naturally produced VHOCs are listed. CHB_3 and CH_2I_2 are considered the most interesting of naturally produced bromide VHOCs, as they contribute to the atmosphere the majority of halogen species.

Through a variety of mechanisms, VHOCs degrade into inorganic halides and other reactive species, mostly through photochemical degradation. The majority of the reactions involving

---

$^{\nabla}$ To avoid confusion, the reader should note that the research presented in Part II only concerns the production of alkane-derived VHOCs, which are just one type of volatile halocarbons. Other volatile halocarbons also exist, for instance, volatile aldehydes[242-243], but are not covered in the scope of this work. Henceforth, unless specified otherwise, the author refers to 'halocarbons' or 'VHOCs' as meaning alkane-derived VHOCs. Such VHOCs are defined by their having atmospheric lifetimes shorter than half a year[244].

biogenic VHOCs happen in the troposphere. The formed radicals and inorganic volatiles are then involved in a wide range of reactions, including oxidation of mercury, cloud formation, ozone destruction, methane destruction, and several climate-relevant atmospheric processes (e.g. NO_x and oxidation of sulphur(IV))[248-256].

The contribution VHOCs have to the troposphere and stratosphere indicates that further studies should take place with the main aim of elucidating the mechanism behind their production and by which parameters this production is affected.

VHOC PRODUCTION

In marine environments, the emissions of VHOCs have both abiotic and biotic sources. The biotic emissions involve a wide range of marine taxa, including micro- and macroalgae and bacteria (**Figure 13**).

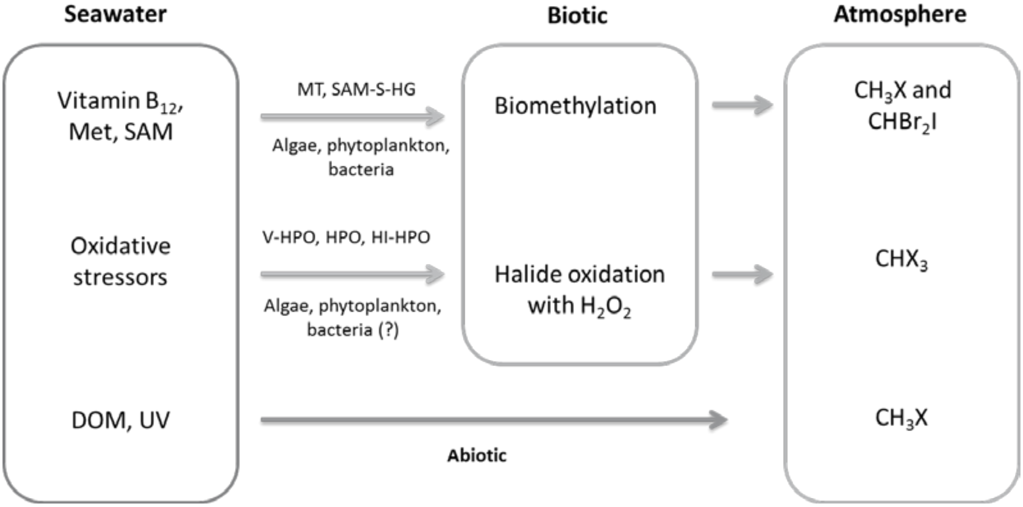


Figure 13. A schematic illustration of some of the known biotic and abiotic processes that contribute to VHOC emissions. Met = methionine; SAM = S-adenosyl-methionine; DOM = dissolved organic matter; MT = methyltransferase; SAM-S-HG = S-adenosyl-methionine-dependent methyl-halogenase; V-HPO = vanadium dependent haloperoxidase; HPO = cofactor-free haloperoxidases; HI-HPO = heme iron dependent haloperoxidase.

The involvement of biotic and abiotic sources in VHOC emissions is therefore a complex process which connects the production, degradation and transport of VHOCs in marine environments[257]. The reader should note that the separation between the definition of biotic and abiotic production is not clear cut. Biogenic or biotic production may refer to the primary production of VHOCs, i.e. that enzymatic processes are necessary for their production. The biogenic production may also be secondary in the sense that products produced by enzymatic reactions react with something else to produce VHOCs. Therefore, some secondary biogenic production mechanisms are called abiotic in the literature.

Biotic production of VHOCs occurs in a significant number of organisms and the reason behind this production varies between them. In macroalgae, VHOCs may functions as antioxidants[258], while cyanobacteria[259-261] and phytoplankton[262-263] may use VHOC production as a way to excrete surplus of halide ions[264-265].

The biotic production or halogenated organic compounds is chiefly associated with six different enzymes families. One class of enzymes – vanadium dependent haloperoxidases (V-HPOs) – is examined in Chapter 3. Other halogenating enzymes include cofactor-free haloperoxidases (HPOs), nonheme iron-dependent halogenases (NI-HG), S-adenosyl-methionine-dependent methyl-halogenase (SAM-S-HG), heme iron dependent haloperoxidases (HI-HPOs), nonspecific methyltransferases, and flavin-dependent halogenases (F-HG)[266]. The difference between halogenases and haloperoxidases is that the latter oxidises halides with the aid of H_2O_2, while the former depends on other co-substrates, such as Fe(II) and flavin adenine dinucleotide hydroquinone ($FADH_2$)[267]. Since algae have been the subject of research in this part of the thesis, the relevant enzyme reactions involve V-HPOs, SAM-S-HG and nonspecific methyltrasferases. HPOs and HI-HPOs are considered, but only in relation to results in Chapter 4. These enzymes may be behind VHOC emissions from alga-associated bacterial communities and may have contributed to the results observed in Chapter 4.

PRODUCTION BY ALGAE

Algae produce both stable and volatile halogenated compounds, all of which play a role in their defence and function. Production of several non-VHOC halogenated volatiles by algae and cyanobacteria has also been connected with allelopathic effects, i.e. the release of halogenated species influences the populations of other organisms through the disruption of, for example, their growth and reproduction[268]. Another example is the presence of hormone-like derivatives which are involved in cell-to-cell signalling and which store organic iodine, monoiodotyrosine and diiodotyrosine. Unlike many other organohalogens, these iodinated tyrosines are formed spontaneously, without enzymatic catalysis[269-270] and do not seem to contribute to the formation of VHOCs. Apart from iodinated hormone derivatives, iodine metabolism on the whole seems to be closely connected to V-HPO activity in both macroalgae and phytoplankton[264, 271-273]. Other halogenated non-volatile halogenated species in algae include phloroglucinols, phlorotannins, polysaccharides, terpenes, and fatty acids[269].

To date, the only known enzymatic mechanisms in algae[∅] specifically involved in the production of alkane-derived VHOCs are the production of monosubstituted halomethanes by SAM-S-HG, nonspecific methyltransferases generating di-and polysubstituted halomethanes, and formation of haloforms assisted by V-HPO-generated HOX[274-276]. The purpose of SAM-S-HG and methylation mechanisms is to ensure rapid antioxidant action by detoxifying ROS'. This is also supposedly done to regulate apoplastic reserves of iodine and bromine[277].

The formation of monosubstituted halomethanes (also named as methyl halides and monohalogenated methanes) by SAM-S-HG is shown in the equations below (**Equations 11** and **12**, where X is a halide):

$$CH_3RS^+ + CH_3 \rightarrow CH_3SR + CH_3^+ \qquad \text{(Eq. 11)}$$

$$CH_3^+ + X^- \rightarrow CH_3X \qquad \text{(Eq. 12)}$$

The formation CH_3X is made possible by the donation of CH_3^+ from sulfonium compounds (CH_3RS^+) such as S-adenosyl-L-methionin[275] and dimethyl-b-propionthetin[278].

Di- and polysubstituted volatile halocarbons such as dibromoiodomethane ($CHBr_2I$) are produced by a different mechanism that likely involves methylation by non-specific methyltransferases[279]. The mechanism of non-specific methyltransferases is hypothesised to involve a successive halogenation of enols at C=O groups, followed by the loss of C1. This produces secondary products, such as CH_2X_2 and CHX_3[280]. Manley[275] has called this mechanism a "biochemical accident" as methyltransferases lack substrate specificity. Because of this non-specificity, a plausible hypothesis has been suggested. This is that if methyl halides are indeed synthesised by nonspecific methyltransferases, their production should be widespread among different types of organisms and may be related to the availability of necessary methyl donors. This has been confirmed by finds of methyltransferases not only in algal taxa, but also in marine bacteria[281] (see the section **Production by bacteria** for more). An alternative proposition is that the enzymatic catalysis involved in di- and polyhalogenation should be attributed to V-HPOs[282], but it is not clear whether V-HPOs are involved as primary or secondary biogenic producers of these VHOCs.

Another halogenating mechanism in algae involves the oxidation of halides by H_2O_2 catalysed by V-HPOs. In plant cells, H_2O_2 originates from two main sources. Intracellular H_2O_2 originates predominantly as a by-product of photosynthesis, i.e. photorespiration and thylakoid electron transport chain. The other sources are extracellular and stem from the apoplasts, where the H_2O_2 pool is generated by the cell wall, by oxidases located in the plasma membrane, and by secretory peroxidases[283]. In seaweeds, nicotinamide adenine dinucleotide

---

[∅] It should be stressed that the consensus surrounding these formation mechanisms has been established only for brown macroalgae, but the assumption is that this is true for other phyla.

phosphate (NADPH) oxidase is understood to be a contributory factor unlike secretory peroxidases, which are not present in seaweeds[284-285].

Haloforms (CHX_3) seem to be a result of secondary biogenic production by V-HPO, as they form as a result of reactions of HOX with organic compounds in seawater. In addition, haloforms form as an abiotic decay product from other VHOCs. Nevertheless, algae from temperate, polar, and tropical regions have been identified as greater contributors of $CHBr_3$ to the atmosphere compared to other sources of the haloform[286-287].

The production of VHOCs is strongly species and phyla dependent. Brown, green, and red algae have all been seen to contribute to VHOC emissions. Of those, brown algae are considered to be the most active producers, although this stipulation probably results from the fact that more studies have been performed on brown algal species[239]. Production within a phylum varies between species. For example, an order of the brown algae, Fucales (e.g. *Fucus* and *Ascophyllum* genera) has much higher bromination activity compared to their iodination activity. For Laminariales, which are also brown algae (e.g. *Laminaria* and *Saccharina* genera) the situation is reversed[288-290]. The variation also depends on the geographical location. VHOC emissions are generally lower in polar species of algae than in temperate and subtropical species[291-295].

Vanadium Dependent Haloperoxidases (V-HPOs)

V-HPOs have been found in terrestrial organisms, such as the fungus *Curvalaria inaequalis* and the lichen *Xanthoria parietina*, but they have proved to be most abundant in a number of marine species, such as the flavobacterium *Zoebellia galactanivorans*, and, most notably in the vast majority of marine algae. V-HOPs have been isolated from red (Rhodophyta), brown (Phaeophyta), and green (Chlorophyta) macroalgae and from several species of microalgae living both in open oceans and in the sea ice. Despite not being associated with extremophile organisms, V-HPOs are remarkably stable both *in vivo* and *in vitro*. Their resistance to high temperatures, to the presence of detergents, organic solvents, and high levels of reactive oxygen species (ROS) such as H_2O_2[286, 296-297], makes V-HPO attractive for potential applications in industry[288].

The exact function of V-HPOs is a matter of speculation. Some have suggested that V-HPOs assist in the management of a surplus of halides[298], while others have attributed their VHOC production to antibacterial and antifouling activity[265, 268, 275, 286, 299-300]. In macroalgae, the reactive intermediates produced by V-HPOs seem to facilitate the cross-linking of phlorotannins and alginates both of which are necessary for the attachment of macroalgal zygotes to the sites of growth[301]. Another hypothesis, which at the time of writing is the most prevalent one, is that V-HPOs function as protectors against oxidative stress by acting as reactive oxygen species (ROS) scavengers[275, 302-303]. The truth may lie somewhere in between the two and is depend on the organisms in which V-HPOs are present.

All V-HPOs are distinguished by the presence of the vanadium (V) cofactor at their active site and catalyse the oxidation of halides by hydrogen peroxide, H_2O_2. At the time of writing, three algal V-HPO structures have been fully characterised. The first structure was determined for *Acsophyllym nodosum*[304], and it was later followed by structures for V-HPO from two red marine macroalgae, *Corallina officinalis*[305] and *Corallina pilulifera*[306].

The V-HPO-catalysed oxidation of halides by H_2O_2 produces reactive intermediates, such as hypohalous acids (HOX, where X is a halogen) (**Equation 13**).

$$X^- + H_2O_2 + H^+ \rightarrow HOX + H_2O \qquad \text{(Eq. 13)}$$

Due to their reactive nature, these intermediates react with nucleophilic acceptors, a reaction which in turn generates a wide variety of halogenated organic species, including alkane- and alkene-derived VHOCs[307]. In spite of V-HPOs having been known to science for several decades, there is currently no thorough understanding as to how the reactions occur at their active sites[288]. The initial step of the catalytic cycle involves the coordination of H_2O_2 to the vanadate, which forms a stable peroxo-vanadate intermediate[308-310]. As the redox state of the vanadium does not change during the turnover, it has been proposed that the vanadium cofactor activates H_2O_2 by acting as a Lewis acid[311]. The following step, the oxidation of the halide, remains unclear. Possibly, one of the oxygens in H_2O_2 is protonated which allows the initial halide oxidation. Then, this is followed by a nucleophilic attack of the halide, which breaks the peroxide bond and forms HOX and XO^-. There are no indications to suggest that the organic substrates that react with HOX are bound to V-HPOs active site. A kinetic study indicated that this catalytic cycle of V-HPOs follows a bi-bi ping-pong mechanism[312]; however, this was reported only for one of the three possible V-HPO subclasses.

V-HPOs are divided into subclasses based on the most electronegative halide they can oxidise. However, the origin for this halide selectivity in the V-HPO family is uncertain[312-313]. Chloroperoxidases (V-ClPOs) have so far been found in a variety of terrestrial organisms, while V-BrPOs and V-IPO are prevalent in marine species. V-ClPO can oxidase all three halides (Cl^-, Br^- and I^-). V-ClPOs are inhibited by chloride at lower pH, suggesting a competitive mechanism, while at high pH the enzyme follows Michaelis-Menten kinetics[314]. In a difference between the two other V-HPO variants, V-ClPOs' active site seems to contain phenylalanine instead of a second histidine residue. Bromoperoxidases (V-BrPOs), which oxidize Br^- and I^- have been attributed the above mentioned bi-bi ping-pong mechanism. Iodoperoxidases (V-IPOs) can only oxidise I^-, and have lower affinity for the halide substrate but a higher rate of oxidation rate than V-BrPO[289].

The decades of studies dedicated to V-HPOs have amassed a large quantity of data. The data from any one source cannot for the most part be compared with the data from any other source. As Tarakhovskaya et al.[290] point out, detailed studies of the enzyme have been predominantly focused on those that were isolated from a few brown macroalgae species, and the protocols used vary greatly between different studies. From the analytical point of view, these protocols are also poorly validated. Thus, there remains a considerable **knowledge gap** regarding V-HPOs, their interspecies variation, and their contribution to the global production of VHOCs both from natural habitats and from the expanding seaweed cultivation industry.

The Influence of Biotic and Abiotic Stressors on VHOC Production in Algae

Adaptation to environmental stressors has an impact of generation of H_2O_2 and other ROS[284, 315-316], thus affection the secondary biogenic production of VHOCs by V-HPOs. Iodine and bromide anions functions as antioxidants in response hydroxyl radicals generated from H_2O_2 during oxidative stress. Küpper et al.[317] suggest that the algae bind halides to store them as organohalides in order to be accessible in case of a stress scenario. It has also been reported that in response to stressors, the production of VHOCs by V-HPOs is upregulated, as well as the expression of V-HPOs in different tissues. For instance, the increase in gene expression of a V-IPO has been observed in *Laminaria digitata* subjected to predation[318]. A connection between breakdown products of algal cell walls following a bacterial attack also have been observed to oxidative stress and, in extension, increased VHOC production[319-320]. Some other stressors include changes in light condition, temperature variation, and CO_2 and nutrient limitation[321].

The fact that irradiance levels trigger increased VHOC emissions suggests that it is connected to photosynthesis, due to higher release of H_2O_2[239]. The connection to photosynthesis was shown by Manley and Barbero[322]and Goodwin et al.[323], where the addition of a photosynthetic inhibitor decreased the production of VHOCs. Photosynthesis is also impacted by changes in temperature[324-325]. Increasing temperatures of seawater has been connected to increased photosynthetic activity in some of the seaweeds[243, 326], and associated therefore with increased H_2O_2. However, the influence of temperature seems to be strongly species-dependent[239], and be a function of short-term induced stress by changed temperature[327-328]. Too little studies are published where the effect of long-term seasonal variations on algae was examined. This in turn makes it difficult to assess the impact of increasing ocean temperatures associated with climate change on biogenic VHOC production.

Changes in pH of the surrounding seawater have also been quoted to influence the performance of the V-HPOs. Within an optimum range between pH 4 to 8.3, the enzyme remains functional[279]. Upon acidification, VHOC production has been reported to increase under laboratory condition[329-330] due to induced oxidative stress, while mesocosm studies did not report any noticeable effects of ocean acidification on VHOC production[331-333]. However, at the time of writing of this thesis, the studies concerning the influence of pH on VHOC production are limited to the ones quoted here. Yet another parameter that affects photosynthesis and photorespiration in algae is the changes in salinity[334-335] – hypersaline[336] and hyposaline conditions have both been attributed to increased VHOC production. For instance, hyposaline conditions have been observed to increase the production of several iodinated VHOCs[327]. Some of the known stressors have been multivariatly explored in Chapter 4.

An increase in H_2O_2 in the surrounding seawater has been cited as a cause of increased V-HPO activity. The stage of an alga's development also influences production. A recent study by Lemensheva et al.[316] found dramatic increase of V-HPO activity during zygote development in *Fucus vesiculosus*, where the activity of brominating V-HPOs was of particular note as it was close to the activity observed in adult individuals. This was attributed to the increase of H_2O_2 during the fertilisation stage, likely due to extracellular sources.

The reader should have noticed by now that all studies dedicated to the impact of environmental parameters on algal VHOC production, are single-factor studies. In a recent review, Keng et al.[239] have pointed out the lack of studies where the influence of multifactor interactions on VHOC emissions is investigated. This **knowledge gap** is addressed by a study in Chapter 4, which describes a first-of-a-kind study about the simultaneous impact of several environmental factors on biogenic VHOC emissions.

The Localisation of the V-HPOs in Algae

V-HPOs can be found in the cytosol and the apoplast, where the latter is located outside the plasma membrane and is used by chloroplasts during photosynthesis. Their location coincides in part with the sites of H_2O_2 production in algal cells, which may explain their inherent resistance to high H_2O_2 levels[337-338]. V-HPOs contribute to the intracellular regulation of H_2O_2 in algae, and are, in fact, more effective H_2O_2 scavengers when compared to intracellular ascorbate peroxidase and catalase enzymes[290]. In some algal species, it has been shown that iodine has a decreasing gradient from the meristoderm to the apoplasm, reaching highest levels in the external cell layers (ca 80 % of the total iodine content). This iodine distribution makes it easily accessible for mobilisation against induced oxidative stress[288, 339]. A similar extracellular speciation of bromine has been observed in other brown algae, while in red algae bromine was found to be stored intracellularly[340]. As these halide storages are species dependent, the author of this thesis hypothesises that they might coincide with the location of V-HPOs, as no explicit explanation for this have been made to date.

Although the location of the V-HPOs is established on cellular level, it is still not known exactly where in the algae's thallus they can be found. Wever and van der Horst[286] have located V-BrPOs near the surface of a green macroalga. Other sources also suggest that some V-BrPO and V-IPO are present in extracellular compartments[288]. However, seaweeds are very different in their physiology making interspecies comparisons unreliable. The extent of the variation of V-HPO presence in the different thalli of algae has been elegantly demonstrated by Tarakhovskaya et al[290]. For instance, in *Fucus serratus*, the algae used in the research reported in Chapter 4, the majority of V-HPO activity was found in the middle thallus. In addition, as has been speculated by some sources cited in the above section, it is possible that stressors may induce varying expression of V-HPOs in different tissues: this makes the questions regarding the localisation of these enzymes of further interest.

To address this **knowledge gap**, enquiry discussed in Chapter 3 was dedicated to establishing a spectroscopic and chemometric method for the qualitative detection of a V-HPO. As will be shown, the Raman spectroscopic signal of the vanadium bromoperoxidase from *Corallina officinalis* (C-VBPO) was not sufficiently specific. This in turn meant that it would not be possible to spectroscopically localise the enzyme inside its algal matrix, as was the aim of research presented in Chapter 3. To circumvent this issue, the author focused instead on the products of the enzymatic reaction, i.e. on the halogenated organic compounds. The premise was that if the conversion could be detected with Raman spectroscopy, it would then be possible to track the placement of vanadium bromoperoxidases in living algae specimens. As the compounds of interest, VHOCs, are volatile, it was deemed unfeasible to develop method for tracking their production spectroscopically. Instead, the starting point for the enquiry was to use methods that would allow spectrocolourimetric detection; when the enzyme catalysed the halogenation of a dye, the dye would change colour, thus enabling the tracking of this change inside algae with normal Raman. This was indeed shown with *o*-dianisidine, where haloperoxidase activity was detected as a development of a dark reddish brown colour where *o*-dianisidine was iodinated by the haloperoxidases[274, 341]. This process could be developed further by studying the resulting staining with Raman spectroscopy by establishing a quantification method for the detection of the conversion of *o*-dianisidine to halogenated *o*-dianisidine. However, *o*-dianisidine is highly toxic and its use is regulated by the Swedish Work Environment Authority (AFS 2014:43; 47-49 §)[342].

As the University of Gothenburg actively works for the phasing-out of dangerous chemicals, other alternatives for staining were sought. An alternative colorimetric assay has been reported where a solution of isolate vanadium peroxidase enzyme from macroalgae was able to brominate thymol blue (ThB) thus converting it into bromothymol blue (BThB)[343]. The conversion was detected with UV-Vis. As it is plausible that the behaviour of an isolated enzyme is not representative of its behaviour in its natural matrix, it would be of interest to study the halogenation by vanadium dependent haloperoxidases with their matrix (the algae) intact.

Since both ThB and BThB have distinct Raman spectra[344], it was expected that a data set with good SNR would be acquired to be used in multivariate modelling .The primary aim was therefore to observe the conversion of ThB to BThB, visually and spectroscopically, and then attempt to quantify the conversion with Raman spectroscopy. In order to account for variation that may stem from several factors, the experiments were designed using an FF design where the concentrations of reactants needed for ThB-to-BThB conversion (H_2O_2, Br^-, and ThB) were varied simultaneously. The results of these experiments were inconclusive and are therefore not reported in this thesis. The issue of enzyme localisation thus remains an interesting analytical challenge.

PRODUCTION BY BACTERIA

As can be seen from the literature overview in the sections above, much of the research regarding biogenic VHOC production has been focused on macroalgae, cyanobacteria and some microalgal species. Several studies have been made of the halogenation reactions of ethers, aldehydes, esters, aromatics etc. by marine bacteria[345-346], but less research has been focused specifically on the possible contribution of marine bacteria to VHOC emissions. Halogenation processes in bacteria are attributed to HPO and HI-HPO enzymes[347]. In marine strains, however, the production of VHOC stems from different sources.

Bacteria have been reported to degrade methyl halides and also to use them for growth[348-350]. However, studies by Fujumori et al.[351] and Klein[281] indicate that bacteria are capable of producing methyl halides. Both these sources have shown that different strains of marine α-proteobacteria produced CH_3I, CH_3Cl, and CH_3Br. Additional studies have shown the production of CH_3I by marine strains[352-353]. Monohalogenated methanes were also observed to be produced in three flavobacteria, one γ-proteobacteria[281]. CH_3I was the most common among the produced monohalogenated methanes. The laboratory studies performed by Klein were also supported by long-term field observations, where it was noted that CH_3I was the most common VHOC in surface seawater during winter and spring upwelling. This production was attributed both to bacteria and photosynthesis-dependent organisms present in the near-surface layer. However, the contribution of bacteria was deemed higher.

Flavobacteria *Vibrio campbelli* was the only bacterial strain that showed a production of $CHBr_3$, albeit a low one. Similarly, a low $CHBr_3$ contribution to total emissions was observed in the field. The biochemical process behind $CHBr_3$ in bacterial synthesis is unknown, but appears to be associated with growth rate[281]. This suggests that although bacteria may contribute to $CHBr_3$ production, algae remain the chief contributors of $CHBr_3$ to the atmosphere. On the other hand, depending on the season, the bacterial contribution to CH_3I seems to be of greater importance than the algal contribution.

Apart from the contributions of algae, emissions of CH_2Br_2 have been connected to the bacterial decomposition of $CHBr_3$ through hydrolysis[354] (see the section **Bacterial degradation** below). The concentrations of CH_2Br_2 in Klein's[281] study however did not show a proportionate production of $CHBr_3$, which suggest that CH_2Br_2 from photoplankton and algae was not its only source. It has been stipulated that this could be due to presence of V-HPOs in bacteria or to the nonspecific phosphatase enzymes[281], which have a high resemblance to the active site of V-HPOs and have been shown to have a limited haloperoxidase activity[266, 355]. Although the presence of V-HPOs was not indicated in genomes of the bacterial strains studied by Klein, V-IPO from marine flavobacteria *Zobellia galactanivorans* has been successfully isolated and structurally solved[356], indicating that some species of flavobacteria may possess V-HPO activity. The writer of this thesis also speculates that the contribution may have stemmed from bacterial HPOs and HI-HPOs.

In summary, the role of bacteria in VHOC emissions is still unclear, although there are indications that they contribute to the production of iodinated VHOC. This contribution has seasonal variations and is related to the rate of bacterial growth.

OTHER FORMATION MECHANISMS

As mentioned earlier, haloforms, i.e. bromoform ($CHBr_3$) and iodoform (CHI_3), are postulated to be the result of the reaction between V-HPO-formed HOX and organic substrates. The organic substrates that are suspected to be involved in a reaction with HOX are ketones present in seaweed or dissolved organic matter (DOM)[286, 357]. The hypothesis, based on ketone involvement, comes from observations of *Bonnemaisonia hamifera*, which produced $CHBr_3$ and CH_2Br_2. Ketones supposedly decay via haloform reaction to form halogenated VHOCs (**Equation 14**) and this reaction is similar to the enol halogenations described in the sections above [278, 358-359].

$$Ketone + 3X_2 + OH^- \rightarrow CHX_3 + Carboxylic\ acid + 3X^- + 2H^+ \qquad \text{(Eq. 14)}$$

If ketones are involved in the VHOC formation, such mechanism would require V-HPOs to be located at or close to the surface of the thalli. The presence of these enzymes on the surface of the thalli has indeed been shown for some macroalgae[286]. It is however possible that there are other isoenzymes present throughout the algae which contribute to VHOC production through different mechanisms.

Another possibility is that HOX is released into surrounding seawater, as it is lipophilic and easily diffuses through cell membranes[302]. Once in seawater, HOX reacts with DOM to form unstable halogenated compounds, which then decay into haloforms[286] and other brominated compounds. The reaction of HOX with DOM is shown in **Equations 15** and 1**6**.

$$HOX + DOM \rightarrow DOM(X) \qquad \text{(Eq. 15)}$$

$$DOM(X)_n \rightarrow CHX_3 + DOM(X)_{n-3} \qquad \text{(Eq. 16)}$$

Other organic substrates have been suggested as sources of polyhalogenated methanes, such as methyl compounds present in seawater (e.g. methionine and dimethylsulfoniopropionate (DMSP))[360].

Apart from HOX, other halogen radicals are produced via photochemical degradation. This degradation occurs near the surface of the seawater, and the resulting halogen radicals then react with DOM to produce monohalomethanes[361-362]. This mechanism is applicable to $CHBr_3$ and CH_3I emissions[362-363]. There is more on degradation mechanisms in the next section.

VHOC DEGRADATION

HALIDE SUBSTITUTION

In seawater, VHOCs can undergo a number of degradation reactions, which decompose them completely, or which lead to the formation of other VHOCs. The high abundance of chloride in seawater causes the nucleophilic substitution of brominated and iodinated halomethanes (**Equation 17**) and haloforms (**Equation 18** and **19**).

$$CH_3X + Cl^- \rightarrow CH_2Cl \qquad \text{(Eq. 17)}$$

$$CHX_3 + Cl^- \rightarrow CHX_2Cl \qquad \text{(Eq. 18)}$$

$$CHX_2Cl + Cl^- \rightarrow CHXCl_2 \qquad \text{(Eq. 19)}$$

The rate of the reaction varies depends on temperature and which VHOCs are involved. For iodine substitution, the reaction is faster than for bromine[364]. This type of degradation lasts from a few years to several decades, and is slower at lower temperatures[365-366].

HYDROLYSIS

As was shown in **Equation 14**, ketones, enols, and other smaller organic molecules[367] can contribute to the formation of halomethanes and haloforms. However, the formation of such compounds can be due to the hydrolysis of halogenated ketones etc.[358, 368], rather than to them being by-products of the halogenation in **Equation 14**. Monohalogenated methanes and haloforms themselves also undergo degradation via hydrolysis.

In the case of CH_3X, the H_2O or OH^- act as nucleophiles and the reaction is similar to the halide substitution (**Equation 20**).

$$CH_3X + H_2O \rightarrow CH_3OH + H^+ + X^- \qquad \text{(Eq. 20)}$$

Haloforms on the other hand follow a different hydrolysis mechanism, suggested by Hine[369] and later refined by Robinson[370] (**Equations 21-24**).

$$CHX_3 + OH^- \rightarrow CX_3^- + H_2O \qquad \text{(Eq. 21)}$$

$$CX_3^- \rightarrow :CCl_2 + Cl^- \qquad \text{(Eq. 22)}$$

$$:CCl_2 + H_2O \rightarrow CO + 2HCl \qquad \text{(Eq. 23)}$$

$$CO + H_2O \rightarrow CO_2^- \qquad \text{(Eq. 24)}$$

Compared to halide substitution, hydrolysis seems to occur at lower temperatures[371]. This may suggest that when the temperature of the oceans increases because of climate change the halide substitution mechanism may become more frequent.

PHOTOLYSIS

Photolysis of VHOC occurs both in the oceans and in the atmosphere. In the case of the latter, degradation is faster in the atmosphere than degradation in the oceans[371]. In both cases, the reaction starts with the cleavage of the C-X bond. The halide radical then formed goes on to react with oxygen in the atmosphere, which leads to the formation of inorganic halide species (X, HOX, HX, XO, etc.). In the oceans, photolysis occurs mainly in the first few meters below the sea's surface[372] and plays a role in the degradation of brominated and iodinated compounds (**Equations 25-28**).

$$CH_2X_2 + h\nu \rightarrow CH_2X\cdot + \cdot X \qquad \text{(Eq. 25)}$$

$$CH_2X\cdot + \cdot X \rightarrow CH_2X - X \qquad \text{(Eq. 26)}$$

$$CH_2X - X + H_2O \rightarrow CH_2XOH + HX \qquad \text{(Eq. 27)}$$

$$CH_2XOH + H_2O \rightarrow CH_2O + HX + H_2O \qquad \text{(Eq. 28)}$$

The rate of this reaction is unknown, but has been estimated to be ca 30 days in seawater[372]. Unlike the atmospheric reactions, the cleavage of C-X bond is followed by a recombination of the $CH_2X\cdot$ and $\cdot X$ radicals, which produces an unstable intermediate, CH_2X-X. This intermediate then reacts with the hydroxyl ions in water to produce halomethanol, which, finally, degrades because of the reaction with water[373-374].

BACTERIAL DEGRADATION

The bacterial degradation of VHOCs in the oceans is a process that is even less understood than the bacterial VHOC production[375]. An example of a marine bacterium *Methylobacter marinus* was shown to degrade $CHBr_3$, but not CH_2Br_2[376]. A similar conclusion was made for three other strains of marine proteobacteria, which produced both CH_2Br_2 and $CHBr_2Cl$ as products of $CHBr_3$ degradation[377]. CH_2Br_2 , in turn, was degraded by other strains through bacterial oxidation[348]. The degradation rates of VHOCs observed in the Arctic Ocean suggest that the degradation mechanism must be bacterial[378]. Few studies have specifically been made on marine bacteria. The majority of studies about dehalogenation by bacteria focus on the degradation of anthropogenic halogenated organics in terrestrial environemnets[375]. However, the dehalogenation displayed by strains from other environments may give an indication of possible degradation mechanisms in natural environments.

Allard and Nielson[379] mention in their overview several strains of methylotrophs[380], nitrotrophs and other bacteria that are capable of VHOC degradation. Further examples show the degradation of $CHBr_3$ to other polyhalogenated VHOCs by non-marine bacteria[381]. Monohalogenated methanes, for instance, may be subjected to degradation by bacterial methyltransferases by substituting the halogen with another functional group or with another halogen[382]. A strain isolated from agricultural soil has shown the reduction of CH_3Br to CO_2[383], suggesting hydrolysis, possibly by hydrolytic dehalogenases[375]. Degradation of CH_3Br was also observed in mixed soil bacterial flora[384]. The degradation mechanisms of halomethanes by bacteria has been attributed to enzymatic reactions of methane monooxygenases and corrin-dependent enzymes[379].

To the author's knowledge, there are currently no publications that consider the bacterial degradation of polyhalogenated VHOCs in the marine environment. Similarly, there are no studies to date that focus on the degradation of iodinated VHOCs.

RAMAN SPECTROSCOPY

In Chapter 3, similarly to Chapters 1 and 2, Raman spectroscopy and its variations were used alongside machine learning methods. The theory for these methods can be found in the introductory part to Part I of this thesis.

GAS CHROMATOGRAPHY

Gas chromatography (GC) is the separation technique in which the separation of compounds is based upon the partition of the analytes between a gaseous mobile phase and a liquid or solid stationary phase through a separation column. It was developed in the late 1950s and early 1960s, with the biggest breakthrough for environmental analysis being the development of bonded fused silica capillary columns in the 1980s. Another considerable contribution to the established of GC as an technique in the environmental analysis was the development of electron capture detector (ECD)[31]. The development of the ECD detector and fused silica capillary columns allowed for analysis of halogenated compounds in ppb and ppt concentration.

For easy transport of the compounds through the column, they must be sufficiently volatile[31,32]. Separation of compounds takes place in the column and the compounds elute from the column depending upon their retention by the stationary phase. Compounds which are greatly attracted to the stationary phase will tend to elute later on the column than those less attracted. The gases used should be inert, dry and free of oxygen to prevent damage to the column. Typically nitrogen, helium, or hydrogen gases are used as carrier gases.

PURGE AND TRAP (PT)

In Chapters 4 and 5, VHOCs produced by marine algae were quantified with a GC-based method. In seawater, the concentrations of these compound is low, ranging from amolL^{-1} to pmolL^{-1} [27,28]. Since VHOC concentrations are low, it was necessary to pre-concentrate the volatiles from the artificial seawater (ASW) into the gaseous phase. Different methods for the determination of volatile organic compounds in water have been compared previously, such as membrane inlet mass spectrometry (MIMS), purge and trap (PT) gas chromatography-mass spectrometry (GC-MS), and static headspace gas chromatography. Of the three, MIMS and PT showed the lowest detection limits[29]. PT is a dynamic headspace technique in which VHOCs are purged out of the sample matrix by a stream of inert gas and carried into a sorbent trap where they are concentrated and thereafter introduced into a GC for analysis[30].

The PT technique is suited for low molecular weight and slightly water soluble volatile organic compounds[30]. PT has three main processes. It starts with extraction of volatile compounds from the sample matrix by purging it with an inert gas with simultaneous adsorption of the volatiles. These would be followed by desorption, either by use of an elution solvent or heating. In this thesis, heating was used for desorption step. Certain variables such as properties of the analytes, sample temperature, flow rate, duration of purging, sample matrix, etc. determine the purge efficiency. The last process is the trapping of the compounds, the efficiency of which is dependent on the properties of the target compounds and the sorbent materials used in the trap. The efficiency is further increased by cooling the trap.

ELECTRON CAPTURE DETECTOR (ECD)

The detector used in GC setup in this thesis was an electron capture detector (ECD), which is one of the detectors commonly used for environmental analysis. The internal chamber of the ECD is lined with a radioactive β-emitter, normally ^{63}Ni or ^{3}H. The carrier gas is ionised by the β-radiation as it passes through the detector to generate free electrons[33], which would move to the anode side of the detector due to potential gradient. This will generate an ionisation current. This makes it highly sensitive for compounds containing electronegative elements like halogens, and hence ECD is particularly suited for VHOC analysis[31,33].

Ekdahl and Abrahamsson[27] showed a simple and sensitive method for determination of VHOCs in seawater with an automated PT pre-concentration and subsequent separation and detection with GC-ECD and GC-MS. The detection limits of amolL^{-1} to fmolL^{-1} and fmolL^{-1} to pmolL^{-1} for GC-ECD and GC-MS were obtained, respectively. GC-ECD combined with PT as proposed by Ekdahl and Abrahamsson was the method applied in Chapters 6 and 7.

CHAPTER 3

CONCEPTUAL APPLICATION OF DESIGN OF EXPERIMENTS, PCA, OPLS, AND T-OPLS FOR DISCRIMINATING PROTEIN SIGNAL FROM BUFFER MATRIX

INTRODUCTION

Marine environments are dynamic biotopes, subject to complex biogeochemical processes. To further understand these, there is interest in developing methodologies that would allow the study of biogeochemistry *in situ* in e.g. living organisms. Marine algae are one of the largest groups of organisms that contribute to the complexity of the seas; for instance, they correlate with a number of climate-relevant gases, such as dimethyl sulphide (DMS)[385], CO_2[386], and halocarbons[387-388].

In the case of halogenated compounds, marine phytoplankton and macro-algae have been recognised as the primary biogenic source of volatile halogenated compounds (VHOC) into the atmosphere[288]. Due to the number of implications which VHOCs have for atmospheric chemistry and global biogeochemical processes[302, 322, 389], such as the destruction of tropospheric and stratospheric ozone, they have drawn wide scientific interest[390]. These compounds are formed in part through the activity of vanadium-dependent haloperoxidases which were first discovered in the brown alga *Ascophyllum nodosum* over three decades ago[391]. Since then, they have been isolated from and studied in several algal species[282, 305, 392-396]. However, there is still no information regarding the tissue location of these enzymes, and, currently, there are no established methods for tracking the enzymes responsible for VHOC production *in situ*. Additionally, the influence of climate change on the formation of VHOC, and the associated enzyme activity, is yet to be fully understood[397] (see Chapter 4 for further investigation of this issue). In order to further our understanding of underlying mechanisms behind the biogenic VHOC production, there is a need to develop new methods for studying the behaviour of vanadium-dependent haloperoxidases.

A well-studied representative of these enzymes is the vanadium bromoperoxidase from the red alga *Corallina officinalis* (C-VBPO). The structure and activity of the isolated C-VBPO have been studied with multiple techniques over the years[305, 398-400], but no attempts have so far been made to develop methodologies for *in situ* study of this enzyme. Raman-based methods present an attractive alternative for studies of protein structure and behaviour[186, 401-403], as well as for *in vivo* studies[85]. However, since biological matrices often produce weak Raman signals[404], signal enhancement methods can be applied to address this issue. Among the frequently utilised Raman spectroscopic methods used for this purpose is surface-enhanced Raman spectroscopy (SERS), which relies on surface plasmon resonance on metal surfaces for the signal enhancement[405]. SERS has been successfully applied several times in the investigation of protein structure and catalytic activity[401, 403, 406-408], thus demonstrating the potential of the technique in studying C-VBPO.

To establish a viable *in situ* methodology, we sought to establish a protocol that would allow separating the C-VBPO signal from its algal matrix. For this purpose we selected Raman spectroscopy as we, in the long term, aimed to establish a method that would be non-destructive and which would impose minimum damage upon living organisms. To achieve this, we wished to establish the spectral fingerprint of C-VBPO as well as to evaluate what factors may have affected its response if surface-enhanced Raman spectroscopy (SERS) was involved. Enhancement surfaces commonly used in SERS include e.g. metal films, metal-coated electrodes, colloidal suspensions, surfaces prepared with nanolithography, etc[409]. As the aim of our enquiryinvestigation was to offer a methodology that would contribute to the further development of *in situ* study of VBPOs, colloidal nanoparticle suspensions were used. These are often applied in studies of living cells as they can disperse evenly through the sample and can also be internalized by cells through passive uptake mechanisms[116, 410-412]. However, as algae do not have the same passive uptake mechanisms as mammalian cells, one of the few ways of ensuring SERS within living algal cells is through the intercellular biosynthesis of colloidal gold nanoparticles (AuNPs) [413-414]. Since such nanoparticles are unmodified per default, we wanted to study how C-VBPO responses would behave when interacting with unmodified AuNPs.

Acquired SERS responses were evaluated using chemometric methods. As a first step, we applied statistical design of experiments in a simpler system investigating fundamental factors during SERS measurements of C-VBPO. For this purpose, a composite face-centred design (CCF) was used to investigate the influence of enzyme concentration, time and number of AuNPs. The experiments were evaluated by intensities of discrete Raman peaks as responses. We also expanded the evaluation to involve spectral profiles. In the latter case, spectral interpretation was made by principal component analysis (PCA)[16], and orthogonal projections to latent structures (OPLS)[45]. In addition to standard OPLS, we also calculated transposed OPLS (T-OPLS) models[52, 54].

METHODS

CONSUMABLES

Vanadium bromoperoxidase from the red alga *Corallina officinalis* (C-VBPO), 5 M H_2SO_4, and Trizma® base were bought from Sigma Aldrich (Stockholm, Sweden) and 60 nm citrate-stabilised gold nanoparticles (AuNPs) were purchased from BBI solutions (Cardiff, UK). 8-welled sticky-slides for mounting on microscope slides were purchased from ibidi GmbH (Planegg/Martinsried, Germany). Raman grade CaF_2 microscope slides were acquired from Crystran Ltd. (Poole, UK).

PREPARATION OF ENZYME SOLUTIONS

50 mM Tris-SO_4 was prepared by dissolving Trizma® base in Milli-Q water (18.2 Ω) and then brought to pH 8.3 with 5 M H_2SO_4. Adequate buffer capacity is often only reached at concentrations higher than 25 mM. But higher concentrations can inhibit enzyme activity due to ionic strength. In addition, higher ionic strength influences aggregation of AuNPs. Suitable initial concentrations were therefore between 10 to 25 mM. After addition of the enzyme, the pH value changed by more than 0.05 units. The concentration of the buffer was then increased to 50 mM. Up to this concentration, the buffer was not expected to interfere with biological systems. Lyophilized C-VBPO powder was dissolved in 50 mM Tris-SO_4 to a final concentration of 1000 μg/L, and was then further diluted with Tris-SO_4 to 500 μg/L and 0.1 μg/L. Both the enzyme and the Tris-SO_4 were allowed to equilibrate to RT (ca 22 °C) before mixing. The C-VBPO solutions were split into aliquots and stored at -20 °C until used.

INSTRUMENTATION

Spectral data was collected using a Czerny-Turner[415] Raman spectrometer (Dilor Labram INV, Horiba) outfitted with an inverted confocal microscope (Olympus IX70). Spectral read-out was visualised with LabSpec 5 spectroscopy suite software (Horiba). The instrument was equipped with a holographic grating (950 grooves/mm) and detection was facilitated by a charged coupled device (CCD). All measurements were performed in a ca 5 μm laser spot within each well. Pinhole aperture was set to 400 μm, slit to 100 μm, and the laser was further focused through a microscope objective (X10). The excitation source was a 632.8 nm He/Ne laser with laser power ~2 mW at the sample. For each sample, a total of 10 spectra were collected for 30 seconds each within the 291-2222 cm^{-1} spectral range and were then averaged into one spectrum per sample. The instrument was calibrated against a Si signal before each set of measurements were performed.

SAMPLE PREPARATION AND SERS RUNS

The experiments were performed according to the generated CCF design. The run order was modified and split over the course of several days to make experimental execution feasible. Thus, 5 min, 24h and 48 h samples were run on three separate days. Each day, the samples were run in a random order. Similar division was done for blanks.

An 8-welled sticky-slide was mounted on top of a Raman grade CaF_2 slide. New slides were prepared when necessary. The enzyme solutions were distributed in the wells together with varying amount of AuNPs, so that the volume ratio between the sample solution and AuNPs suspension was always 1:1. For samples with $2.6 \cdot 10^9$ particles, the AuNPs solution was added as is to C-VBPO samples. The samples with $5.2 \cdot 10^9$ and $1.04 \cdot 10^{10}$ particles were prepared by first spinning down double and quadruple amounts of AuNPs for 5 min at 5000 rpm, discarding the supernatant, and then re-suspending the resulting pellet with Milli-Q water. The resuspended AuNPs were immediately mixed with samples. The analysed final concentrations of C-VBPO after dilution with AuNPs were 0.05 μg/L, 250 μg/L, and 500 μg/L. These concentrations correspond roughly to 0.004, 18, and 35 enzyme units. After mixing the samples with AuNPs, the mixtures were incubated at RT (~22 C°) for 5 min, 24h, or 48h. Blanks were prepared with Tris-SO_4 in a similar fashion as the samples.

DESIGN OF EXPERIMENTS

The experimental design matrices were generated and analysed in MODDE 12 (Sartorius Stedim Data Analytics AB, Umeå, Sweden). Two composite face-centred (CCF) design matrices were created; one for C-VBPO samples and one for blanks. The designs were created separately to better illustrate the spectral contribution from the blank signal, which might have been overlooked if it had been included in a design with samples. Three factors were examined in the designs, each having three levels – the enzyme concentration (0.004, 18, and 35 units), the incubation time of the sample with AuNPs (5 min, 24 h, and 48 h), and the amount of AuNPs present (ca $2.6 \cdot 10^9$, ca $5.2 \cdot 10^9$, and ca $1.04 \cdot 10^{10}$). The particle numbers were logged before input into the design matrix. In the design matrix of blanks, the concentration factor was omitted.

The designs included 4 replicates for each combination of factors, as well as 4 centre points. This gave rise to 60 C-VBPO samples and 40 blanks. An additional 4 centre points were measured for samples to account for possible variability in precision between days. Prior to modelling, the data was pre-treated with a rolling-circle filter (RCF)[63] with circle radii set approximately to 98 cm^{-1}. This was done in order to avoid the inclusion of signals below the baseline so that they would not be interpreted as true signals, and to compensate for baseline drift. Responses were measured as the intensities of peaks representing enzyme and Tris-SO_4 signals. All models for CCF designs were fitted with multiple linear regression (MLR)[416] in the mean auto-scaled mode. Additionally, models were normalised through log transformation. Non-significant factors that caused a deterioration of the predictive abilities of the regression model were removed. Q^2 was used as a measure of the goodness of fit based on generalized cross-validation[417].

DATA PRE-TREATMENT AND ANALYSIS

Orthogonal projections to latent structures (OPLS) algorithm and its variations were run in SIMCA 15.0.1 (Sartorius Stedim Data Analytics AB, Umeå, Sweden). File conversions and data-pre-treatments not available in SIMCA were performed with in-house Python 3.4 scripts.

All modelled data was first pre-treated with RCF. Model diagnostics can be found in **Appendix 1.1** (**Table 1-A**).

RESULTS AND DISCUSSION

DESIGN OF EXPERIMENTS

The choice of using a designed experimental space was motivated by the need to investigate how different factors of the experiment (time, amount of C-VBPO, and number of AuNPs) influence the spectral response of C-VBPO. This in turn, would provide the information necessary to make further optimization to allow for *in situ* monitoring of the enzyme. For that purpose, we selected discrete Raman peaks to represent the C-VBPO and Tris-SO_4 responses.

The number of AuNPs was selected as a factor to enable us to understand in which way the enhancement-available surface would affect the C-VBPO enzyme signal, as well as to account for any possible AuNPs-induced denaturation of the enzyme. The factor incubation time, was to examine how long the AuNPs and the enzyme needed to equilibrate to reach a discernible spectral signature. Lastly, the number of enzyme units was evaluated in order to see whether there is a linear dependency in the SERS response.

Using a designed experimental matrix has several advantages; the factors become orthogonal, which means that the influence of each factor on a chosen response can be evaluated separately. Additionally, when spectral variations are scrutinized using OPLS modelling, the fact that the variations are made in a design improves the ability to get a contrast between generally varying spectral features in the experiments and the spectral features correlating with the Y matrix in OPLS[48].

A design based on a quadratic model was given preference in order to identify factors that may cause curvature. A composite face-centred (CCF) design was chosen to be able to model square terms. CCF design is an extension of a 2^k full factorial design[27] augmented with centre points and star points[33, 38].

EXPERIMENTAL OBSERVATIONS

Through subtraction of the average spectrum of all blanks from the average spectrum of all samples (**Figure 1.1**) it is apparent that despite several overlapping signals across the measured spectral range, C-VBPO containing samples had signals sufficiently different compared to the blanks. For tentative assignments, see **Table 1.1**.

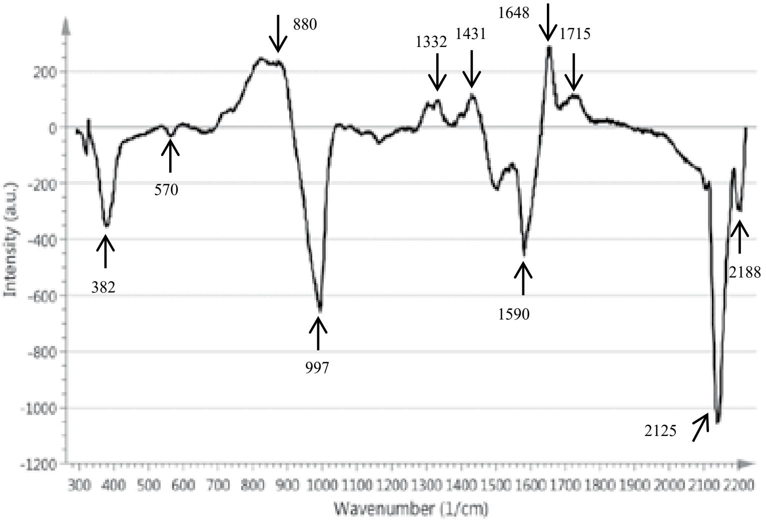


Figure 3.1. Difference between the average spectra of all C-VBPO samples and all Tris-SO_4 blanks. Negative peaks belong primarily to the Tris-SO_4 signal while positive peaks stemmed from C-VBPO.

The peaks assigned by us as Tris-SO_4 were in good correlation with the literature[418-421]. Peaks stemming from SO_4^{2-} and HSO_4^- species[422-423] and protein peaks were also detected[116, 424-427] . Based on **Figure 3.1** we selected the peak 570 cm^{-1} to represent the pure Tris-SO_4 signal and 1651 cm^{-1} to represent the C-VBPO signal in subsequent CCF modelling. In general, few of the C-VBPO peaks were well separated from overlapping buffer signals.

Table 3.1. Tentative assignments of Tris-SO_4 and C-VBPO peaks based on subtraction of average spectra of all blanks and all samples. Abbreviations used in the table: n. a. = not assigned; δ = in-plane deformation; ν = stretching; φ= bending; γ = out-of-plane deformation. The table continues on the next page.

Blanks (Tris-SO_4)	*C-VBPO - Tris-SO_4*	*Tentative assignment*
	373	C-VBPO (n.a.)
384		Tris-SO_4 (n.a.)
488		Tris-SO_4 (n.a.)
	562	C-VBPO (n.a.)
570		Tris-SO_4 $\nu_4(SO_4)$ [422]
671		Tris-SO_4 (n.a.)
766		Tris-SO_4 (n.a.)
811		Tris-SO_4 (n.a.)
	819	Protein C-C stretch[424]
	872	Pro, Val, hydroxyl-Pro, Trp C-C stretch[424]
880		δ(HCC) Tris; HSO_4^-[418, 422]
	995	Phe[424]
997		Tris δ(HNC)+ν(CC); SO_4[116, 423]
1074		ν(CO) Tris[418, 421]
	1076	Carbohydrates[425]
	1141	C-VBPO (n.a.)
1176		Tris-SO_4 (n.a.)
1265		ν(CC) Tris[421]
1290		δ(OH) Tris[421]
	1303	δ(CH2) twisting, wagging, collagen (protein assignment), Amide III[39]
1328		Tris-SO_4 (n.a.)

Blanks (Tris-SO_4)	*C-VBPO - Tris-SO_4*	*Tentative assignment*
	1379	CH_3 rocking (enz)
1385		Tris-SO_4 (n.a)
	1398	C-H rocking; C=O sym stretch, CH2 deformation, NH deformation[39, 424]
1430		Tris-SO_4 (n.a.)
	1432	CH_2 vibrations (enz)
	1495	C-C stretch benzenoid ring[424]
1545		Tris-SO_4 (n.a.)
	1580	δ(C= C), phenylalanine, C= C olefinic stretch in proteins; Phe, Tyr[116, 424]
1595		Deprotonated Tris δ(NH_2); δ(NH_2) [420-421]
1617		Tris-SO_4 (n.a.)
1630		Tris δ(NH_2)[420]
	1651	Amide I; amide I α-helix [424]
	1714	Leu, Asp, Tyr; C=O[424]
1720		Tris-SO_4 (n.a.)
	1732	C-VBPO (n.a)

EVALUATION OF THE CCF DESIGN

First, we examined how different factors influence the signal of C-VBPO by examining the intensity of the amide I peak at 1651 cm^{-1} in a contour plot (**Figure 3.2**).

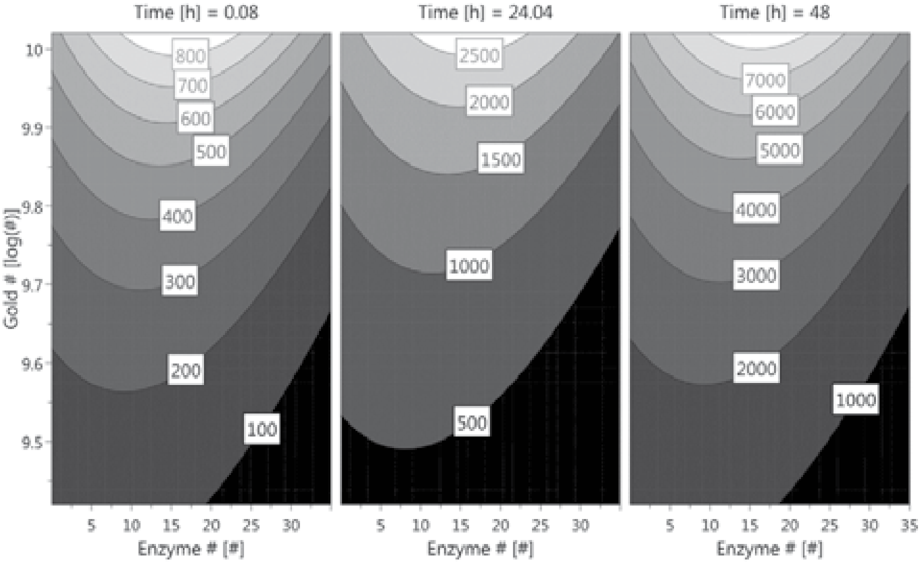


Figure 3.2. A contour plot shows how the signal intensity of the 1651 cm^{-1} peak (C-VBPO) varied with time and number of enzyme molecules present, when the Time [h] were 5 min (left), 24 h (middle), and 48 h (right). AuNP value, Gold [#], was the logged particle numbers, and Enzyme [#]was the number of enzyme units (0.004-35 units). The intensity values for 1651 cm^{-1} peak are given as numbers within the plot.

The contour plot in **Figure 3.2** shows a nonlinear response in the intensity of the 1651 cm^{-1} peak (amide I vibration), which consists of the C=O stretching of the amide group the N-H and C-N vibrations of the peptide backbone[428]. The intensity always increases with longer incubation times (comparing plots from left to right), as well as with higher numbers of AuNPs (compared along the y-axis). The model for the 1651 cm^{-1} peak, has linear, quadratic, and cross-term dependencies of the factors in the design (**Figure 3.3**).

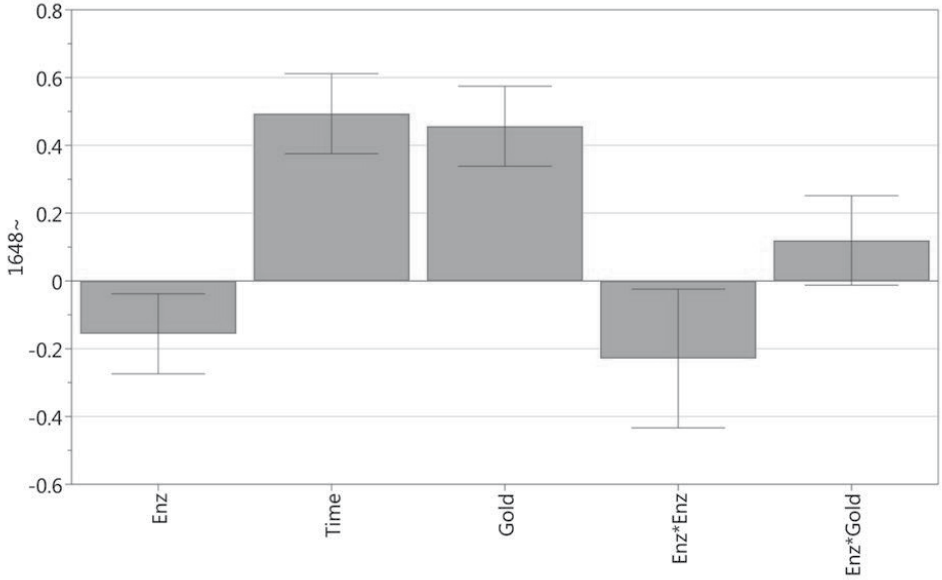


Figure 3.3. Coefficient plot displaying the significant terms of a model describing the variation in intensity at 1648 cm^{-1} (C-VBPO). The model was log-transformed 10log(Y).

Thus, the signal at 1651 cm^{-1} was linearly dependent on the incubation time (*Time*), and amount of gold (*Gold*). The signal also had a quadratic dependency on the enzyme concentration (*Enz*Enz*) and there was a small synergy effect between *Enz* and *Gold*. The linear dependencies for *Time* and *Gold* in the model had a positive sign, i.e. the signal increases with longer incubation time and more AuNPs. For high gold amounts, the enzyme response had a relative increase up to the mid-point of the design, after which there was a decrease of the relative response. This was consistent with the negative sign of the quadratic term for the enzyme factor which indicates the presence of a convex surface in the contour plot (**Figure 3.2**). This convex surface was also slightly affected by the *Enz*Gold* cross-term; increased elevation at higher gold amounts caused the skewed ridge in the contour plot.

The expected behaviour would have been that there was more response the more C-VBPO we add. However, the results suggested otherwise. A similar behaviour of the *Enz* term was also observed in other peaks stemming from both Tris-SO_4 and the enzyme (results not shown). Interestingly, with lower number of AuNPs, the apex of the C-VBPO's amide I response was shifted towards lower enzyme concentrations. This may indicate that there was a saturation of the enzyme on the surface of the AuNPs. To make a rough estimation of the number of enzyme units per nanoparticle, some geometrical assumptions were made. Assuming that the C-VBPO units and the AuNPs were spherical, a kissing number[429] could be estimated, i.e. a number of non-overlapping spheres that touch one common sphere. If the spheres had the same diameter, 12 enzymes would be arranged in an icosahedron formation around one

AuNP. But the reported diameter of C-VBPO was of course smaller, ca 150 Å[305], compared to the 60 nm nanoparticle. Thus, the size ratio is 1:4, and there ought to be 16 times more enzymes units per AuNP. This gives a kissing number of 192 enzyme units/AuNP. Therefore, even at the highest concentration of enzyme used in this study, i.e. 35 units, there was a gross surplus of AuNPs in all samples. Thus, the saturation of AuNPs' surface by C-VBPO was impossible. This could mean turn that the available enhancement surface was dominated by Tris-SO_4.

Another possibility was the interaction of the buffer with the C-VBPO backbone; earlier studies have confirmed that such interaction exists and that it influences the position of the amide I band, especially at high concentrations of Tris[430]. The concentration of Tris-SO_4 in this study, however, was kept constant at 25 mM, and so the interaction with the enzyme should have been moderate and less variable. Therefore, it can be concluded that the interaction causing the nonlinear behaviour did not stem from either the saturation of AuNPs by C-VBPO, or from a variation in Tris-SO_4 concentration, as it was kept constant.

Next, Tris-SO_4 peak at 570 cm^{-1} was modelled to represent the blank signal. The peak displayed linear behaviour since no significant quadratic terms were present. This resulted in straight lines in the contour plot (**Figure 3.4**).

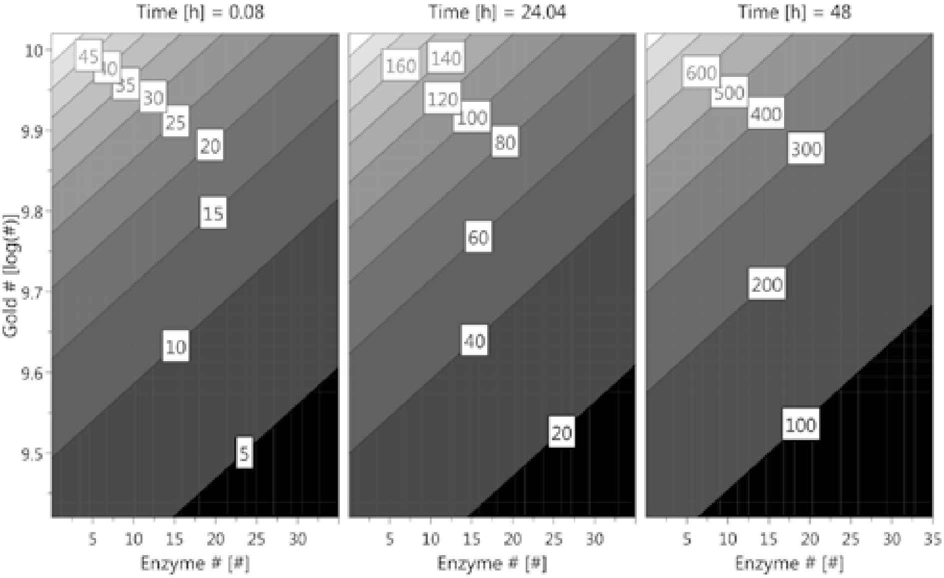


Figure 3.4. A contour plot shows how the signal intensity of the 570 cm^{-1} peak (Tris-SO_4) varied with time and number of enzyme molecules present, when the Time [h] were 5 min (left), 24 h (middle), and 48 h (right). AuNP value, Gold [#], was the logged particle numbers, and Enzyme [#] was the number of enzyme units (0.004-35 units).The intensity values for 570 cm^{-1} peak are given as numbers within the plot.

The magnitude of 570 cm^{-1} peak increases therefore, just like in the case for 1648 cm^{-1} peak, with longer incubation times (*Time*) and higher amounts of gold (*Gold*) (**Figure 3.5**).

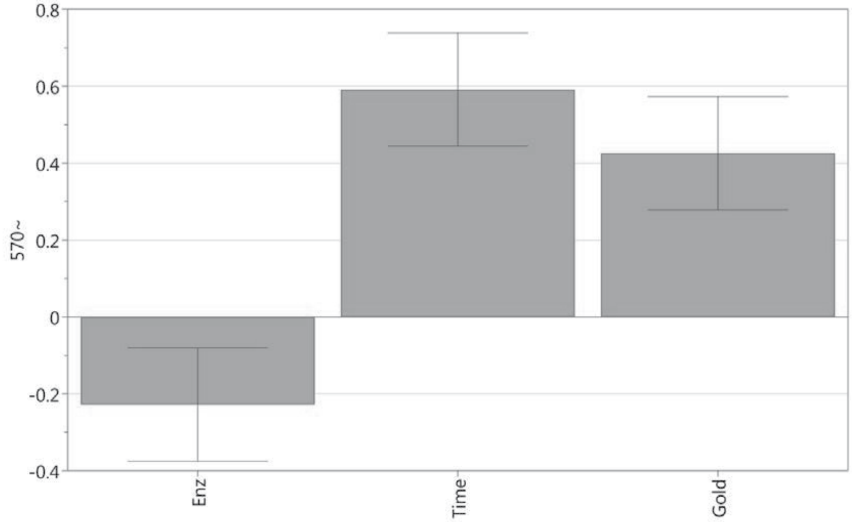


Figure 3.5. Coefficient plot displaying the significant terms for the intensity at 570 cm^{-1} (Tris-SO_4). The model was 10Log (Y+0.1) transformed.

Somewhat surprisingly however, there also was a linear dependency on *Enz*, suggesting that the amount of enzyme added was negatively correlated with the magnitude of the Tris-SO_4 peak at 570 cm^{-1}. More specifically, the intensity of 570 cm^{-1} decreased when the amount of C-VBPO increased. A similar behaviour was noticed when other Tris-SO_4 peaks were modelled. In general, there was a trend observed for both C-VBPO and Tris-SO_4, where peak intensities decrease with more enzyme added. This was observed despite the fact that the Tris-SO_4 concentration was constant and was higher than the enzyme concentration. One hypothesis would be that it is some other factor than the interaction between Tris-SO_4 and C-VBPO caused this decrease in relative response as it affects both the Tris-SO_4 and the enzyme in the same direction, with the exception of the initial increase of 1651 cm^{-1} response for high gold amount before the degradation of the response sets in.

PCA ANALYSIS

PCA OVERVIEW OF THE ENTIRE DATA SET

The nonlinearity of SERS signal becomes further apparent when performing multivariate modelling with PCA. The PCA algorithm has been explained in detail[16] in literature and earlier in this thesis, and will therefore be noted here briefly in relation to our data.

Comparing two or several models through loading vectors **p**, which represent how the variables relate to the samples in the modelled data, highlights spectral differences. Since PCA is a matrix decomposition method where in essence loadings are decomposed spectra. Loading vectors capture the main variation in the data, of which the first vector **p1** represents the largest variation, which in our case was contributions from both the spectral as well as intensity variation. The loadings that follow – **p2**, **p3**, and so on – can also contain valuable information. We observed that later loadings were able to separate overlapping peaks without the need for deconvolution as well as highlight smaller peaks. We therefore recommend that less-significant loadings are examined as much as possible in order to understand the spectral variations. However, for the sake of brevity, only the first loading vectors from PCA models are thoroughly discussed within the scope of this chapter.

Normally, PCA loadings are normalized separately for each model. To be able to compare the magnitudes of loadings between models, the loadings were de-normalised, i.e. multiplied by their eigenvalues. Through the de-normalisation it becomes possible to compare the relative contributions in signal between, say, different enzyme concentrations. Also, the models were not centred or scaled. In this way the **p1** are roughly representative of the average spectrum in each point in the design space.

The first loading vectors for PCA models performed on all blanks and all samples respectively were compared (**Figure 3.6**). Model diagnostics for this and the PCA models that will follow were reported in **Table 3-B** in **Appendix 3.1**. One experiment with the combination of factors 24 h, 18 enzyme units, and $1.04 \cdot 10^{10}$ particles was, after an examination of the score plots (results not shown), identified as a strong outlier, i.e. it deviated greatly from the general variation in the data. Due to its unique spectral shape it was excluded as it skewed the models significantly. The reason for this deviation was due to an unidentified sample contamination, which introduced alien spectral features into the model's loadings in the 700-940 cm^{-1} range, masking the contribution from the 880 cm^{-1} peak.

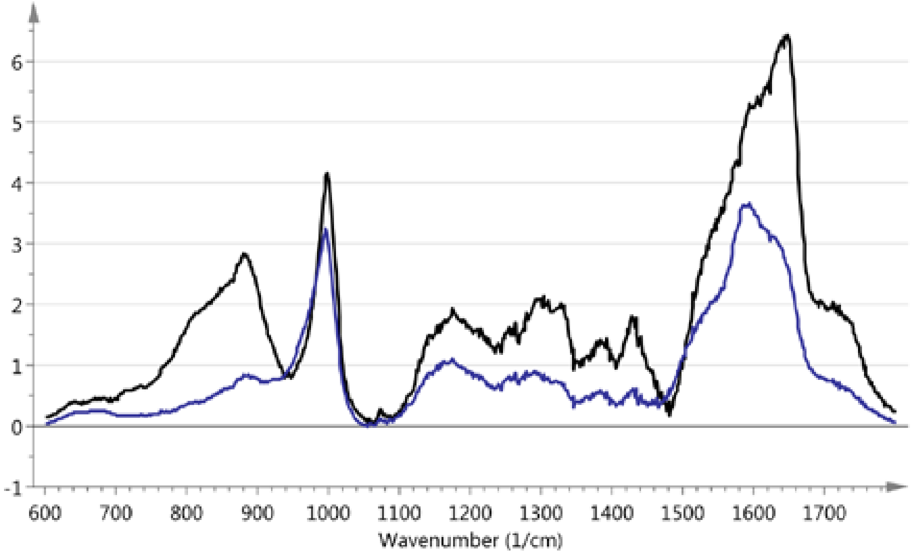


Figure 3.6. De-normalised loading line plot comparing the first two loading vectors **p1** from a PCA model performed on all samples containing C-VBPO (black) and all blanks (blue). Data in these models was pre-treated with RCF and ranged from 602-1800 cm^{-1}.

The spectral range was limited to 600-1800 cm^{-1} for PCA analysis as it contained all of the C-VBPO's peaks. **Figure 3.6** shows that the loading vector of models that contained C-VBPO-displayed more pronounced features in 700-940 cm^{-1} and amide III regions (1200-1500 cm^{-1}), as well as at 1715 cm^{-1}. Despite the partial overlap with the C-H deformation of Tris-SO_4 at 1430 cm^{-1}, the signal at 1430 and the shoulder peak at 1441 cm^{-1} for C-VBPO had a larger magnitude compared to blanks. Lastly, the C-VBPO loading shows a clear amide I peak at 1648 cm^{-1}, while the $\delta(NH_2)$ vibration of Tris-SO_4 at 1590 cm^{-1} is more pronounced in the blank loading.

C-VBPO is classified as an α+β type protein and has been reported to comprise 19 α-helices, 8 3_{10}-helices, and 14 β-strands (mostly hairpin-type) in each of its 12 subunits[305, 431]. We, therefore, expected to see indications of chiefly α-helical structures in our spectra. Conversely, the lack of α-helical vibrations would serve as an indication of structural changes.

To demonstrate the above changes and to increase the understanding of studied system, we will present in sections below results from analyses with PCA with respect to the three main factors in the design – enzyme concentration, time, and number of AuNPs.

Dependency on Concentration

First, we examined how the spectrum of C-VBPO changed with varying concentrations. A difference was observed in 700-900 cm^{-1} region in enzyme-containing samples, a range which is expected to contain a number of protein vibrations (**Figure 3.7A**).

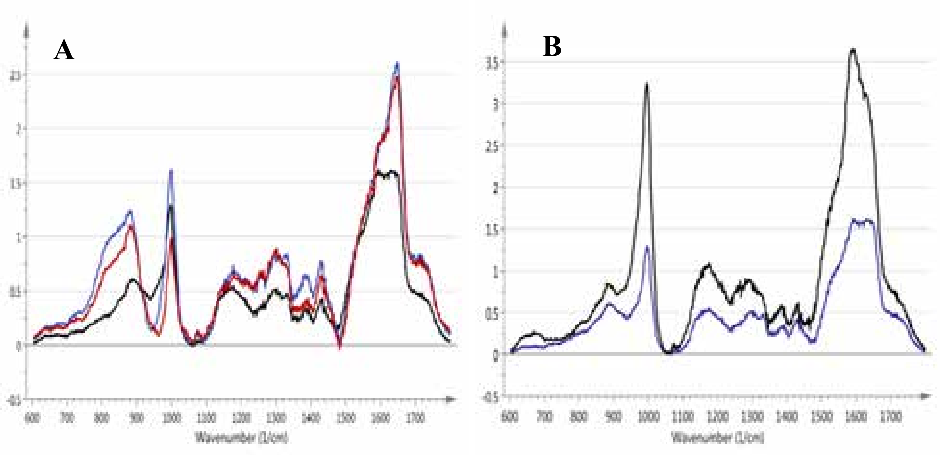


Figure 3.7. *(A)* De-normalised loading lines **p1**s for concentration dependency; 0.004 units (black), 18 units (blue), and 35 units (red). *(B)* De-normalised loading lines **p1**s comparing the spectral pattern for all blanks (black) and the lowest concentration of C-VBPO (blue). The two spectral patterns are not identical, but demonstrate high degree of similarity.

A rough calculation of the amino acid composition of C-VBPO[432] indicated that the most abundant (≥ ~5 %) residues are – in decreasing order - Ala, Leu, Gly, Ser, Asp, Glu, Phe, Val, Ile, Pro, Arg, and Asn. The vast majority of these residues have medium to strong, and often broad, bands within the specified range[433]. A strong amide I band for α-helical structure is often reported to be around 1650 cm^{-1} or in the 1640-1658 cm^{-1} range, which resonates with the 1651 cm^{-1} observed by us. At this wavenumber, a contribution from protein-associated water was also expected[434]. The high intensity of the amide I band at 1651 cm^{-1} suggested a high abundance of α-helical structures. This was further supported by the fact that the signal for this peak was much higher than in the 1235-1240 cm^{-1} range, where β-sheet and random coil structures have their vibrations[427].

All the C-VBPO signals assigned by us (**Table 3.1**) - 880, 1430-1441, 1648, and 1715 cm^{-1} - increased with increasing concentration, showing that a concentration dependency exists, although not a linear one. This was illustrated by the intensity of the 1651 cm^{-1} peak which first increases from 0.004 units to 18 units, but then decreased when concentration increased to 35 units. Other enzyme peaks behaved similarly. We relate this to our observations in MODDE models for the 1651 cm^{-1} peak, where there were significant cross-terms and quadratic terms dependent on the number of enzyme units (**Figure 3.2**). The low concentration loading in **Figure 3.7A** resembled the average spectrum for the Tris-SO_4 blank (**Figure 3.7B**), suggesting that concentration of C-VBPO at 0.004 units falls below the detection range for this method.

STABILITY AND BEHAVIOUR OF C-VBPO OVER TIME

The spectrum of a protein may change over time[401] as a consequence of nanoparticle-induced denaturation and/or changes in the secondary structure[435]. To identify potential changes, we first examined if any structural changes in C-VBPO had occurred over time by comparing three models; each model was based on how long the enzyme was incubated with gold (**Figure 3.8**).

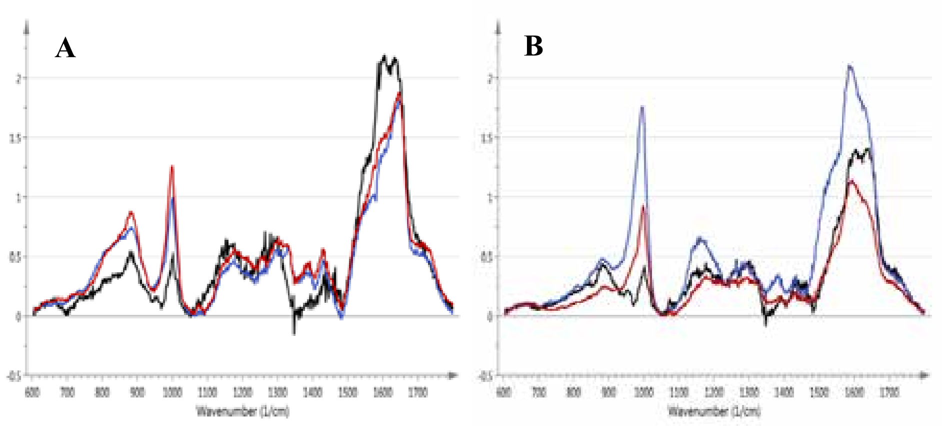


Figure 3.8. *(A)* De-normalised loading lines of the first loading vectors for C-VBPO samples incubated for 5 min (black), 24 h (blue), and 48 h (red). *(B)* Time dependency in blanks, 5 min (black), 24 h (blue), and 48 h (red).

Our PCA models demonstrate that there were several variations in intensity occurring over time. The peaks 880, 999, 1431-1441 and 1715 cm^{-1} all increased over time in enzyme samples (**Figure 3.8A**), whereas in the blanks these peaks reach a maximum at 24 h before reducing in intensity at 48 h (**Figure 3.8B**). This suggests that the main contribution to these bands stems from C-VBPO vibrations. The Tris-SO_4 peak at 1590 cm^{-1} peak decreased over longer incubation times in samples which can be seen as an indication of a stabilisation of the buffer-protein-gold interaction. Indeed, it has been reported for other proteins that the initial adsorption to the surface of AuNPs is completed within the first 10 min, yet the conformational changes continue for at least 48 h[436]. For the 1651 cm^{-1} peak, however, the loadings show that the signal is at its highest in 5 min samples, which contradicts what has been observed in MODDE the models (**Figure 3.2**). This could mean that the intensity at this wavenumber depends on contributions from Tris-SO_4 and C-VBPO signals at 1590 cm^{-1} and

1651 cm^{-1}, respectively. We still interpret the band at 1651 cm^{-1} as an enzyme band, despite the presumed interference of Tris-SO$_4$ signal.

Next came the question of the stability of the protein over time. If the majority of protein had lost its structure, other changes in the spectrum would be expected, e.g. a red-shift for the amide I peak to 1660-1670 cm^{-1} indicative of a decrease in α-helical folds[434, 437]. Such a shift has not been observed by us. This was further supported by the fact that some protein peaks became clearer only after longer incubation times, such as 1441 cm^{-1}. Furthermore, the 880 cm^{-1} peak of α-helix/β-sheet Trp as well as the amide I peak did not decrease upon longer incubation, suggesting yet again that C-VBPO's higher structures are conserved.

A reliable way of indicating that a protein has not denatured is that the disulphide (S-S) bridges remain intact and show in a SERS spectrum[438]. Disulphide bridges ought to be present in bromoperoxidases, with each monomer containing eight such bridges[311]. However, the typical wavenumbers associated with S-S (ca 430-550 cm^{-1}) were absent in our spectra. We concluded that this was not due to denaturation, but rather due to the fact that vanadium bromoperoxidases from the *Corallina* species do not contain disulphide bridges[431].

With the results reported here, it can be concluded that C-VBPO has not lost its higher structural features over time. We base this claim on the fact that the intensity of enzyme peaks changes over time, but not their position. An exception to this was the peak at 1651 cm^{-1}, which could be confounded with a signal from Tris-SO$_4$.

DEPENDENCY ON AUNP NUMBER

Next, PCA models were performed to see how the amount of AuNPs affected the SERS signals. The signal for 880 cm^{-1} seems to initially increase up to 5.2 $\cdot 10^9$ particles but decreased dramatically at 1.04 $\cdot 10^{10}$ particles (**Figure 3.9**).

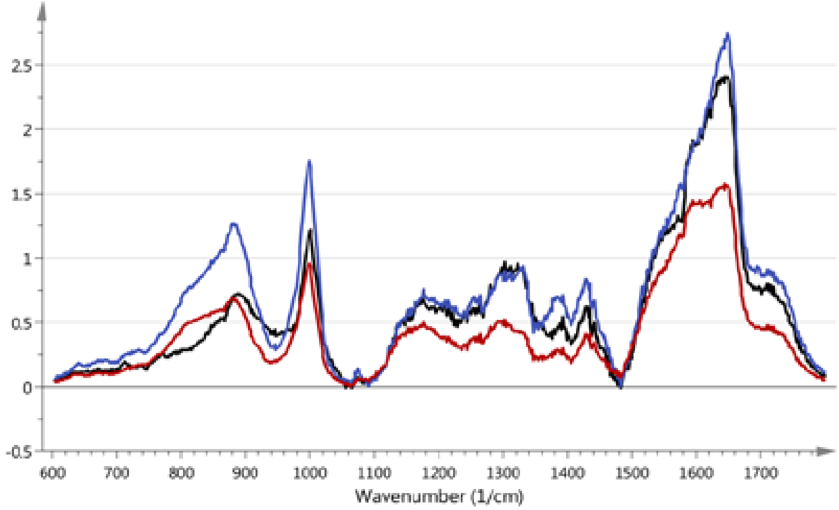


Figure 3.9. De-normalised loading lines, **p1**s, for amount of AuNPs. 2.6 $\cdot 10^9$ (black), 5.2 $\cdot 10^9$ (blue), and 1.04 $\cdot 10^{10}$ (red) particles.

A similar behaviour was observed for the C-H deformation peaks 1430 and 1441 cm^{-1}, the amide I peak at 1651 cm^{-1}, and the 1715 cm^{-1} peak. The majority of the peaks attributed to Tris-SO$_4$ displayed similar behaviour. This effect could be attributed to a saturation of the gold surface, and, thus, a LOD$_{max}$ for the method. Alternatively, protein-induced aggregation could have been the cause. However, even at the highest concentration of C-VBPO measured (35 units) there should have been plenty of surface area available in light of the large surplus of AuNPs (see earlier discussion about kissing number). In fact, higher concentrations of C-VBPO seem instead to have had a stabilising effect on AuNPs aggregation (**Figure 3.10**).

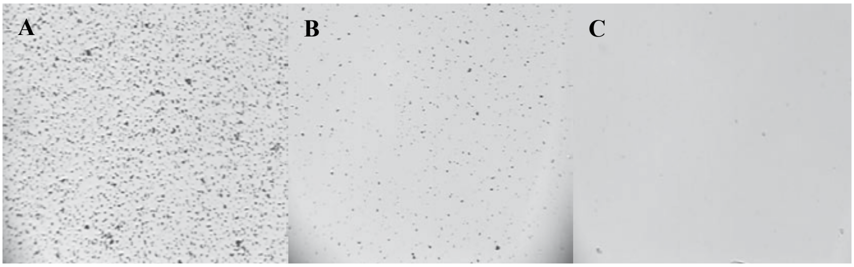


Figure 3.10. Aggregation of AuNPs as visualised under x10 magnification in sample vs. blank solutions after 24h incubation with 5.2 $\cdot 10^9$ particles. The biggest clusteres have been measured to be ca 0.5 µm. *(A)* Blank (Tris-SO$_4$), *(B)* 0.004 units C-VBPO, *(C)* 35 units C-VBPO.

Aggregation was compounded further upon longer incubation times in samples that contained little to no enzyme, as well as at higher amounts of AuNPs. Hence, it was unlikely that the aggregation has been caused by protein-gold interaction. We concluded therefore that the reduction of the signal was due to aggregation induced by the Tris-SO_4 buffer. The non-linear response in the PCA model in **Figure 3.9** could have been skewed by the inclusion of low-concentration samples, which have, as mentioned above, a spectral profile virtually identical to blanks (**Figure 3.7B**).

Since aggregation could have had an unfavourable effect on the signal, another PCA model was made by excluding all 0.004 unit samples. Upon the exclusion of low-concentration samples, the model loadings did not change significantly. The time-dependency PCA models were also changed to exclude low-concentration samples. The model diagnostics improved slightly, but the shape of the loadings remained virtually the same (results not shown).

To summarise the findings above: our attempt to compare the intensity variations using PCA models for each level in the design did not reveal sufficient detail to enabled us to identify the source of the variation that was confounded between Tris-SO_4 and the enzyme.

OPLS ANALYSIS

In **Table 3.1** we listed the peaks identified through an examination of the average spectra of blanks and samples, respectively. As this comparison only illustrated the difference between blanks and samples without any reference to a factor, and since the first evaluation of the design was made on selected peaks, we looked for a more rational method to enable us to study all peaks at the same time. For this purpose we selected OPLS analysis.

OPLS has the advantage over PLS and net analyte signal (NAS)[192] that it separates systematic variation from the sought response, i.e. the Y matrix[45]. OPLS operates on the entire design response vector, i.e. the entire spectral profile in contrast to design modelling, where the responses are evaluated one peak (or any other univariate feature) at a time. For further discussion relating to NAS, see **Appendix 3.2** at the end of this chapter.

In the OPLS models that follow, we set each of the varied factors in the CCF design – number of enzymes, number of AuNPs, and time – as Y variables. If there was a linear variation in the data that was stemming solely from one factor, OPLS ought to have been able to isolate it into a separate predictive component, **p**. The remaining variation, i.e. unrelated to Y, is described by orthogonal components, **po**. All models were calculated only for enzyme containing samples, blanks were excluded. For purposes of qualitative spectral interpretation, mean centring of models was optimal.

As a sanity check, we calculated a model for blanks where Y was set as *Enz*. The model failed to calculate components – this was expected as the blanks did not contain any C-VBPO and hence no variation dependent on *Enz* was found by OPLS. The reason behind this action was the previously observed peak dependencies in Chapter 1; OPLS and T-OPLS algorithms have been observed by us to find the analyte pattern where there ought to be none.

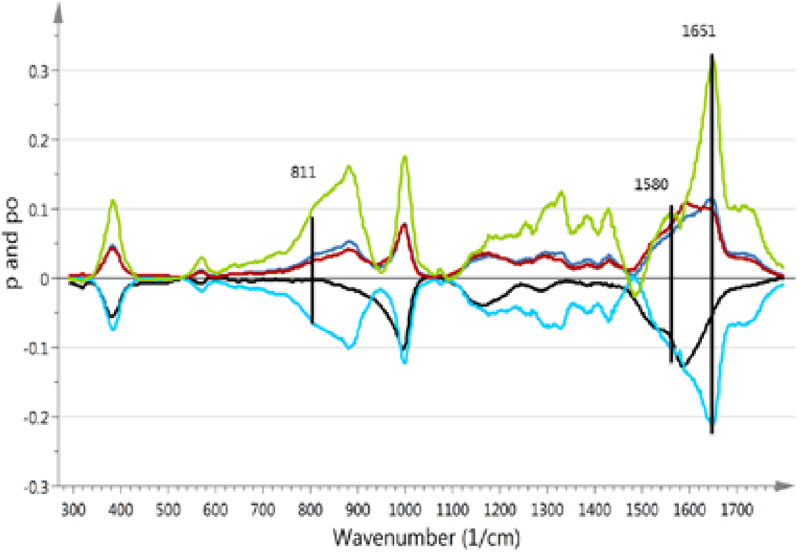


Figure 3.11. Predictive loadings **p** and orthogonal loadings **po** of an OPLS model calculated on enzyme samples, where Y was the enzyme concentration (**p** =black and **po**=light blue), time (**p**=dark blue), or number of AuNPs (**p**=red and **po**=green). No orthogonal loadings were calculated by the model where Y = time variation. Marked peaks exemplify some of the more prominent differences found between the models. All models were mean centred, including the Y vector.

Several peaks appeared both in predictive and orthogonal loadings (**Figure 3.11**). This bleed-over of spectral features from predictive to orthogonal components in OPLS has been noted earlier for nonlinear systems[48]. This confirmed the conclusion from the first evaluation of the design, that we have a nonlinear response. For Tris-SO_4 signals we were not able to pinpoint a single explanation as to why some peaks followed a linear behaviour, while other behaved nonlinearly. Several peaks appeared only in orthogonal loadings in sample and/or blank models, suggesting that there are factors influencing the spectral profile which were not covered by our design. Such factors could be day-to-day temperature fluctuations or buffer's pH. Example of the peaks include 714, 811, 898, 1141, 1245, and 1580 cm^{-1}.

T-OPLS ANALYSIS

Transposed OPLS (T-OPLS) is a method where an OPLS model is created between the X matrix and the spectrum of a pure component of interest as the Y matrix [52, 54]. Here, T-OPLS was used to separate the enzyme signal by setting the Y as the average spectrum of blanks. Since the **y** in T-OPLS models were based on blank signals only, the predictive scores **t** ought to represent signals that are related to the pattern of **y**, i.e. that of Tris-SO_4. Orthogonal scores **to**, on other hand, should show spectral patterns that are unrelated to the Tris-SO_4 signal. In other words, T-OPLS was used here as a more sophisticated way of background subtraction.

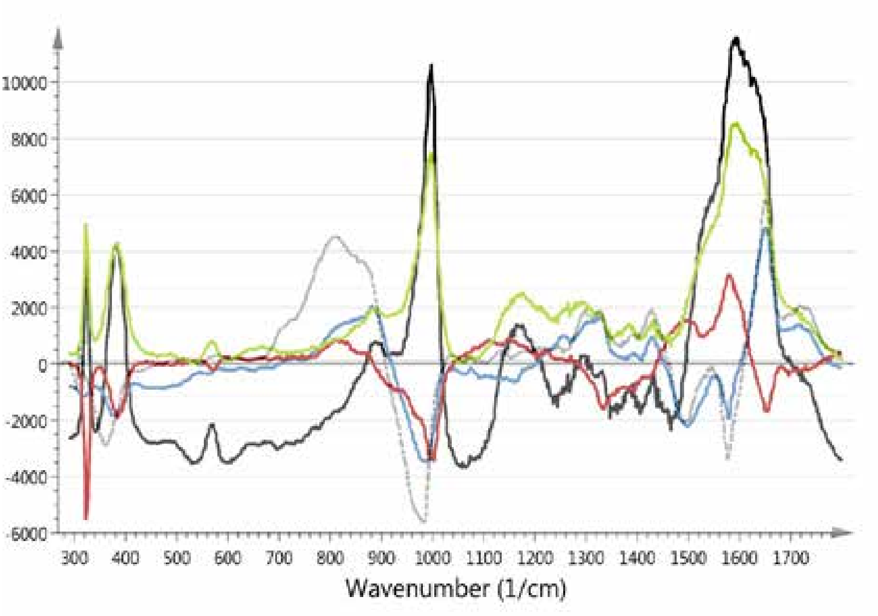


Figure 3.12. T-OPLS model performed on all samples, where y was the average signal of all blanks (green). The predictive score vector **t** (black) describes the variation related to the blank, i.e. Tris-SO_4 signal only. The orthogonal score vectors **to**1 and **to**2 (blue and red) are signals unrelated to Tris-SO_4. The dashed grey line shows the subtracted averages from **Figure 3.1**. The model was mean centred, including the y vector.

All peaks in T-OPLS predictive and orthogonal scores were noted and analysed. In the orthogonal scores, the spectral region containing amide I signals (~1500-1700 cm^{-1}) was showed better resolved details compared to previous OPLS models (**Figure 3.12**). For instance, the peak of 1580 cm^{-1}, which was previously observed in orthogonal loadings of sample and blank OPLS models, had in T-OPLS been isolated as unrelated to Tris-SO_4. Also, the peak at 1651 cm^{-1} was isolated by T-OPLS, which was previously observed in both **Figure**

3.1 and in OPLS. Notably, there were an extensive number of peaks that appeared both in **t** and **to**, further indicating the nonlinear behavior of some spectral features. These vibrations may stem from the matrix effect of Tris-SO_4 and C-VBPO.

Nevertheless, the majority of the identified peaks were unique to orthogonal scores of the T-OPLS model. T-OPLS was capable of isolating peaks previously unobserved either the subtractive spectrum (**Table 3.1**) or in OPLS models. These include OH-bending of proteins[424] (587 cm^{-1}), vibrations of amino acids Met, Phe, Pro, Val, and Tyr in α-helical and β-sheet structures (700, 856, 973, 1040, 1106, and 1188 cm^{-1}) [116, 424, 427], an amide III vibration[424] at 1223 cm^{-1}, and C-C stretch of a benzenoid ring[424] at 1499 cm^{-1}. The majority of the protein vibrations were attributed to Pro, Phe, and Tyr.

Comparisons between peaks noted in OPLS and T-OPLS reveal that many of the peaks that have been identified are confounded. For instance, peaks marked as stemming from C-VBPO in **Table 3.1**, such as 562, 1141, 1332, and 1714 cm^{-1}, behave in OPLS and T-OPLS models in a way that suggests that these signals have a contribution from Tris-SO_4. The peaks 819, 995, 1076, 1398, 1495, 1580, and 1732 cm^{-1} also have an unclear origin, although comparison between OPLS and T-OPLS models suggests that C-VBPO contributes primarily to these peaks. Only a few peaks could have been said to stem solely from C-VBPO (1303 and 1651 cm^{-1}) and Tris-SO_4 (1617 and 1720 cm^{-1}). For more information see **Table 3.2**.

In OPLS, some peaks only showed in orthogonal loadings indicating that their response was affected by factors other than those included in the design. In T-OPLS, some of those peaks (714, 811, 898, and 1141cm^{-1}), were observed in both **t** and **to**, adding to the argument that those peaks were influenced by factors outside the design.

Table 3.2. Comparison between the average spectra of all blank and the subtracted spectrum from **Table 3.1** with OPLS predictive (**p**) and orthogonal (**po**) loadings, and T-OPLS predictive (**t**) representing variation associated with the Tris-SO$_4$ and orthogonal (**to**) scores, which represent all variation unrelated to Tris-SO$_4$. Abbreviations used in the table: n. a. = not assigned; δ = in-plane deformation; ν = stretching; φ= bending; γ = out-of-plane deformation. Peaks without an assignment (empty rows) could not be identified. The table continues on next page.

Peak (cm^{-1})	***Blank spectrum***	***Subtracted spectrum***	***p samples***	***po samples***	***p blanks***	***po blanks***	***t***	***to***	***Tentative Assignment***
384	yes	no	yes	yes	yes	no	yes	yes	Tris-SO$_4$ + C-VBPO (n.a.)
488	yes	no	yes	yes	yes	yes	yes	yes	Tris-SO$_4$ + C-VBPO (n.a.)
562	no	yes	yes	yes	yes	yes	no	no	Tris-SO$_4$ + C-VBPO (n.a.)
570	yes	no	yes	yes	yes	yes	yes	yes	Tris-SO$_4$ $\nu_4(SO_4)$[422] +
671	yes	no	yes	yes	yes	yes	yes	yes	Tris-SO$_4$ + C-VBPO (n.a.)
766	yes	no	no	no	yes	yes	yes	yes	Tris-SO$_4$ + C-VBPO (n.a.)
811	yes	no	no	no	no	yes	yes	yes	Tris-SO$_4$ > C-VBPO (n.a.)
819	no	yes	no	no	no	no	no	yes	
872	no	yes	no	no	no	no	no	no	
880	yes	no	yes	yes	yes	yes	no	no	Tris-SO$_4$ > C-VBPO (n.a.)
995	no	yes	yes	no	no	yes	no	yes	Tris-SO$_4$ > C-VBPO (n.a.)
997	yes	no	yes	yes	yes	yes	yes	yes	Tris δ(HNC) + ν(CC)[424] > amide III[116]
1074	yes	no	yes	yes	yes	yes	yes	yes	Tris-SO$_4$ > C-VBPO (n.a.)
1076	no	yes	yes	yes	no	no	yes	yes	Tris-SO$_4$ > C-VBPO (n.a.)
1141	no	yes	no	yes	no	yes	yes	yes	Tris-SO$_4$ > C-VBPO (n.a.)
1176	yes	no	yes	yes	yes	yes	yes	yes	Tris-SO$_4$ > C-H bending of Tyr[424]; Tyr, Phe Tris+protein[426]; Phe[116]
1265	yes	no	yes	no	yes	yes	no	no	Tris-SO$_4$ > C-VBPO (n.a.)
1290	yes	no	no	no	yes	yes	yes	yes	Tris-SO$_4$ > C-VBPO (n.a.)
1303	no	yes	yes	yes	no	no	no	yes	C-VBPO, *δ*(CH2) twisting, wagging, collagen (protein assignment), Amide III
1328	yes	no	yes	yes	no	yes	yes	no	C-H$_2$ deformation Tris+protein[426]
1332	no	yes	yes	yes	no	yes	yes	yes	Tris-SO$_4$ > C-VBPO (n.a.)
1376	no	no	no	no	yes	yes	no	no	Tris-SO$_4$ > C-VBPO (n.a.)

Peak (cm^{-1})	*Blank spectrum*	*Subtracted spectrum*	*p samples*	*po samples*	*p blanks*	*po blanks*	*t*	*to*	*Tentative Assignment*
1379	no	yes	no	no	no	no	no	no	Tris-SO_4 > C-VBPO (n.a.)
1385	yes	no	yes	yes	yes	yes	yes	no	Tris-SO_4 C-H rocking > C-VBPO C-H rocking[39]
1398	no	yes	yes	no	no	yes	no	yes	Tris-SO_4 > C-VBPO (n.a.)
1430	yes	no	yes	yes	no	yes	yes	yes	Tris-SO_4 > C-VBPO (n.a.)
1495	no	yes	yes	yes	no	yes	no	yes	Tris-SO_4 > C-VBPO; NH in-plane deformation[424]; CH_2, CH_3 deformation Tris+protein[426]
1545	yes	no	yes	yes	no	yes	no	no	Tris-SO_4 > C-VBPO (n.a.)
1580	no	yes	no	yes	no	yes	no	yes	Tris-SO_4 < C-VBPO; δ(C= C), phenylalanine, C= C olefinic stretch in proteins; Phe, Tyr[116]
1595	yes	no	yes	yes	yes	no	yes	no	Tris-SO_4 > C-VBPO; Deprot. Tris δ(NH_2)[31] +
1617	yes	no	no	no	yes	yes	yes	no	Tris-SO_4 (n.a.)
1630	yes	no	yes	no	yes	yes	yes	no	Tris-SO_4 > C-VBPO (n.a.)
1651	no	yes	yes	yes	no	no	no	yes	C-VBPO; Amide I[424]; amide I α-helix[427]
1714	no	yes	yes	yes	yes	yes	yes	yes	Tris-SO_4 > C-VBPO (n.a.)
1720	yes	no	no	no	yes	yes	no	no	Tris-SO_4 (n.a.)
1732	no	yes	yes	yes	yes	no	no	yes	Tris-SO_4 > C-VBPO (n.a.)
1738	no	no	no	yes	yes	yes	no	no	Tris-SO_4 > C-VBPO (n.a.)

CONCLUSIONS

To evaluate whether Raman spectroscopic methods together with chemometrics can be used to develop *in situ* methodologies for marine sciences, we have here performed an extensive qualitative spectral interpretation of the two-component system of C-VBPO and Tris-SO_4. The chemometric methods applied were able to partially discern wavenumbers that distinguished C-VBPO from its matrix. By first constructing a CCF design, we were able to pinpoint several sources of nonlinearities that stemmed from quadratic behaviour at several wavenumbers, as well as to identify synergies between the studied factors.

Due to the nonlinear behaviour of the system, it is clear that a fool-proof peak assignment cannot be reached with the method used; however, our initial aim was not to scrutinize the chemistry of the system. Rather, we sought to find from which sources the different peaks stem and by what factors they were affected, an aim in which we have succeeded. The issues with peak assignment were identified as: a) overlapping contributions from C-VBPO and Tris-SO_4; b) issues stemming from chemical interaction between the two, and; c) influence of factors which were not accounted for in the design.

T-OPLS combined with a designed experimental space, in particular, makes possible the separation of relevant signals from a multicomponent matrix, which otherwise can be missed during spectral interpretation and PCA and OPLS modelling. Compared to more traditional spectral interpretation strategies, our method facilitates a more accurate evaluation of spectral features, by revealing complex matrix effects. Also, our methods highlight spectral effects such as peak confounding and hidden peaks, which can be easily overlooked during an interpretation of multicomponent spectra, where peaks are investigated one by one.

Our method can be of benefit for *in situ* detection of other algal components, such as lipids and non-volatile halogenated substances[50, 52].

APPENDICES FOR CHAPTER 3

APPENDIX 3.1 – TABLES

Table 3-A. Model diagnostics for five PCA models. Each row stands for a model performed on a group of samples that had the same characteristic, such as that all included samples were measured after 48 h.

Samples	**R^2X(cum)**	**Q^2X(cum)**	**Blanks**	**R^2X(cum)**	**Q^2X(cum)**
0.004 units	0.999	0.994	5 min	0.997	0.984
18 units	1	0.994	24 h	0.999	0.996
35 units	1	0.998	48 h	0.999	0.996
5 min	0.994	0.988	-	-	-
24 h	1	0.998	-	-	-
48 h	0.999	0.996	-	-	-
$2.6 \cdot 10^9$ particles	0.999	0.992	$2.6 \cdot 10^9$ particles	0.997	0.988
$5.2 \cdot 10^9$ particles	0.996	0.988	$5.2 \cdot 10^9$ particles	0.999	0.996
$1.04 \cdot 10^{10}$ particles	0.999	0.992	$1.04 \cdot 10^{10}$ particles	1	0.996

Table 3-B. Cumulative model diagnostics for all OPLS-based models.

Model type	**y vector**	**Components (p+po+u)**	**R^2X(cum)**	**R^2Y(cum)**	**Q^2(cum)**
OPLS	Enz	1+1+0	0.96	0.153	0.125
OPLS	Time	1+0+0*	0.904	0.253	0.23
OPLS	Gold	1+1+0	0.956	0.235	0.191
T-OPLS	Average spectrum of all blanks	1+2+0	0.971	0.99	0.99

APPENDIX 3.2 – COMPARISON BETWEEN OPLS AND NAS

Net-analyte signal (NAS) is a popular method for quantification of analytes. We used NAS as described by Bro and Andersen[439] to see if it could extract further information which would have been relevant to our enquiry, and if NAS was comparable to OPLS for the discrimination of signals from C-VBPO and Tris-SO_4. The research into the subject indicated that NAS alone would have not given us the interpretative power that we sought. This is due to the fact that NAS contains all the variation, regardless if that variation is correlated to the Y response or not. We have thoroughly examined the NAS option and performed NAS calculations before reaching these conclusions. A more extensive rationale for our decision can be found below.

NAS is a predictor of the predictive Y matrix (YPred), which is useful if the Y is not known (i.e. the varying concentrations of the sought analyte). However, due to the design, the YPred is readily available from OPLS modelling (see **Equation 3-A**). Moreover, we did not seek a determination of a Y; rather, we sought to understand how the Y matrix relates to our spectral responses, something that cannot have been achieved with NAS.

$$\begin{aligned}&YPred\ for\ NAS = \\ &Centred\ and\ scaled\ spectra\ \times Regression\ vector\ of\ the\ OPLS\ model = \\ &YPred\ for\ OPLS \qquad \text{(Eq. 3-A)}\end{aligned}$$

NAS calculation as per Bro and Andersen includes the regression vector to predict YPred by multiplying the regression vector with each spectrum (see **Equation 3-B**). Thus, such a calculation incorporates all variation corresponding to the variation in all the loadings acquired by OPLS models. This means in turn that NAS does not discriminate between the predictive component and the orthogonal component, while OPLS does. The latter is therefore more suited for the purpose of our study which was focused on qualitative analysis and signal discrimination.

$$NAS = \frac{(YPred\ for\ OPLS)}{Regression\ vector^2} \times Regression\ vector^T \qquad \text{(Eq. 3-B)}$$

When it comes to the measure of sensitivity, it becomes clear upon the examination of the NAS calculation that the parameter sensitivity is equivalent to the regression vector gained from OPLS (see **Equation 3-C**). Thus, again, no new information could have been gained in applying NAS to our case.

$$Sensitivity_{NAS} = \frac{NAS}{YPred} = Regression\ vector \qquad \text{(Eq. 3-C)}$$

Although prediction was not sought in our article per se, NAS could indeed have been used a deconvolution tactic to improve the quality of interpretation by, for instance, using norm(NAS) also known as the scalar NAS. However, as research in chromatography has revealed (we found that our spectral measurements suffer from the similar issues as the ones described for chromatography, i.e. peak overlap) there are clear drawbacks to scalar NAS application[440], including the nonfulfillment of the Euclidean metric at $k<2$

We also found during our examination of the NAS concept a consensus with previously published critique[441], specifically the realization that NAS failed to represent the behaviour of (O)PLS predictions. Again, prediction was not our aim, but the existing evidence suggests that comparing predictions made by OPLS versus those of NAS would not provide any new insights which would help us achieve the aim of our enquiries.

CHAPTER 4

MULTIVARIATE EXAMINATION OF THE EFFECT OF ABIOTIC ENVIRONMENTAL FACTORS ON THE PRODUCTION OF VOLATILE HALOCARBONS BY MARINE ALGAE

INTRODUCTION

Production of halogenated compounds by algae has been associated with environmental parameters as increasing temperature of ocean waters[319, 327-328, 442]. Change of abiotic factors such as light[327, 330, 442], salinity[327], nutrient concentration[327] and temperature[327-328] have been shown to impact the formation of volatile halogenated organic compounds (VHOCs) by marine macroalgae. In addition, H_2O_2 variations[319, 443] and elevated pH have also been shown to influence response of macroalgae in production of VHOCs. It has also been suggested that the reactive halogen species released due to V-HPOs activity in algae react abiotically with dissolved organic matter (DOM) to form halogenated organic compounds[302, 357]. Global warming and further uncontrolled eutrophication of the oceans may lead to a change in the current conditions. This may result in an unknown escalation of VHOCs emission into the global environment in the future.

To investigate how different environmental factors influence each other in production of VHOCs simultaneously, it would be of benefit to use statistical design of experiments (DoE). DoE permits enhanced assessment of interactions as variables are changed together compared to varying one factor at a time. On investigation of the relative impact of a number of factors on a specified response such as how environmental factors simultaneously influence production of VHOCs, screening designs are one of the DoE approaches often used. The simplest screening designs such as full and fractional factorial designs, consider chiefly linear interactions. By studying literature concerning biogenic VHOC production, the author of this thesis found no concrete evidence that would suggest linear dependency between VHOC production and abiotic factors. It meant therefore that the choice of the design had to consider nonlinear interactions, i.e. the chosen design had to consider cross terms and/or quadratic terms. One of such designs is definitive screening (DS) design proposed by Jones and Nachtsheim. The DS requires few experiments, its main effects are completely independent of two-factor interactions, they are not confounded as their estimation is not biased by second-order effects and enable estimation of quadratic effects if present[40, 444-445]. The DS design calculates therefore only linear and quadratic interactions. Therefore DS was deemed suitable as the initial screening design. Methodology-wise, examining the influence of multiple environmental parameters on biogenic VHOC production by means of DoE and multivariate projection algorithms is unprecedented.

Effects of abiotic environmental factors on production of VHOCs by algae have been well studied as well as methods for analysing the VHOCs in water at very low quantities[446]. Here, an established purge and trap (P&T) gas chromatography electron capture detector (GC-ECD) method was utilized as described by Ekdahl and Abrahamsson[446] to measure VHOCs. In this work, the factors chosen for closer scrutiny were light intensity, dissolved organic matter (DOM), H_2O_2 concentration, salinity, and pH. All these factors have previously been studied without consideration for the possible factor interactions, that is, each environmental factor was considered one at a time, a shortcoming that was also pointed out in the recent review by Keng et al.[239]. Any synergies and/or antagonistic effects have to date not been considered thoroughly. Thus, there is an interest in developing new methods for studying the complexity of biogenic halocarbon production multivariatly. Methodology-wise, examining the influence of multiple environmental parameters on biogenic VHOC production by means of DoE and multivariate projection algorithms as was done here is unprecedented.

VHOC PRODUCTION BY *FUCUS SERRATUS*

Fucus serratus has been shown to have highest activity of V-BrPO and V-IPO in the middle thallus, which are the mature parts of the alga's blades. Compared to other Fucales, *F. serratus* had third highest V-BrPO activity and fourth highest V-IPO activity. Highest H_2O_2 production was also found in the middle thallus followed by the lowest part of the frond axis, and, compared to other Fucale reported by Tarakhovskaya et al.[290], the alga had second-highest production of H_2O_2 in the middle thallus. Here, middle thalli of an alga specimen collected at the shores of Gothenburg were studied. Mechanical damage to algal tissue, which here was induced by cutting the thalli, is expected to increase the production of $CHBr_3$[447] and possibly other halocarbons. However, since this study was not quantitative but aimed to examine the relative distribution of VHOCs, the contribution of cutting was considered irrelevant.

CHOICE OF ENVIRONMENTAL PARAMETERS

Production of VHOCs by algae was observed to significantly reduce after approximately six hours during an incubation study. CH_3I production by *Macrocystis pyrifera* showed a linear response for about six hours[448]. *Meristiella gelidium* production of $CHBr_3$ and CH_2I_2 slowed down after about one to two hours[319]. Sampling was therefore planned at the beginning, after three hours and after six hours from start of experiment. The incubation time was not a factor in the DS design itself.

Temperature was also considered as one of the abiotic factors that affects release of VHOCs[327-328]. During this study, the temperature was kept constant. This was due to shortage of resources – the phytotron that was available for use only had constant temperature function.

With consideration that the DS works at three levels for assessment of curvature[40], the choice of three different levels for each factor were based on below considerations. The normal and/or average values for each factor were taken as midpoints, and two extremes were assigned as the minimum and the maximum. The average values were adopted to reflect the natural habitat of the sampled *Fucus serratus* as closely as possible. The values were also chosen as to enable exposure of algae to stressors but relating it to possible environmental factors variations.

PH

The net uptake of CO_2 by coastal and marginal seas is about 20% of the world ocean's uptake of anthropogenic CO_2[449] resulting in acidification of sea water. Mtolera et al.[330] observed an increase in VHOCs production from pH 8.2 to 8.8. Studying different pH levels may show general influence of VHOC production. The normal pH of seawater is approximately 8.2 pH units[450]. As mentioned in the section **Influence of biotic and abiotic stressors on VHOC production in algae**, the influence of acidification on VHOC production is uncertain. Since mesocosm studies indicated that pH change did not have any influence of VHOC production, it is plausible to assume that pH change alone would not cause a change in VHOC emissions. The author hypothesised therefore that pH may play a role in multifactor interactions, by either increasing or by decreasing VHOC emissions. The maximum was therefore chosen as pH 9.2 to maximise the stress applied. pH 7.2 was chosen as the lower limit for the same reason.

LIGHT INTENSITY

Varying irradiance has been shown to affect VHOC production[327, 330, 442]. In general, it has been found that higher light intensities are associated with higher VHOC production compared to low or dark conditions[327, 330]. In most algal species studied, the production of $CHBr_3$ and other halocarbons increased[292, 322, 372] at higher irradiance with the exception of CH_3I[292]. $CHBr_3$ production was shown to increase almost triple for light compared to dark conditions[322]. For *Fucus serratus*, the production of VHOCs increased from 3 to 12 h of irradiation[451]. It was also observed that di-halogenated compounds had high production rates compared to tri-halogenated VHOCs in longer periods of light[442]. In laboratory incubation studies, Sundström et al.[447] maintained irradiation levels between 600 and 15 μmol photons m^{-2} s^{-1}, while Mtolra et al.[330] had an irradiation range between 400 and 1500 μmol photons m^{-2} s^{-1}. Similar trend with higher VHOC emissions in response to irradiance levels has also been observed in natural habitats[372, 452]. At the sampling site of *Fucus serratus*, the irradiance was estimated post-collection to be ca 169 μmol photons m^{-2} s^{-1}. To simulate dark conditions, the lower level was set to 4 μmol photons m^{-2} s^{-1}; to induce stress the highest level was set to 334 μmol photons m^{-2} s^{-1}.

SALINITY

Salinity of the water has been associated with the rate of production of VHOC by macroalgae[327, 442]. Hyposaline (22 PSU) medium showed increased VHOCs production whereas hypersaline (50 PSU) decreased production of VHOCs. Normal salinity of seawater[450], 36 psu was used as the midpoint, 22 and 50 PSU were taken as the minimum and maximum, respectively, as was used by Bondu et. al[442] in order to stress macroalgae.

H_2O_2 CONCENTRATION

H_2O_2 is a by-product of photosynthesis and is utilized in oxidation of halides to form reactive HOX intermediate in production of VHOC[311, 453]. *Meristiella gelidium* showed increased production of tri-halomethanes, especially bromo-chloro compounds upon addition of H_2O_2. Iodinated compounds showed no significant increase in production on addition of H_2O_2[319]. Addition of H_2O_2 to algae was observed to increase formation of brominated compound[286]. Bromoform production decreased on removal of H_2O_2 while it was enhanced in presence lower concentrations (100 μM) than higher concentrations (1 mM) of H_2O_2 on studying green macroalgae[322]. Artificial stressing of macroalgae with H_2O_2 concentrations more than 3 mM have been shown to cause cell damage[443, 454]. 3 mM was therefore taken as the maximum concentration, 1.6 mM as the midpoint and 0.25 mM as the minimum.

DOM

HOX reaction with DOM is one of the mechanisms that contribute to production of halocarbons[286, 322]. Reactive DOM was shown to be a contributor in $CHBr_3$ and $CH2Br_2$ production. Production of both VHOCs was enhanced for nearshore seawater compared to offshore seawater. The same was observed at the mouth of a river during precipitation period as reactive DOM was suspected to be of terrestrial origin[357]. $CHBr_3$ production by *Ulva lactuca* decreased when DOM was removed from the seawater medium[322]. Use of oxooctanoic acid as an organic substrate in presence of bromoperoxidases and H_2O_2 lead to formation of $CHBr_3$, $CH2Br_2$ and $CH_3(CH_2)_4Br$[455]. To represent the natural environment of *Fucus serratus*, the DOM concentrations of Skagerrak (90-160 μM)[456] were considered since they apply to the shores of Gothenburg. In many aquatic systems, dissolved fulvic acid accounts for ca 40-60% of the DOM. Humic substances varies between 50% in the coastal environment to 10% in the open sea[457]. Humic acid, glycolic acid, and alginic acids, together with other cell metabolites increase the production of brominated VHOCs[458].

To represent DOM humic acid was used. The concentrations for the humic acid were estimated using levels of fulvic acid expected in DOM. The minimum point was obtained by subtracting the difference between the midpoint and the maximum point from the midpoint. The resulting values were 56, 112, and 168 μM.

METHODS

ALGAE

Fucus serratus was collected from the Gothenburg's shores. The alga was sampled in a glass bottle and completely submerged in natural seawater. The container was covered with black plastic bag to reduce effects due to direct sunlight. The alga was kept refrigerated and used the following day. The algae leafs were cut in whole pieces of approximately 1.5 g wet weight and incubated in 150 ml of artificial seawater (ASW) of differing salinities as indicated by the design matrix. 10 g wet algal weight per litre of medium has been shown to be optimal weight[35]. The algae samples were freed of any foreign objects by washing with ASW prior to the placement into the incubation bags.

ARTIFICIAL SEAWATER MEDIUM

36 PSU Artificial seawater (ASW) medium was prepared according to Kester et al[39]. NaF was omitted in the mixture due to its toxicity. The salts used are NaCl (>99.5%Merck), Na_2SO_4 anhydrous (99%, Merck), KCl (99.0-100.5%, Sigma Aldrich), $NaHCO_3$ (99.5%, Merck), KBr anhydrous (≥99%, Sigma Aldrich), H_3BO_3 (>99.5%, Fluka), $MgCl_2$ hexahydrate (Merck, $CaCl_2$ hexahydrate (98%, Sigma Aldrich) and $SrCl_2$ hexahydrate (99-103%, Merck). The actual masses are shown in Appendix 10. The salts were dissolved in ultrapure MilliQ water (Elga Purelab Flex, 18.2MΩ). ASW was then nutrient enriched according to the F2 medium procedure[56]. 22 and 50 PSU ASW was also prepared the same way but with adjustment of reagents to meet the required concentrations as shown in Appendix 10. The salinities of the ASW were verified with a salinity meter (Atago PAL-03S Pocket refractometer). They were found to be slightly higher than expected values. However, this was not a problem because the salinity meter only measures chlorinity and not the sum of chlorinity and conductivity hence higher values observed[39]. The values were 22, 35 and 50 PSU respectively.

pH ADJUSTMENT

In order to obtain the required pH values of 7.2, 8.2 and 9.2, the pH of the ASW media were adjusted with either 0.1M NaOH or HCl. A pH meter (Metrohm 691) was used for measuring. The 0.1M NaOH was prepared from NaOH pellets (>99% Riedel-de Haen) and 0.1M HCl was prepared from concentrated HCl (≥37%, Sigma Aldrich).

INCUBATION OF ALGAE

Tedlar® bags were filled with 150ml of the required media. The bags with media were incubated in the phytotron (CLF Plant Climatics, Model LT-36VL, Emersacker, Germany) overnight to acclimatise to the temperature and the different light irradiances. The light source was white fluorescent bulbs (Philips 700 series, 32 watts Alto 2). The required irradiances were obtained by shading the different sections of the phytotron with black plastic bags. The photon flux rates were verified with an irradiance meter (Model LI-1400 data logger). The actual irradiance values are shown in Appendix 9.

After the media were left overnight in Tedlar® bags, the correct volume of concentrated H_2O_2 (≥30%, Sigma Aldrich) was pipetted into the media to obtain the required concentrations as shown in Table 1. Different masses of humic acid (technical grade, Sigma Aldrich) were then added accordingly as in design matrix. This was followed by addition of algae mass. The Tedlar® bags were then sealed and air space removed using a glass sampling syringe (100ml Perfektum hypodemic, Popper and Sons Inc). This was done to ensure that no VHOCs in the water were lost to the air within the bags. The calculation of volumes of H_2O_2 pipetted and mass of humic acid added are shown in Appendices 7 and 6, respectively.

SAMPLING

After adding all the required ingredients and removing air space in the Tedlar® bag, approximately 40ml of sample aliquot was immediately extracted using a glass syringe (100 ml Perfektum hypodemic, Popper and Sons Inc.). This was the sample for 0 hour. The samples were kept in glass sample vials with no head space sealed with PTFE septa (Chromacol Ltd) to avoid VHOCs escaping from the medium. The tedlar bags were placed back in different light intensities in the phytotron accordingly. Each sample coincided with a blank. The only difference between every sample and its blank was that the blank did not contain algae. Sampling was repeated after 3 and 6 hours for every sample and blank.

MEASUREMENT OF VHOCS

A custom built P&T system with a GC-ECD chromatography system (Varian CP 3800 GC) was used. Samples were filtered through a 0.20µm sterile filter (Sarstedt Acr. & Co.) to remove particulate matter. The P&T GC-ECD system has been explained previously by Ekdahl and Abrahamsson[23]. The P&T system has electrically actuated valves, mass flow controllers and specially designed electronics. All tubing is made from stainless steel. The sample is injected using a syringe fitted with a filter through a valve into a 27.8 ml sample loop. The sample will then be transferred to a purge chamber. The purge chamber is mounted in two aluminium blocks. The lower block is heated to 70°C to increase extraction efficiency while the upper block is cooled to approximately 0°C with a circulating mixture of water/glycol to minimise amount of water vapour in the gas. Ultrapure N_2 was used as purge gas. The trap is made of stainless steel tubing filled with adsorbent (VOCARB, Superlco). It is placed in an aluminium block held at 0°C by an aluminium plate cooled by circulating water/glycol mixture. The cooling plate is pushed out of the trap during desorption. The trap is then rapidly heated to 240°C for desorption of the compounds.

GC-ECD

The separation of compounds was carried out with a Varian CP 3800 GC equipped with a ^{63}Ni ECD. The column used is a 57m capillary column with an internal diameter of 320μm. The column flow was 3.7 ml/min. Film thickness in the column was 1 μm, and the stationary phase was DB-624. Ultrapure N_2 was used as the carrier gas and make-up gas with flow rate of 20 ml/min. The temperature program for the column oven is as shown in **Table 4.2** below. The detector was set to 300°C.

Rate (°C/min)	Temperature (°C)	Time (min)	Total (min)
Initial	240	0.00	0.00
14	24	6.49	14.70
10	70	0.00	19.20
3	80	0.00	22.54
5	95	0.00	25.54
12	160	0.00	30.95
40	230	3.30	36.00

Table 4.2 The temperature program for the column oven.

Following VHOCs have been measured: iodomethane (CH_3I), iodoethane (CH_3CH2I), dibromomethane (CH_2Br_2), dichlorobromomethane ($CHCl_2Br$), chloroiodomethane (CH_2ClI), dibromochloromethane ($CHBr_2Cl$), bromoiodomethane ($CHBr_2I$), bromoform ($CHBr_3$), and diiodomethane (CH_2I_2).

DATA ANALYSIS

PCA and OPLS-DA multivariate statistical analysis was performed using SIMCA software (version 15.0.2 Sartorius Stedim Data Analytics AB, Umeå, Sweden). Data was pre-treated by centering. The experimental design matrix (DSD) and contour plots were generated in MODDE 12 software (Sartorius Stedim Data Analytics AB, Umeå, Sweden).

RESULTS AND DISCUSSION

PCA ANALYSIS

A PCA model was calculated including samples measured immediately after the introduction of algae into Tedlar® bags (referred to as 0 hours from this point onward) and at 3 hours after incubation, and blanks 0 and 3 hours after incubation. The samples collected at 6 hours and the corresponding blanks were not included into the model as they reached instrumental saturation. Unfortunately, the sample from experiment number 1 (see **Table 4-A** in **Appendix 4.1**) could not be measured due to instrumental error. The blanks followed the same DS matrix as the samples but were without any added algae. The resulting PCA model had 4 components with R^2X(cum) value of 0.977 and Q^2(cum) value of 0.792 (**Figure 4.1**).

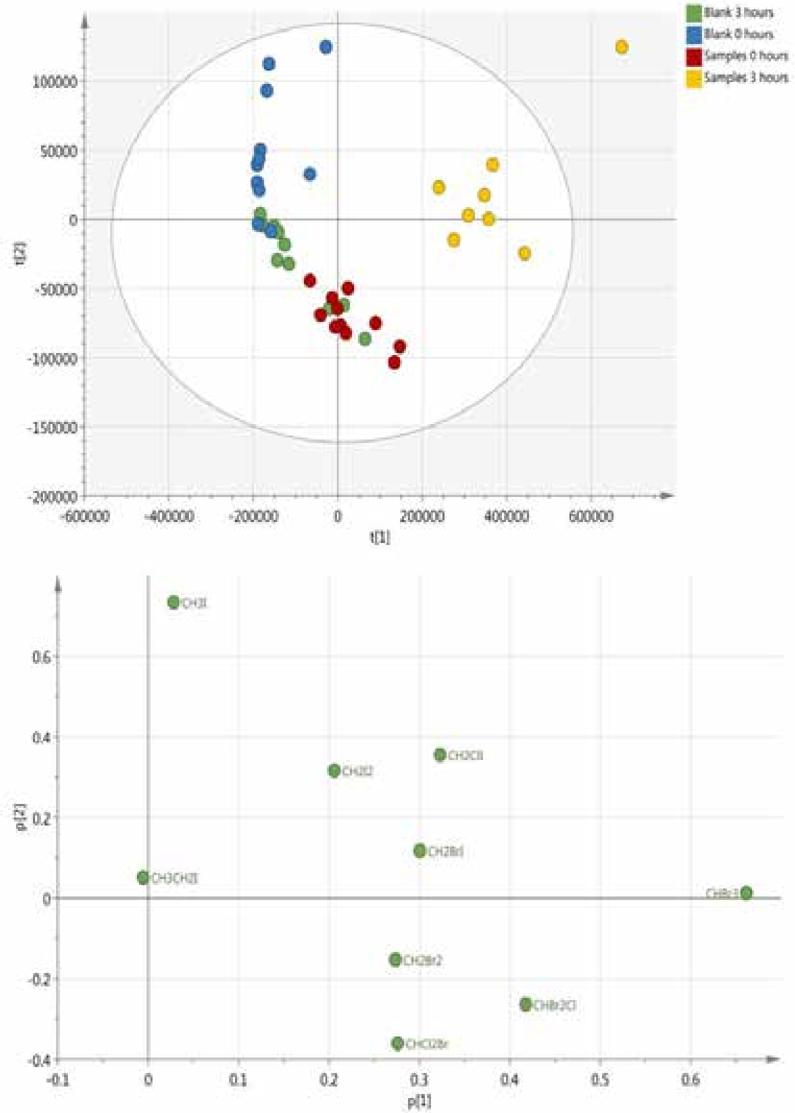


Figure 4.1 Score (top) and loading (bottom) scatter plots showing the first two score and loading vectors, respectively. The score plot shows blanks at 0 hours (blue), blanks at 3 hours (green), samples at 0 hours (red), and samples at 3 hours (yellow). The loading plot shows instead the how different VHOCs have affected the distribution seen in the score plot.

Blanks at 0 and 3 hours follow a similar trend and were notable because of a higher production of CH_3I, which then decreased after 3 and 6 hours of incubation. The PCA model showed that most samples containing algae had a relative increase in VHOC concentration compared to the blanks, with the exception of CH_3I and CH_3CH_2I. It should be noted that there was a slight production of all other halocarbons in the blanks. There was, however, a net production of VHOC in all samples, which was calculated by subtracting the blank (results not shown).As 3 hours passed, the samples displayed a significant increase in the VHOC production (**Figure 4.2**).

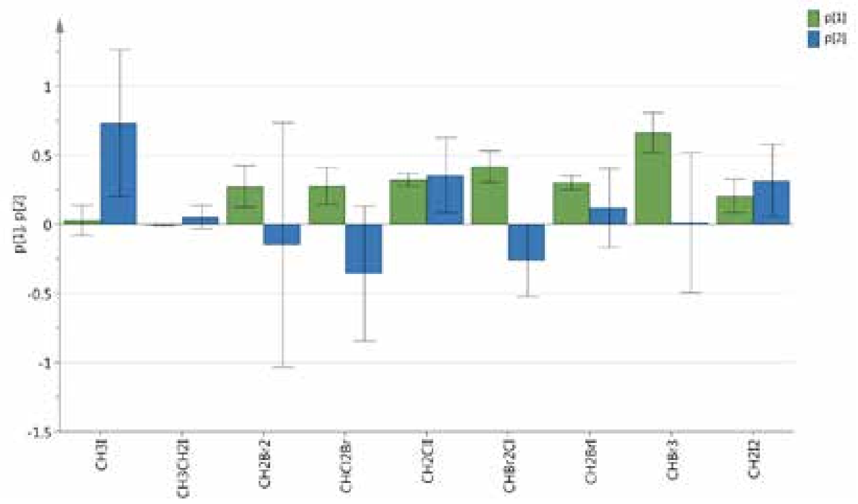


Figure 4.2. Loading column plot showing the influence of each VHOC on the model, where green column represent the impact of VHOCs along the first loading vector **p1**, and blue columns represent the impact of VHOCs along the second loading vector **p2**.

Figure 4.2 shows which of the VHOCs had the greatest influence on the model and how significant they were (indicated by confidence interval bars). Along the first loading vector **p1** the formation of all VHOCs, except CH_3I, was significant. The formation of CH_3I was instead significant along the **p2 vector**. CH_3CH_2I was not significant in any vector, and was therefore disregarded from further analysis. The VHOCs that had the greatest influence on the separation along the first PC were (in decreasing order of importance) $CHBr_3$, $CHBr_2Cl$, CH_2ClI, CH_2BrI, CH_2Br_2, $CHCl_2Br$, and CH_2I_2. The separation observed along the second PC was instead caused by (in decreasing order of importance) CH_3I, CH_2ClI, and CH_2I_2. It was clear that the reason for the separation of blanks and 0 hour samples from the 3 hour samples was caused by a higher production of iodinated VHOCs (CH_3I, CH_2ClI, and CH_2I_2) by blanks, and other VHOCs, mainly $CHBr_3$, in case of the 3 hour samples. These results agree with the well-established formation mechanisms of brominated and iodinated VHOCs, where HOX, a by-product of photosynthesis, reacts with DOM (for formation mechanisms, see **Part II**). The PCA model also suggests that CH_3I follows a different formation pathway, which has been shown previously in the literature (see **Part II** on VHOC formation mechanisms).

To further investigate the dependencies between different abiotic factors and VHOC formation, OPLS and OPLS-DA models were calculated. For OPLS models, the Y was one of the factors at a time (pH, irradiance, humic acid concentration, H_2O_2 concentration, and salinity). However, regardless of Y, OPLS failed to calculate any components suggesting that there was no linear relationship between the observations (blank and sample measurements) and the selected Y. OPLS-DA did not show any difference between, for instance, high and low concentrations of H_2O_2 in samples. The only OPLS-DA models that detected a difference between subsets of samples and/or blanks were dependent on time, i.e. if the measurement was taken at 0 or 3 hours. Thus, OPLS-DA did not provide any more information than was already observed in PCA.

As no OPLS model was possible to calculate, it stood to reason that the design would reveal quadratic interactions and/or contribution from both linear and quadratic terms. Based on the observations gained from the PCA model, a selection of the more significant VHOCs was made for further detailed analysis of the DS design. The coefficient and contour plots for all remaining VHOCs can be found in **Appendix 4.2**.

DS ANALYSIS

The analysis of how different factors have contributed to the formation of VHOCs was done through the interpretation of coefficient and contour plots. The responses of several VHOCs were gathered following the design matrix (**Appendix 4.1**), where samples were taken at 0 and 3 hours. To arrive at the coefficient and contour plots presented here, the 0 hour samples were subtracted from the 3 hour samples. The resulting difference was used to calculate DS models. All DS models were fitted with PLS. No models could be calculated for 6 hours samples due to an oversaturation of the GC-ECD detector. Normally, this could be addressed by diluting the samples. However, the sample volumes were too small and no replicates had been made. Therefore, 6 hour samples were not modelled. A similar issue of too high response occurred in 3 hour samples for CH_2Br_2 and $CHCl_2Br$, but since not all the measurements were oversaturated, these two VHOCs remained part of the design. The used design matrix can be found in **Appendix 4.1** together with model statistics.

The coefficient plots show whether factors are linear (X) or quadratic (X^2) and if they have a negative or positive correlation with the sought response, which in this case is the production of VHOCs. In the coefficient plot, the significance of the factors is usually shown with confidence bars similar to **Figure 4.2**. However, no confidence bars could be calculated, as the model had too few degrees of freedom. One of the reasons behind this was the loss of one of the experiments spanning the design space due to instrumental error (experiment number 1 in **Table 4-A**, **Appendix 4.1**). The loss of one of the experiments in a DS design may potentially lead to confoundings between terms. Another reason for the lack of confidence bars was the lack of a sufficient number of centre points in the design space.

The contour plot shows the influence of both linear and quadratic factors. As shall be illustrated, no models were free from the influence of quadratic terms, meaning that the contour plots always display some degree of curvature. If the terms in the model were only linear, then the contour plot would have displayed straight lines (see example **Figure 3.4** in Chapter 3). More on how quadratic terms should be interpreted can be found in section **3.1.2** in the beginning of this thesis. The redder the colour in the contour plot, the higher the estimated production of VHOC, while the bluer the colour, the lower the estimated production. The patterns of the contour plots at different salinities were identical for 22, 36, and 50 PSU for all VHOCs. Therefore, only the 36 PSU plots were shown in **Figure 4.3** and contours plots that follow. The reader should note that the 0 and the negative values shown in the contour plots did not represent measured concentrations. Instead, these values were values calculated by the design model to span the design space.

Table 4.2 shows a summary of the factors used in the design and their levels (low, medium, high).

Factor	Shorthand	Low	Medium	High
pH	pH	7.2	8.2	9.2
Salinity (PSU)	Sal	22	36	50
Humic acid concentration (mM)	DOM	0.56	1.12	1.68
H_2O_2 concentration (mM)	H2O2	0.25	1.63	3
Irradiance (μmol photons m^{-2} s^{-1})	Lig	4	169	334

Table 4.2. The factors used in the calculation of DS design models.

Earlier investigations have shown that marine macroalgae produce all of the known VHOCs, including the ones examined here[294-295, 319, 454]. Current research indicates that most of VHOCs are produced as by-products of photosynthesis (see **Part II**) through scavenging of H_2O_2 and other ROS. This would mean that the response to different environmental stressors should be the same, which would have led to a similar response in the design space. Also, the formation of CH_3I would, according to theory, differ from the formation of other VHOCs. During the investigation of coefficient and response contour plots it was evident that brominated VHOCs, with the exception of $CHBr_3$, seemed to have been formed through similar mechanisms, while CH_3I formed following a different pathway. The other iodinated VHOCs, CH_2ClI and CH_2I_2, had a similar pattern in the design space as the brominated VHOCs. The response contour plots also revealed that the formation of VHOCs was far more complicated than had previously been thought, which was indicated by the presence of curvature in the design space.

INTERPRETATION OF THE DS DESIGN

The VHOCs CH_2I_2, CH_2Br_2, $CHCl_2Br$, $CHBr_2Cl$, CH_2BrI, and CH_2ClI all showed a similar circular concave pattern. The maxima observed in the contour plots corresponding to those compounds was dependent on the negative quadratic terms for pH (pH*pH) and humic acid concentration (DOM*DOM). The position of the maxima was determined by the magnitude of the linear terms for pH and DOM (**Figure 4.3** and **Appendix 4.2**). Thus the quadratic terms for the concentration of humic acid and pH had on the whole a significant positive contribution to the production of VHOCs compared to other factors in the design. Among the linear terms, there was a positive relationship between an increased VHOC formation and pH and DOM terms. The influence of DOM reported here supports the findings of laboratory experiments performed with the V-HPO enzyme and different DOM, where the production of brominated VHOCs was dependent on the chemical composition of DOM[458]. The exception from this was CH_2I_2 which had a negative linear pH term (**Figure 4.3**).

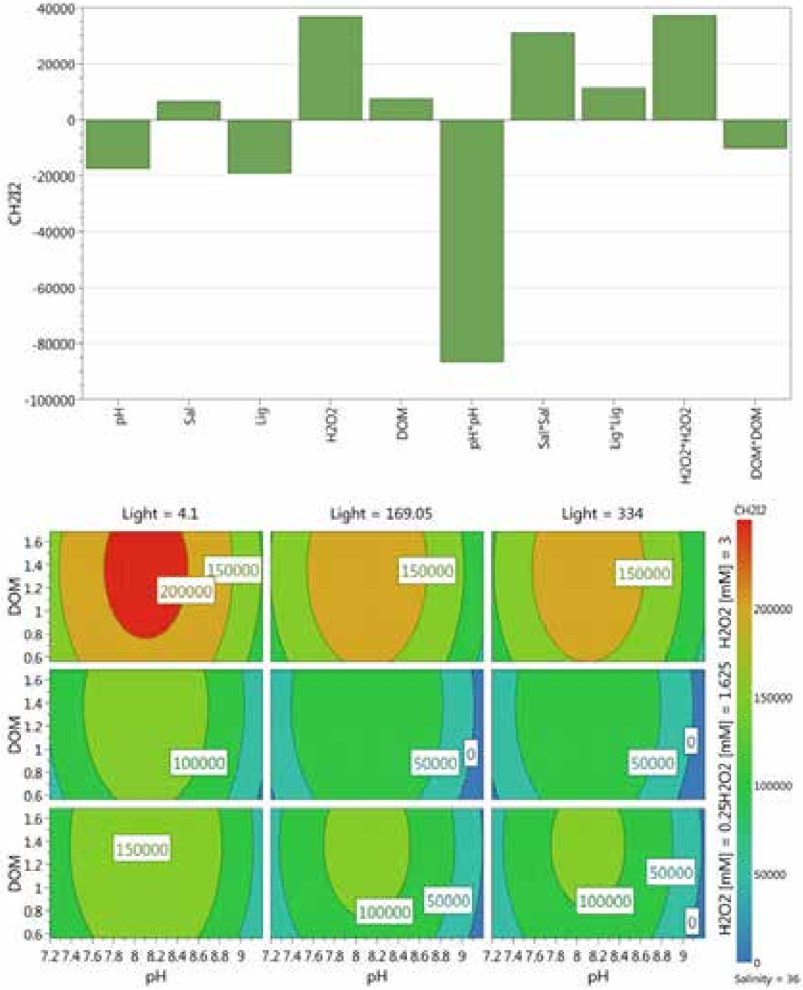


Figure 4.3. Coefficient (top) and response contour (bottom) plots showing the influence of five different factors on formation of CH_2I_2. The contour plot shows the influence of four factors at salinity 36 PSU on the production of CH_2I_2.

To exemplify how the design model for a VHOC can be interpreted and related to existing literature, the results observed for CH_2I_2 were interpreted as follows. The formation of CH_2I_2 specifically was predicted to have a maximum at low irradiance, pH range from ca 7.6 to 8.6, low-medium to high humic acid concentrations, and high H_2O_2 concentrations. As for most VHOCs, CH_2I_2 formation followed a circular concave pattern. Production was estimated to drop as irradiance increased, and to halt completely at medium and low H_2O_2 concentrations and high pH.

Similarly to many other VHOCs produced by algae, the increased production of CH_2I_2 has been associated with light-induced stress[330, 459] through the production of H_2O_2 during photosynthesis. Although H_2O_2 in the design was indeed connected to increased CH_2I_2 formation, it was not associated with high irradiance. Thus, H_2O_2 produced through photosynthesis, which was independent from the H_2O_2 added to the incubation bags, could not have caused the observed maximum. The mechanism for CH_2I_2 formation could be dependent on other parameters than light-induced stress. The stressor inducing the CH_2I_2 production could instead be induced by other ROS. An increased ROS production may have been caused by oxidative stress induced by bacterial infection. Küpper et al.[258] have shown that CH_2I_2 was produced as a result of a burst in ROS scavenging to mitigate bacterial growth. It has also been demonstrated that iodinated compounds such as CH_2I_2 tend to have higher emissions in response to the presence of oligoguluronates, which are generated during the degradation of the cell walls by bacteria[460].

Therefore, at lower irradiance, the extracellular H_2O_2 added to the incubation medium may have induced stress, but it was less likely to be connected to photosynthesis. At medium irradiance and salinity of 36 PSU, the production of CH_2I_2 was estimated to still be high, although lower than at the predicted maximum. The decrease in CH_2I_2 at those 'normal' conditions could have been caused by competing degradation mechanisms. The degradation of CH_2I_2 via photolysis is the primary degradation pathway of this VHOC in seawater[461]. According to earlier reports, the degradation of CH_2I_2 happens after 12 to 30 minutes[462-463].

Another explanation of the production of CH_2I_2 by *Fucus serratus* under darker conditions could be connected to the activity of nonspecific methyltransferases, as has been shown for the production of CH_2Br_2[280]. However, there has been to date no concrete evidence to suggest that involvement of methyltransferases in algae can result in CH_2I_2 production. The author of this thesis speculates that the bacteria themselves could have also contributed to CH_2I_2 production, possibly by methylation of CH_3I. This assumption was based on the earlier observations of CH_2Br_2[348, 352, 376, 464].

The production of CH_2I_2 has been shown to have strong correlation with total chlorophyll (chl *a*) and total bacteria counts together with the acidification of seawater in mesocosm studies performed in the Arctic[462]. This study showed an interaction between bacterial abundance and the acidification level, which the author interprets as a nonlinear behaviour of CH_2I_2 production. It was shown that there was a negative correlation between bacterial abundance and CH_2I_2 production, while at the same time CH_2I_2 had increase production per bacteria upon acidification. It was concluded that the observed increase of CH_2I_2 in the seawater was either due to an increased CH_2I_2 production by bacteria or by a decrease in bacterial degradation as a response to acidification. As evident from

Figure 4.4, the pH*pH term had a negative correlation with CH_2I_2 production, which could support the connection between acidification and increased CH_2I_2 emissions. It was however not possible to correlate this to bacteria directly as no bacterial analysis of the samples was performed.

An interesting exception from the circular concave pattern observed for CH_2I_2 and the majority of VHOCs, was $CHBr_3$, which had a unique formation predicted by the design. Instead of the circular concave pattern, the formation of $CHBr_3$ followed an ellipsoidal concave pattern (**Figure 4.4**).

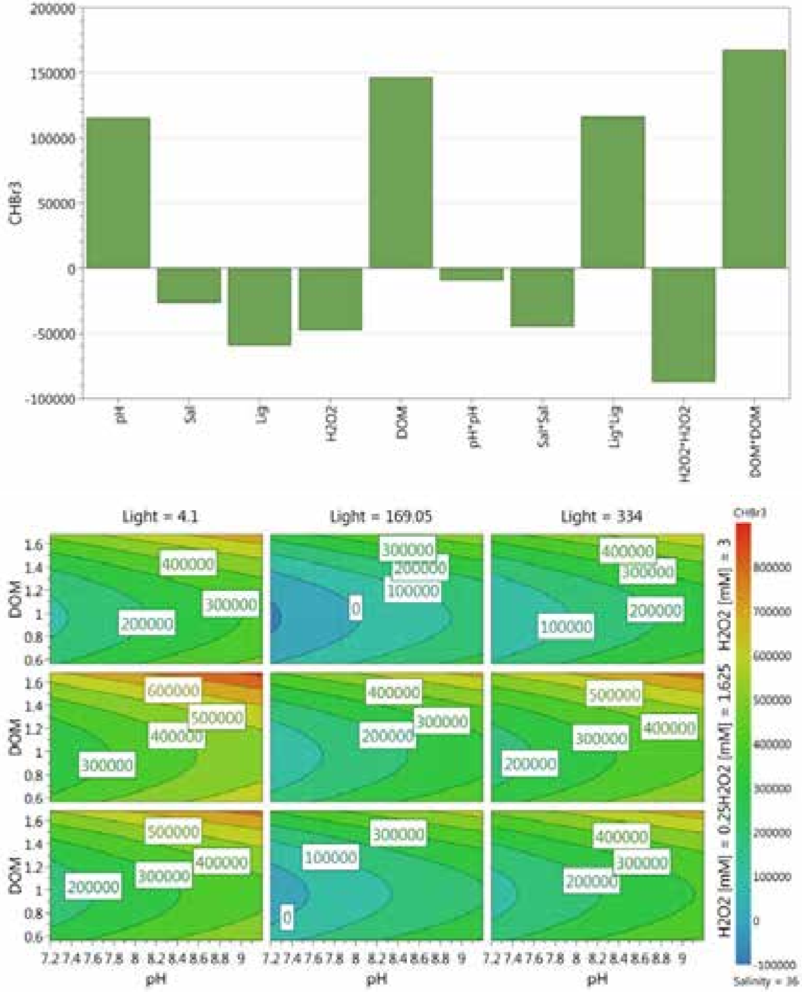


Figure 4.4. Coefficient (top) and response contour (bottom) plots showing the influence of five different factors on formation of $CHBr_3$. The contour plot shows the influence of four factors at salinity 36 PSU on the production of $CHBr_3$.

After 3 hours of incubation, maximum production was predicted at high humic acid concentrations, medium H_2O_2 levels, low irradiance, and high pH. The production of $CHBr_3$ was estimated to stop completely at medium irradiance, low to medium pH, medium humic acid concentrations, and low and medium H_2O_2 concentrations. The deviation from the behaviour observed for all other brominated VHOCs was caused by a negligible negative pH*pH term and a strong DOM*DOM term. In addition, compared to other brominated VHOCs, $CHBr_3$ displayed stronger dependence on linear terms for pH and for humic acid, which had a positive correlation with $CHBr_3$ production. As the pH term was positive, it meant that the $CHBr_3$ emissions were predicted to increase upon increasing pH. The most dominant quadratic contributions were Lig*Lig (positive), H2O2*H2O2 (negative), and DOM*DOM (positive).

In the literature, the main formation mechanism for $CHBr_3$ has been ascribed to the activity of V-HPO enzymes in seaweeds[286]. V-HPO produce reactive halogenated intermediates such as HOX, which then react with ketones and/or DOM. Indeed, Liu et al.[458] showed that the presence of humic acid increased the production of brominated VHOCs, including $CHBr_3$. This resonates with the observation in **Figure 4.4**, where the highest production was observed for high humic acid concentrations. The connection between high pH and high $CHBr_3$ production in samples was harder to explain. In previous publications, it was observed that an increase from pH 8.0 to 8.8 led to higher $CHBr_3$ levels. However, at low irradiance, the exact opposite was true[330]. Here, the highest production was estimated to be found at high pH levels and low irradiance, which does not conform to previously reported results.

Production decreased between 4.1 μmol photons m^{-2} s^{-1} and 169 μmol photons m^{-2} s^{-1}, but then increased again at 334 μmol photons m^{-2} s^{-1}. Higher irradiance did therefore influence the production of $CHBr_3$ as interpreted by the positive quadratic term Lig*Lig. This increase could be due to light induced stress, which in turn increased the production of H_2O_2 and the activity of V-HPOs. The dependency on H_2O_2 concentration was shown to have a strong association with the production of brominated VHOCs[460]. It was nevertheless puzzling that the highest production was observed at low irradiance. Contrary to what was found here, the evidence in the literature states that higher irradiance causes high production of $CHBr_3$[292, 322, 372, 459]. For instance, the increase of concentration of $CHBr_3$ has been observed to increase between 3 and 9 hours incubation with high irradiance exposure, and then decrease when reaching the 12 hour mark[451]. With the short incubation times reported here, the highest production of $CHBr_3$ ought to have been estimated at highest irradiance, which was not the case. Alternatively, the observed lower production at 334 μmol photons m^{-2} s^{-1} could be explained by there being competing degradation mechanisms at the highest irradiance, thus decreasing the overall $CHBr_3$ estimation at that irradiance. $CHBr_3$ has been reported to be degraded by bacterial hydrolysis[376-377], abiotic hydrolysis[370], and photolysis[286, 397, 465]. Although photolysis and hydrolysis may have seemed to be a likely degradation pathways to an explanation of the observed results, it is too slow[372] to have occurred in the incubations performed here.

The dependence on low irradiance could indicate that bacterial production of $CHBr_3$ could have been involved. However, although some strains have been shown to produce $CHBr_3$, the bacterial contribution was deemed too low to explain the observed $CHBr_3$ emissions in a field study[281]. Therefore, the chief contribution of $CHBr_3$ observed here must have stemmed from *Fucus serratus*.

A VHOC that has been reported to follow a different formation mechanism was CH_3I. After 3 hours (**Figure 4.5**), the production of CH_3I was predicted to be mainly dependent on the linear terms for pH (positive), Lig (negative), and DOM (positive). With the exception of the linear DOM term, the highest influence was estimated to originate from quadratic terms, notably pH*pH (negative), Lig*Lig (positive), and DOM*DOM (positive).

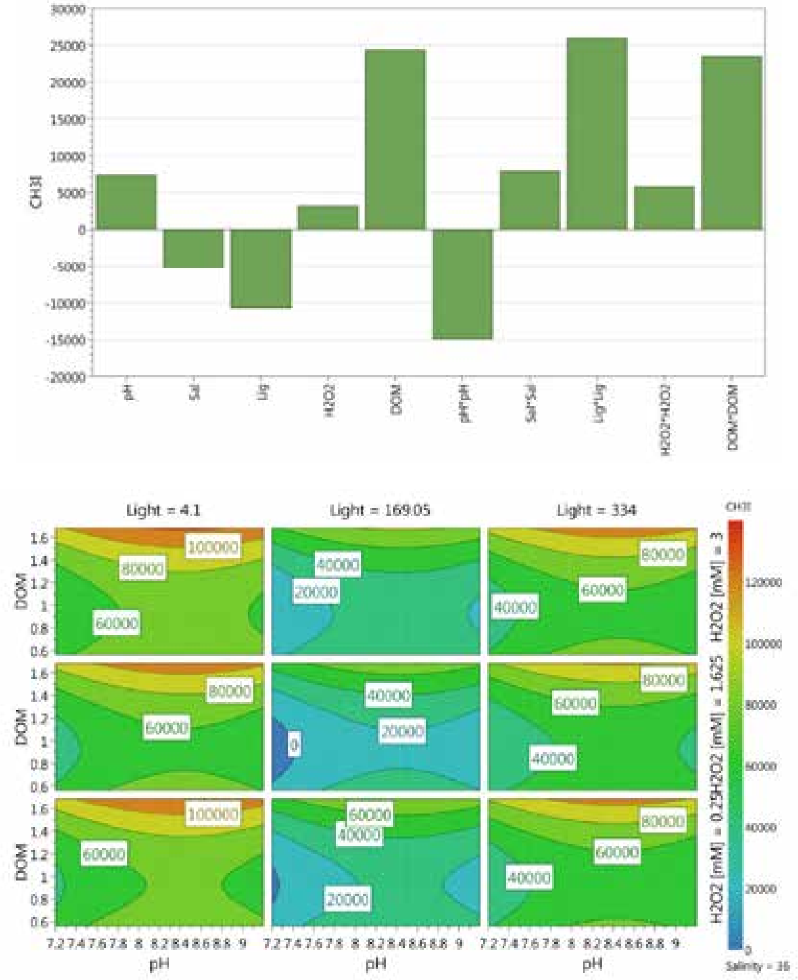


Figure 4.5. Coefficient (top) and response contour (bottom) plots showing the influence of five different factors on formation of CH_3I. The contour plot shows the influence of four factors at salinity 36 PSU on the production of CH_3I.

The relationship between CH_3I formation and DOM was more complicated since the enzymatic pathway differs from the other VHOC through the methylating enzymes. However, the contour and coefficient plots reveal that the formation of CH_3I was positively dependent on DOM in a convex pattern while having a somewhat weaker negative dependence on pH in a concave pattern. The involvement of DOM could be attributed to abiotic hydrolysis of halogenated humic acid, which can result in the production of CH_3X[358, 368]. Although the humic acid molecule indicates susceptibility to both acidic and basic hydrolysis, it is not known whether the hydrolysis of halogenated DOM species is fast enough to have occurred within the frame of a 3 hour incubation. Production of CH_3I increased roughly by two orders of magnitude after 3 hours of incubation, suggesting that production of this VHOC was linked to biological processes by algae and/or bacteria, both of which have been identified as sources of CH_3I[239, 281]. The optimal combination of factors for highest CH_3I production was found to be at pH 8 to 8.4, highest humic acid concentration, and low and high H_2O_2 concentrations, and lowest irradiance. The influence of the linear and the quadratic H2O2*H2O2 terms was low compared to other factors. As one of the factors contributing to highest CH_3I production was low irradiance, this author stipulates that production of this VHOC was not solely connected to photosynthesis. Nightingale[292] observed that for the brown algae *Asocphyllum nodosum* high irradiance did not result in increased CH_3I emissions, which seems to suggest that the production of CH_3I was not strictly linked to photosynthesis. On other hand, phytoplankton and cyanobacteria have been identified as main contributors of CH_3I[466].

At medium irradiance (169 µmol photons m^{-2} s^{-1}), the production of CH_3I decreased and halted at low pH, but increased again at highest irradiance (334 µmol photons m^{-2} s^{-1}). This suggests that if photosynthesis was involved, its contribution would stem primarily from a stress-related response to high irradiance, similarly to that observed for $CHBr_3$. The decrease in production of CH_3I due to low pH has been previously noted in the literature[329], and was reflected by the negative quadratic term pH*pH in **Figure 4.5** which indicates a decrease at both low and high pH values. A possible contribution to CH_3I production from algae not associated with H_2O_2-dependent pathways could be the production of CH_3X by SAM-S-HG enzymes. In near-dark conditions (4 µmol photons m^{-2} s^{-1}) the CH_3I production may be instead explained by bacteria[351-353]. This is feasible as the algae were not pre-treated in any way before being incubated – therefore, any bacterial strains associated with the *Fucus serratus* would have been present in the sample. The results observed by this author were therefore in line with the observations made by Manley and Dastoor[448] and Klein[281], which identified CH_3I production as stemming both from the algae themselves and from the associated bacterial colonies. In particular, CH_3I production was associated with cellular destruction caused by bacteria, possibly because of the involvement of bacterial nonspecific methyltranspherases[382]. *Vibrio* spp., which was shown by Klein[281] to produce CH_3I, are widespread in coastal waters in Sweden, including the sampling site of *Fucus serratus* that was used for experiments here[467]. These bacteria are thallus associated and are responsible for the degradation of components of the macroalgal cell walls[468], including fucoidan, a polysaccharide found only in brown seaweeds. As the degradation of algal cell walls and the bacterial production of CH_3I have not been indicated to be light dependent, the author hypothesises that the maximum at low irradiance could be caused by algae-associated bacteria.

CONCLUSIONS

The DS design clearly showed that production of VHOCs is not as straightforward as it seems. The connection between increased pH and increased production of brominated VHOC and CH_2I_2 has been shown in earlier laboratory experiments, where for instance an increase from pH 8.0 to 8.8 led to higher $CHBr_3$, $CHBr_2Cl$ and CH_2I_2 levels[330]. The nonlinear formation was indicated in this investigation, where a lowered light intensity decreased the amounts formed, even if light intensity has a large impact on photosynthesis, and, thereby, on the formation of H_2O_2. In this investigation, there was a positive linear dependence between pH and VHOC production, but in general a negative correlation with the quadratic term pH*pH. A decrease in production of CH_3I due to low pH has also been previously noted in the literature[329].

Surprisingly, most of the VHOC did not show a high dependency on light intensity contrary to earlier studies. Investigations have shown that the increase in production with increasing light intensity could be due to light induced stress, which in turn increased the production of H_2O_2 and the activity of V-HPOs. The dependency on H_2O_2 concentration was shown to have strong association with the production of brominated VHOCs[460]. Contrary to what was found here, the evidence in the literature states that higher irradiance causes higher production of VHOCs[292, 322, 372, 459]. For some of the other VHOCs, the influence of irradiance was more pronounced, which could indicate competing formation/degradation mechanisms. For CH_2I_2 it should be noted that there was a possible degradation mechanism as it has been reported to occur within the time frame of the 3 hour incubation[462].

The influence of H_2O_2 is related to light-induced stress, and it has been shown that algae respond rapidly to light-induced stress with an increased activity of V-HPO[286],[330, 459], and thereby with the formation of HOX. In the case of the research reported here, the influence of H_2O_2 was not at all clear cut. The responses for the individual VHOC all show different dependencies with both negative and positive linear relationships as well as with positive and negative H2O2*H2O2 terms. This indicates that more factors have been involved in the production of VHOCs which were not considered in the design space.

Another uncertain influence on the observed results was the role of bacteria. The dependence on low irradiance for the majority of VHOCs could indicate that bacterial production and/or degradation of VHOC occurred in the experiments. This is a possibility since the algae and the ASW medium were not anexic.

Although the author of this thesis attempted to connect the observed behaviour to existing knowledge regarding production and degradation mechanisms, it was obvious that the available knowledge on how certain VHOCs behave is insufficient. The results obtained here indicate that some production mechanisms, such as the high production of $CHBr_3$ caused by high irradiance, may depend on abiotic factor in a different manner than previously established in the literature. Similarly, some of the achieved results contradicted already established mechanisms or indicated that previously unknown mechanisms may have been involved. These discrepancies were interpreted by the author as stemming from nonlinear factor interactions, and possibly interactions that were not considered by the DS design, i.e. two- or multifactor cross-terms (for example Sal*pH). As the study of this type was unprecedented, previously published results indicate a fallacy in how biogenic VHOC emissions should be considered. The results of DoE presented here should serve as a strong incentive to encourage the multivariate paradigm in VHOC-dedicated research. The design used here was unfortunately not able to assist in distinguishing which mechanism could have caused an increase or decrease in VHOC production. For the majority of analysed VHOCs, the salinity seemed to influence *Fucus serratus* less on the whole compared to, for instance, humic acid. There were some similarities in production patterns (i.e. response contour plots), which may suggest that these VHOCs follow similar or co-dependent production mechanisms. An example of such similarity was the production of CH_2BrI and $CHCl_2Br$, which followed an identical pattern (see contour plots in **Appendix 4.2**). CH_3I have been shown by this enquiry to follow a completely different formation pattern, which was consistent with earlier reported results in the literature.

Keng et al.[239] point out that the majority of incubation studies are short term, lasting from 30 minutes to 3 months. Such short-term stresses cannot be representative of the full life-cycle of algae and in turn how that influences the dynamics of VHOC production. In our study, the short-term incubation and other resource-limiting factors produced results which are far from exhaustive. However, the insights provided here are the first to illustrate previously unknown relationships between various environmental factors. In response to the need of more research regarding multifactor interactions[239], the principle behind our methodology can be germinated into a wide range of future laboratory and mesocosm studies devoted to elucidating the effects on biogenic VHOC production. In addition to the benefit of studying multifactor interactions, a usage of DoE and other multivariate methods may contribute to more standardised procedures within the field. Further, the DS analysis also indicated that many mechanisms could be dependent or co-dependent on the production and degradation by bacterial communities associated with algae. This calls for more research regarding the contribution of bacteria to emissions of VHOCs from the seas, as too few mechanisms are currently known.

It is the hope of the author that the knowledge amassed from future multivariate laboratory and mesocosm studies could be then used to study the production of VHOCs in the algae's natural environment.

APPENDICES FOR CHAPTER 4

APPENDIX 4.1 – THE DESIGN MATRIX

Exp No	Run Order	pH	Salinity	Light	H2O2	DOM
1	3	8.2	50	334	0.25	0.56
2	9	8.2	22	4	3	1.68
3	11	9.2	36	4	0.25	1.68
4	1	7.2	36	334	3	0.56
5	4	9.2	22	169	3	0.56
6	6	7.2	50	169	0.25	1.68
7	7	9.2	22	334	1.63	1.68
8	8	7.2	50	4	1.63	0.56
9	5	9.2	50	334	3	1.12
10	10	7.2	22	4	0.25	1.12
11	2	8.2	36	169	1.63	1.12

Table 4-A. The DS design matrix consisting of 10 experimental runs and 1 centre point run (exp No 11). Experiment number 1 could not be measured due to instrumental error. The factors considered were pH, salinity (psu), irradiance (light, μmol photons m^{-2} s^{-1}), H_2O_2 concentration (mM) and humic acid concentration (DOM, mM).

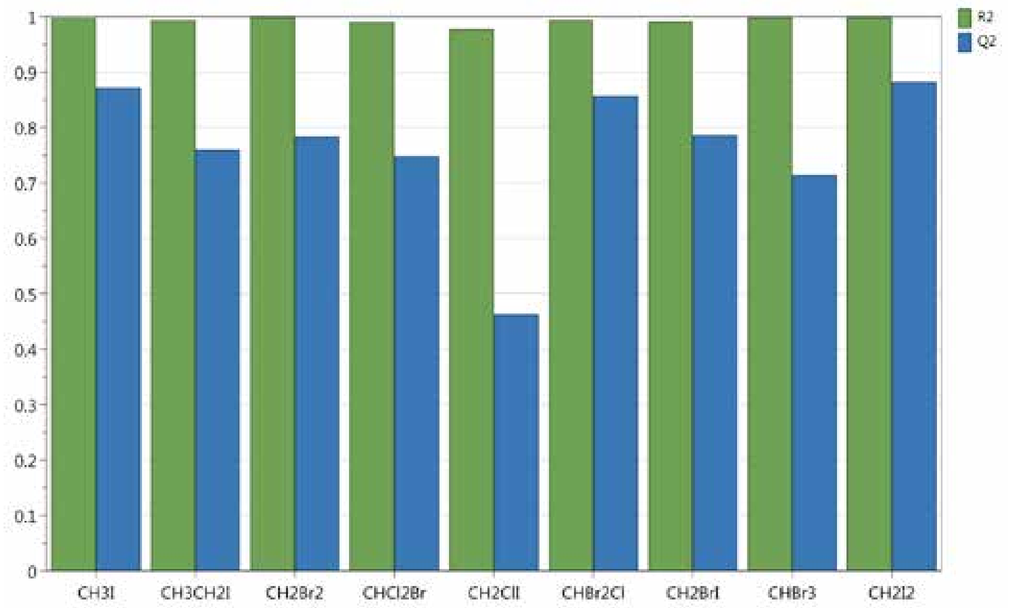


Figure 4-A. Summary of fit for all VHOCs considered in the DS design. Green columns show R^2 values and blue the Q^2 values.

	R^2	Q^2
CH^3I	0.99778	0.871397
CH^3CH^2I	0.993851	0.759485
CH^2Br^2	0.999637	0.783173
$CHCl^2Br$	0.990333	0.748869
CH^2ClI	0.977219	0.463065
$CHBr^2Cl$	0.994105	0.857217
CH^2BrI	0.991355	0.786265
$CHBr^3$	0.999009	0.714184
CH^2I^2	0.999538	0.882499

Table 4-B R^2 and Q^2 value corresponsing to Figure 4-A.

Appendix 4.2 – Response Contour Plots for Other VHOCs

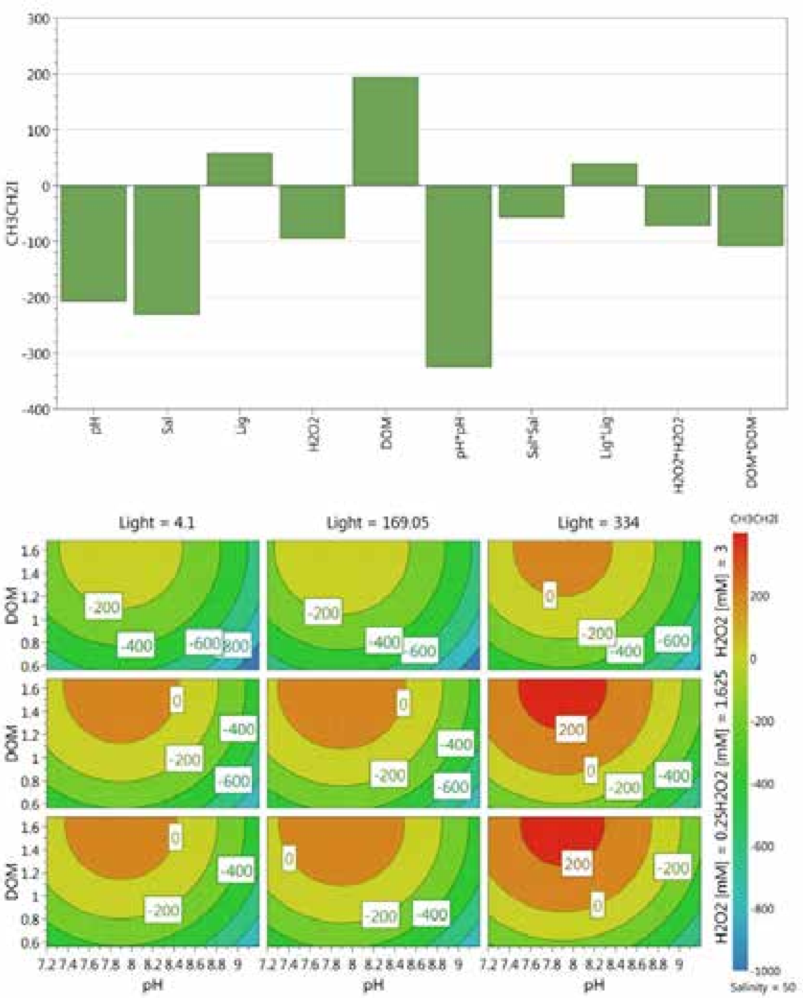


Figure 4-B. Coefficient plot (top) and contour plot (bottom) for CH_3CH_2I.

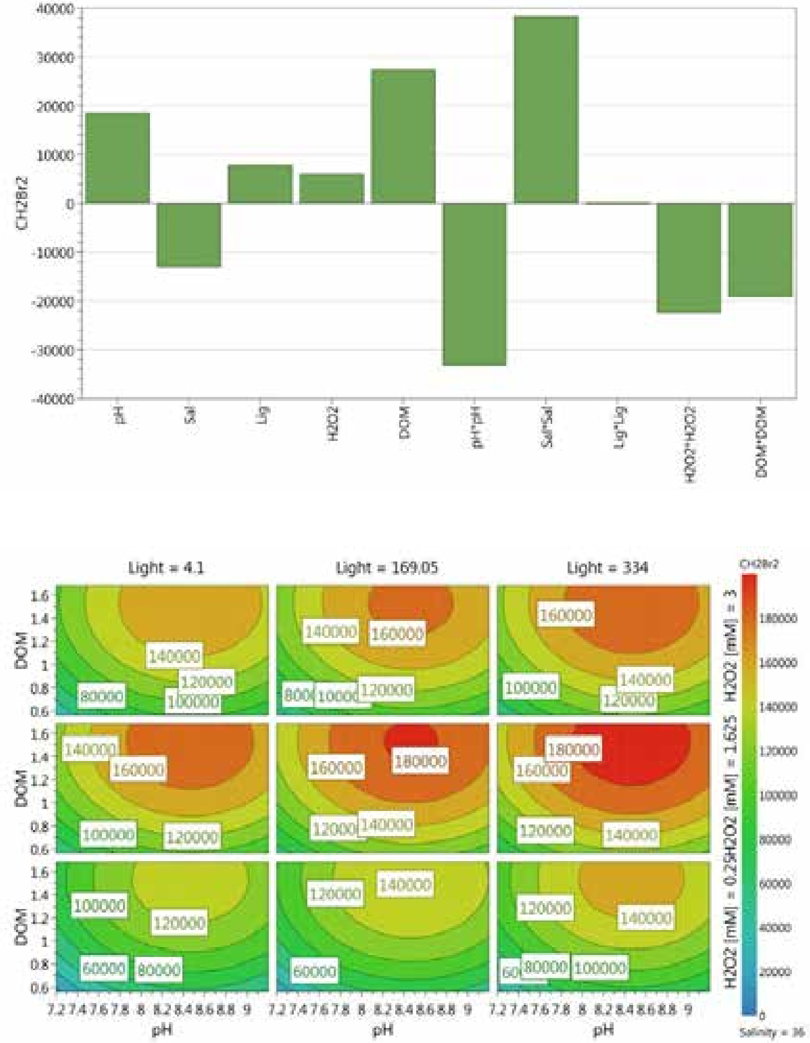


Figure 4-C. Coefficient plot (top) and contour plot (bottom) for CH_2Br_2.

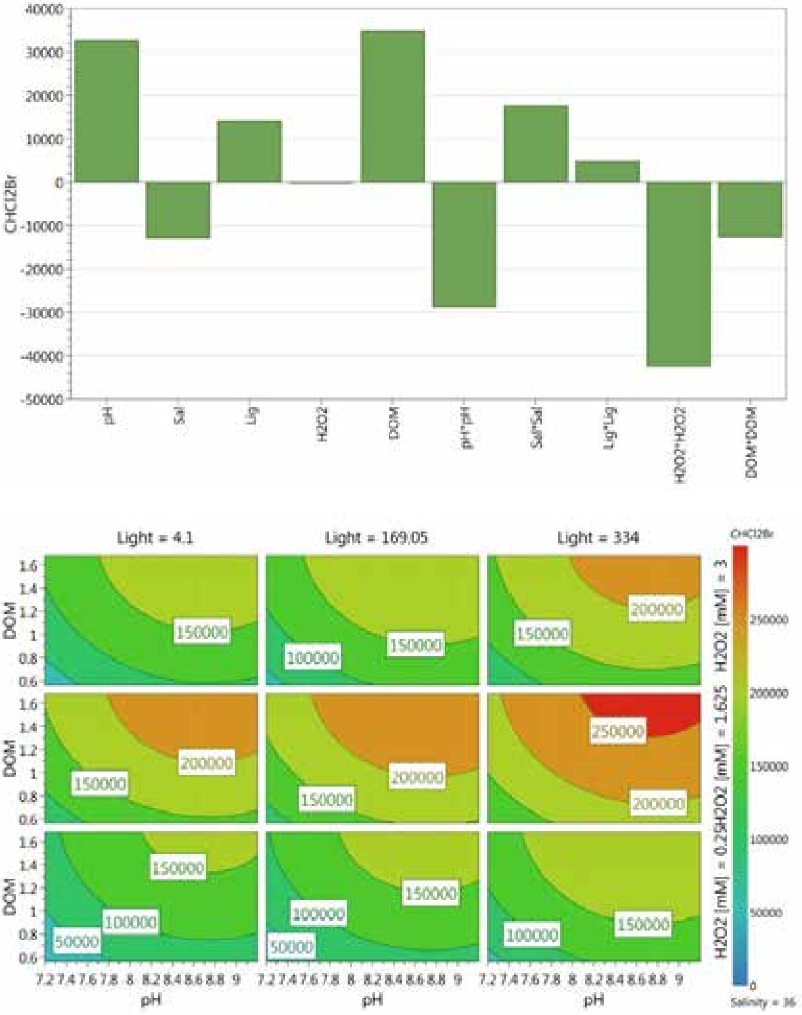


Figure 4-D. Coefficient plot (top) and contour plot (bottom) for $CHCl_2Br$.

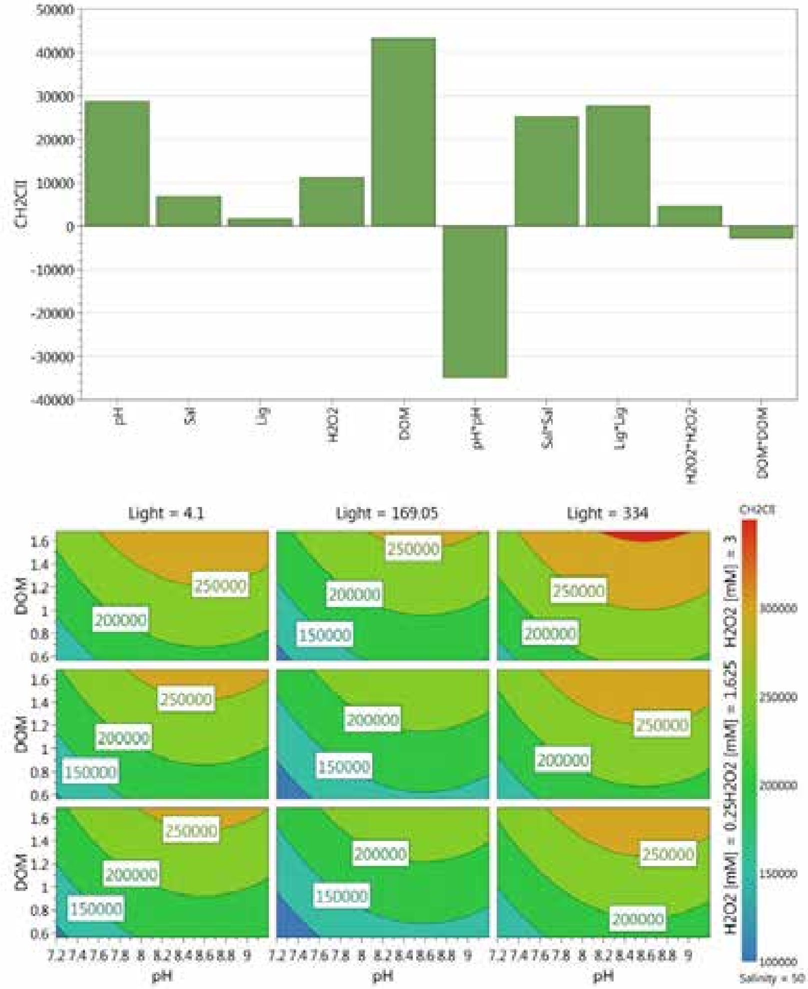


Figure 4-E. Coefficient plot (top) and contour plot (bottom) for CH_2ClI.

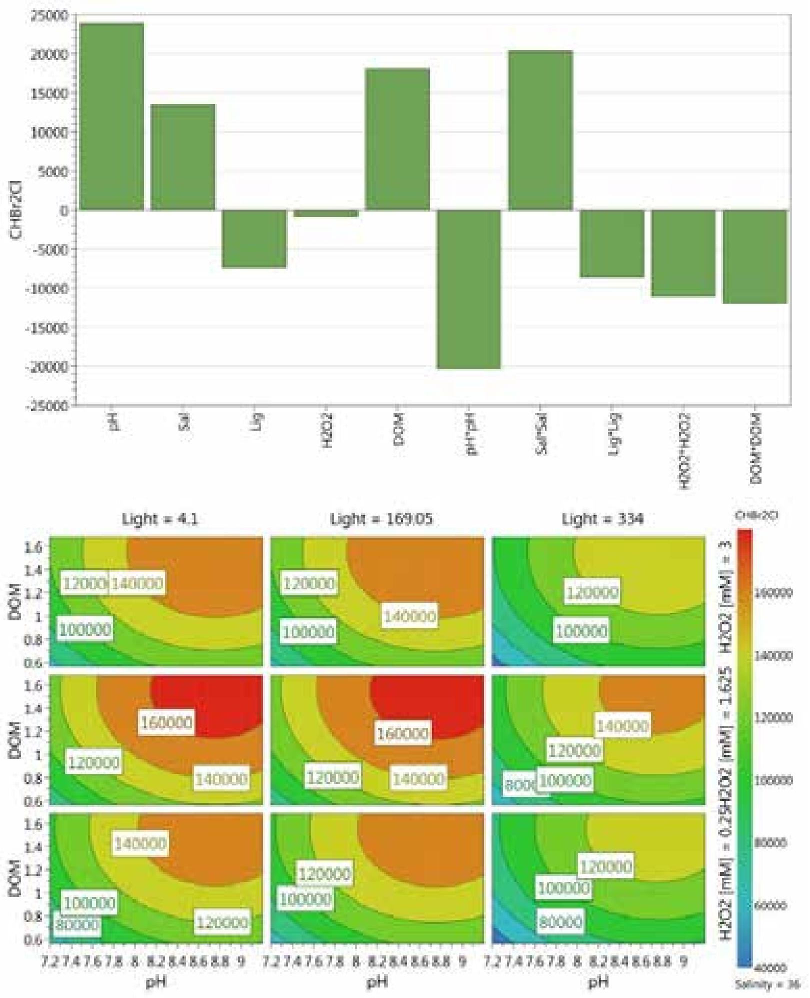


Figure 4-F. Coefficient plot (top) and contour plot (bottom) for $CHBr_2Cl$.

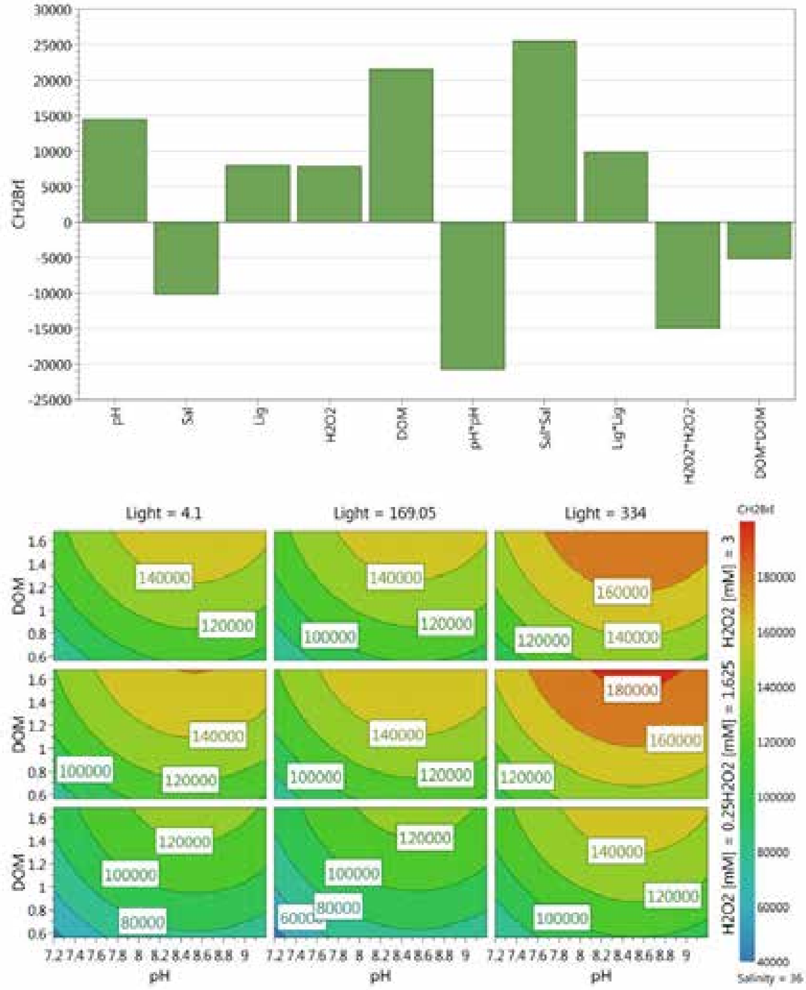


Figure 4-G. Coefficient plot (top) and contour plot (bottom) for CH_2BrI.

CONCLUSIONS AND LOOKING TO THE FUTURE

My name is Truth and I am the most elusive captive in the universe.
Carl Sandburg

The research presented in this thesis represents a comprehensive contribution to understanding how complex biological matrices and environmental phenomena can be approached through machine learning methods. The achieved results have shown several benefits of using DoE and linear projection algorithms in the study of complex multicomponent systems. In addition, the results presented here were of assistance in identifying several knowledge gaps, while also contributing to a deeper understanding. Several issues found in machine learning methods when applied to data generated from complex measurements have been in focus in this thesis, such as non-selectivity and nonlinearity; how these issues impact the performance of the algorithms; and how methods should be developed to address them.

Although the nonlinearity issue was not resolved with linear projection algorithms per se, the author concludes that these algorithms are suitable tools in identifying nonlinear behaviour. From an analytical point of view, applying such algorithms before turning to more complex nonlinear algorithms (e.g. neural networks, regression trees, etc.) allows for experiments that are more resource-efficient. The balancing act between generating information that is of as high quality as possible and aiming to reduce resource consumption is, after all, one of the social responsibilities of the modern analytical chemist[469-470]. Linear projection algorithms perform most strongly, however, if the experimental space was statistically designed and if, when interpreting, several different algorithms are used in combination. In spectral analysis particularly, such interpretative approach is of great benefit, as it gives a more accurate representation of the analyte and matrix behaviours – as has been shown throughout this thesis, linear algorithms combined with DoE facilitate the detection of hidden peaks, nonlinearities, false-positives, and complex interactions between chemical species. This kind of EDA-based thinking is in stark contrast to a common fallacy in spectroscopy, where a few peaks are analysed or 'convenient' spectral regions are examined, thus missing potential false-positives and other features which threaten the validity of a method.

Unfortunately, the insights linear projections algorithms offer are limited if the studied system is too complex, regardless of whether DoE was used or not. This was illustrated in Chapters 1 and 3, which shed light onto the possibilities and shortcomings of T-OPLS and also were the first thorough examination of this algorithm. T-OPLS was like OPLS capable in separating overlaps. In addition, as was shown in Chapter 3, T-OPLS offers more in-depth information for qualitative spectral interpretation compared to solely using PCA and/or OPLS. T-OPLS is therefore an excellent tool for highlighting hidden spectral features and complex matrix effects. T-OPLS was, however, not deemed capable of pinpointing where the nonlinearities originate and how they should be interpreted. The presence of co-dependencies in turn caused the issue of false-positives, which suggests that T-OPLS was not sufficiently selective unless certain conditions were met. In Chapter 1, the non-selectivity between the IS and the analytes, DOP and DOX, was caused in part by the lack of statistical design, thus making T-OPLS unsuitable. The nature of the performed analysis would not have made DoE possible. However, the research reported in Chapter 1 illustrates how a method can fall short if no proper design is in place.

A designed experimental space was analysed with T-OPLS in Chapter 3. Although the analysis of the design made it clear that several peaks followed nonlinear behaviour, T-OPLS could not be fully relied on to corroborate the observations made in the analysis of DoE. The design did however contribute to the understanding of where the co-dependencies may have originated – something T-OPLS alone could not do. The investigation of T-OPLS reported here was by no means exhaustive. More studies where the experiments are statistically designed would be of great benefit in understanding T-OPLS – in particular, such a design would clarify whether the co-dependency in signals stems from linear or nonlinear sources. The author of this thesis strongly recommends a systematic examination of T-OPLS using simulated Raman spectral responses, which follow a DoE matrix. The results in Chapters 1 and 3 stress yet again the need for designed experiments to facilitate interpretation of calculations performed by linear projection algorithms, especially in systems with low SNR and nonlinear effects. A possible future way of working to get fully functional nonlinear modelling with Raman and SERS data would be to have a well-developed general nonlinear deep-learning ANN model. In order to generate enough data for such a model, it could be built through a collaborative effort of academia and industry. This model would supply the foundation for how Raman spectra behave in different situations. To adapt to the new research task, it would be possible to add a couple of hundreds of spectra preferably from designed experiments and to retrain the last layers of the ANN. Alternatively, the last layers of the ANN may be replaced by an OPLS model as a shortcut to interpretability. This way of working already exists for image analysis and is called transfer learning.

DoE is a cornerstone for increased understanding of complex systems and phenomena in general and lends itself to the study of both linear and nonlinear systems in a well-defined and efficient manner. Similar to the sentiment expressed during the Faraday Discussions[1], the author of this thesis wished to put more emphasis on the importance of experimental design and on knowledge of the sample matrix. This was achieved by highlighting the benefits of a designed experimental space in Chapters 2 and 4. In the case of Chapter 2, a mixture design combined with PLS proved to be a robust method for quantification of an analyte in a complex biological matrix. In Chapter 4 DoE helped to illustrate the complexity of an environmental phenomenon in such a way previously assumed knowledge may be called into question. The research in Chapter 4 in particular shows that even for a natural system of high complexity, it is possible to construct representative design spaces. Although this particular application was a short-termed laboratory study, it is more than plausible to use DoE on a much larger scale, both for laboratory, mesocosm, and even for field studies. In situations where designed experiments are in need of complementary tools, it would be ideal to have nonlinear modelling methods that provide easily interpretable results. One obstacle in the experiment interpretations has been that such methods have not been readily available during the thesis work.

One of the basic premises behind DoE is the reduction of the required experimental runs; application of DoE should therefore ameliorate the practical disadvantages of large scale, long-term environmental studies. Further, DoE offers a more comprehensive and objective interpretation tool, which provides a solid statistical foundation. The multivariate nature of DoE along with the aforementioned benefits makes it particularly attractive for the study of natural phenomena and complex biological matrices. The author of this work hopes therefore for the expanded implementation of statistically designed experimental spaces in environmental sciences and in analytical spectroscopy.

ACKNOWLEDGEMENTS

I swear by my pretty flower bonnet, I will end you.
Malcolm Reynolds, *Firefly*

Said and done, Mal – I have now ended this thesis. It is almost a rueful feeling, really. So much has been done and yet there is so much left to do, so many questions I wish to have had time to answer.

As (hopefully) is evident from this thesis, there were numerous proverbial bees in my bonnet. Now that the majority of them have been crushed into ink with which this thesis was written (not my blood, I promise), I can finally take off my bonnet in gratitude to those who made this journey possible.

First, I would like to extend my warmest thanks to my supervisors, Prof. Katarina Abrahamsson and Dr. Mats Josefson. Katarina, thank you for your patience and never-failing high spirits, which complemented (and sometimes negated) my brooding pessimism. Mats, I never thought I would say this, but I have grown to love multivariate statistics so much that I am not sure how I will manage without them now that my academic carrier is over.

I would like to extend my thanks to all the students I had the pleasure to teach during advanced analytical chemistry courses in the years during and prior to my PhD appointment. Thanks to you, I have discovered the joy and fulfilment that pedagogics can provide. A was also fortunate to supervise multiple bachelor and master students, whose take on research topics I dedicated my PhD to has been inspiring – Annie Jonson, Boban Kesic, Saba Kazemi, Taymaa Yalil, Johan Lewing, thank you all! I would like to extend a special thanks to the bachelor student Anton Wimnell Macke, who provided interesting insights to the ALL project and has become a dear friend; to the master student Åke Henrik-Klemens, whose master thesis work was invaluable both to my own research and to the field of wood conservation, and, lastly; to the master student Kelebileone Kaisara, with whom we were first to introduce multivariate statistical design into the field of biogenic VHOC formation.

Thanks to MDr. Mats Bemark and MDr. Jonas Abrahamsson who were involved in the ALL project. Another thank you goes to Charlotte Björdal, who co-supervised the PEG project, and provided wood conservation expertise together with Åke. I also would like to thank Olga Kroutchenko, whose assistance as research engineer was much appreciated throughout a number of experimental conundrums.

Although this may seem odd, I would like to thank Miyazaki Hidetaka and FROMSOFTWARE for creating the Dark Souls, Bloodborne, and Sekiro: Shadows Die Twice video games. I do not know what I would have done without those masterpieces of games to cool off my boiling brain. 本当にありがとうございます!

Last but not least, my perhaps warmest and most heartfelt gratitude is dedicated to my family. Mike, I do not know what I would have done without your assistance with correcting my thesis. So, thank you, my dear no-nonsense father-in-law! I know better now than to use ‘viz.’, ‘herein’, and a range of Americanisms. Mother, my dearest friend and my most avid supporter – many years were bleak but the years that followed have become ever so brighter thanks to you. Nicholas, my husband, the light of my life, you make this life worth living, every hardship worth to endure. Kira, my little star, you are not born yet, but you should know how much your presence has helped me to pull this off. Maybe one day you will read this thesis and actually understand it better than I do, bemoaning a lack of this and that. Finally, a thank you to my cat, who interrupted my nightly thesis writing by sitting next to the bedroom entrance and calling me incessantly until I came to bed. And for all the unconditional fluffy cuddles, of course.

REFERENCES

1. Goodacre, R., The blind men and the elephant: challenges in the analysis of complex natural mixtures. *Faraday Discussions* **2019,** *218* (0), 524-539.
2. Brereton, R. G.; Jansen, J.; Lopes, J.; Marini, F.; Pomerantsev, A. L.; Rodionova, O. Y.; Roger, J. M.; Walczak, B.; Tauler, R., Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. *Analytical and Bioanalytical Chemistry* **2017,** *409* (25), 5891-5899.
3. Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L., Computerized learning machines applied to chemical problems. Molecular formula determination from low resolution mass spectrometry. *Analytical Chemistry* **1969,** *41* (1), 21-27.
4. Wold, S., Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems* **1995,** *30* (1), 109-115.
5. Esbensen, K.; Geladi, P., The start and early history of chemometrics: Selected interviews. Part 2. *Journal of Chemometrics* **1990,** *4* (6), 389-412.
6. Geladi, P.; Esbensen, K., The start and early history of chemometrics: Selected interviews. Part 1. *Journal of Chemometrics* **1990,** *4* (5), 337-354.
7. Brereton, R. G., A short history of chemometrics: a personal view. *Journal of Chemometrics* **2014,** *28* (10), 749-760.
8. Jurs, P. C.; Isenhour, T. L., *Chemical applications of pattern recognition*. Wiley: New York, 1975.
9. Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L.; Reilley, C. N., An investigation of combined patterns from diverse analytical data using computerized learning machines. *Analytical Chemistry* **1969,** *41* (14), 1949-1953.
10. Malinowski, E. R.; Weiner, P. H., Factor analysis of solvent shifts in proton magnetic resonance. *J Phys Chem-Us* **1970,** *74*, 4537-4542.
11. Massart, D. L.; Janssens, C.; Kaufman, L.; Smits, R., Application of the theory of graphs to the optimization of chromatographic separation schemes for multicomponent samples. *Analytical Chemistry* **1972,** *44*, 2390-2393.
12. Nalimov, V., *Preminenie matematicheskoi statistiki pri analize veshestva (Application of mathematical statistics in the analysis of substaces)*. Fizmatlit: Moscow, 1960.
13. Fisher, R. A., Statistical Methods for Research Workers. In *Breakthroughs in Statistics: Methodology and Distribution*, Kotz, S.; Johnson, N. L., Eds. Springer New York: New York, NY, 1992; pp 66-70.
14. Pearson, K., LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901,** *2* (11), 559-572.
15. Fisher, R. A.; MacKenzie, W., Studies in crop variation II. The manurial response of different potato varieties. *Journal of Agricultural Science* **1923,** *13*, 311-320.
16. Wold, S.; Esbensen, K.; Geladi, P., Principal component analysis. *Chemometr. Intell. Lab.* **1987,** *2* (1-2), 37-52.
17. Kumar, N.; Bansal, A.; Sarma, G. S.; Rawal, R. K., Chemometrics tools used in analytical chemistry: an overview. *Talanta* **2014,** *123*, 186-199

18. Rodionova, O. Y.; Pomerantsev, A. L., Chemometrics: achivements and prospects. *Russian Chemical Reviews* **2006,** *75* (4), 271-287.
19. Kumar, R.; Sharma, V., Chemometrics in forensic science. *TrAC Trends in Analytical Chemistry* **2018,** *105*, 191-201.
20. Rajalahti, T.; Kvalheim, O. M., Multivariate data analysis in pharmaceutics: A tutorial review. *International Journal of Pharmaceutics* **2011,** *417* (1), 280-290.
21. van der Greef, J.; Smilde, A. K., Symbiosis of chemometrics and metabolomics: past, present, and future. *Journal of Chemometrics* **2005,** *19* (5-7), 376-386.

22. Trygg, J.; Lundstedt, T., Chemometric techniques in matabonomics. In *The handbook of metabonomics and metabolomics*, Lindon, J. C.; Nicholson, J. K.; Holmes, E., Eds. Elsevier B. V.: 2007.

23. Eriksson, L.; Antti, H.; Gottfries, J.; Holmes, E.; Johansson, E.; Lindgren, F.; Long, I.; Lundstedt, T.; Trygg, J.; Wold, S., Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Analytical and Bioanalytical Chemistry* **2004,** *380* (3), 419-429.

24. Musumarra, G.; Fichera, M., Chemometrics and cultural heritage. *Chemometrics and Intelligent Laboratory Systems* **1998,** *44* (1), 363-372.

25. Henrik-Klemens, Å.; Abrahamsson, K.; Björdal, C.; Walsh, A., An in situ Raman spectroscopic method for quantification of polyethylene glycol (PEG) in waterlogged archaeological wood. *Holzforschung* **2020,** (0), 20190238.

26. Davies, O. L., *Statistical methods in research and production*. Oliver and Boyd: London, 1947.

27. Leardi, R., Experimental design in chemistry: A tutorial. *Anal Chim Acta* **2009,** *652* (1-2), 161-72.

28. Wold, S.; Berglund, A.; Kettaneh, N., New and old trends in chemometrics. How to deal with the increasing data columes in R&D&P (research, development and production) - with examples from pharmaceutical research and process modelling. *Journal of Chemometrics* **2002,** *16*, 377-386.

29. NIST/SEMATECH, e-Handbook of Statistical Methods. 2012. https://www.itl.nist.gov/div898/handbook/index.htm (accessed 2020-04-30).

30. Gaukroger, S., *Objectivity: a very short introduction*. Oxford University Press: New York, 2012.

31. Daston, L.; Galison, P., *Objectivity*. Zone Books: New York, 2007.

32. Næs, T.; Isaksson, T.; Fearn, T.; Davies, T., *A user-friendly guide to multivariate calibration and classification*. NIR Publications: Chichester, UK, 2002.

33. Esbensen, K.; Schönkopf, S.; Midtgaard, T., *Multivariate analysis in practice: a training package*. Camo AIS: Trondheim, Norway, 1994.

34. Geladi, P., Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2003,** *58* (5), 767-782.

35. Bro, R., Multivariate calibration: What is in chemometrics for the analytical chemist? *Analytica Chimica Acta* **2003,** *500*, 185-194.

36. Mahalanobis, P., The first session of the Indian Statistical Conference. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **1938,** *4* (1).

37. Eriksson, L., *Design of Experiments: Principles and Applications*. Umetrics Academy: 2008.

38. Cavazzuti, M., Design of experiments. In *Optimization methods: From theory to scientific design and technological aspects in mechanics*, Springer Berlin Heidelberg: 2013; pp 13-42.

39. Zhang, Z. H.; Xiaofeng, B., Comparison about the three central composite designs with simulation. *International Conference on Advanced Computer Control : Icacc 2009 - Proceedings* **2009**, 163-167.

40. Jones, B.; Nachtsheim, C. J., A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology* **2011,** *43* (1), 1-15.

41. Olivieri, A. C., *Introduction to multivariate calibration: a practical approach*. Springer Nature: Switzerlang, 2018.

42. Wold, H., Nonlinear estimation by iterative least squares procedures. In *Research papers in statistics*, David, F., Ed. Wiley: New York, 1966; pp 411-444.

43. Kowalski, B. R., Chemometrics. *Analytical letters* **1978,** *11* (10), xi-xiii.

44. Loren, A. Quantitative Surface Enhanced Raman Spectroscopy. Chalmers University of Technology, Göteborg, 2004.

45. Trygg, J.; Wold, S., Orthogonal projections to latent structures (O-PLS). *J Chemometr* **2002,** *16* (3), 119-128.

46. Trygg, J.; Wold, S., Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* **2002,** *16*, 119-128.

47. Abbas, A. Exploration of Multivariate Evaluation Techniques in Surface Enhanced Raman Spectroscopy (SERS). Chalmers University of Technology, Göteborg, 2011.

48. Stenlund, H.; Johansson, E.; Gottfries, J.; Trygg, J., Unlocking interpratetion in near infrared multivariate calibrations by orthogonal partial least squares *Anal. Chem.* **2009,** *81*, 203-209.

49. Feudale, R. N.; Brown, S. D., An inverse model for target detection. *Chemometrics and Intelligent Laboratory Systems* **2005,** *77* (1-2), 75-84.
50. Abbas, A.; Josefson, M.; Abrahamsson, K., Characterization and mapping of carotenoids in the algae Dunaliella and Phaeodactylum using Raman and target orthogonal partial least squares. *Chemometrics and Intelligent Laboratory Systems* **2011,** *107* (1), 174-177.
51. Kvalheim, O. M.; Karstang, T. V., Interpretation fo latent-variable regression models. *Chemometrics and intelligent laboratory systems* **1989,** *7*, 39-51.
52. Abbas, A.; Josefson, M.; Nylund, G. M.; Pavia, H.; Abrahamsson, K., Chemical images of marine bio-active compounds by surface enhanced Raman spectroscopy and transposed orthogonal partial least squares (T-OPLS). *Anal Chim Acta* **2012,** *737*, 37-44.
53. Josefson, M.; Walsh, A.; Abrahamsson, K., Imaging and identification of marine algal bioactive compounds by surface enhanced Raman spectroscopy (SERS). In *Natural products from marine algae*, Stengel, D.; Connan, S., Eds. Humana Press: New York, 2015; Vol. 1308.
54. Dumarey, M.; Galindo-Prieto, B.; Fransson, M.; Josefson, M.; Trygg, J., OPLS methods for the analysis of hyperspectral images-comparison with MCR-ALS. *Journal of Chemometrics* **2014,** *28* (8), 687-696.
55. Souihi, N.; Nilsson, D.; Josefson, M.; Trygg, J., Near-infrared chemical imaging (NIR-CI) on roll compacted ribbons and tablets – multivariate mapping of physical and chemical properties. *International Journal of Pharmaceutics* **2015,** *483* (1), 200-211.
56. Hotelling, H., The Generalization of Student's Ratio. *The Annals of Mathematical Statistics* **1931,** *2* (3), 360-378.
57. Eliasson, C. Applications in Surface Enhanced Raman Spectroscopy. Chalmers University of Technology, Göteborg, 2004.
58. Gautam, R.; Vanga, S.; Ariese, F.; Umapathy, S., Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *Epj Techniques and Instrumentation* **2015,** *2* (8).
59. Geladi, P.; MacDougall, D.; Martens, H., Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat. *Applied Spectroscopy* **1985,** *39* (3), 491-500.
60. Bro, R.; Smilde, A. K., Centering and scaling in component analysis. *Journal of Chemometrics* **2003,** *17* (1), 16-33.
61. Jackson, J. E., *A user's guide to principal components*. John Wiley & Sons Inc: 1991.
62. van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J., Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **2006,** *7* (1), 142.
63. Brandt, N. N.; Brovko, O. O.; Chikishev, A. Y.; Paraschuk, O. D., Optimization of the rolling-circle filter for Raman background substraction. *Appl. Spectrosc.* **2006,** *60* (3), 288-293.
64. Mikhailyuk, I. K.; Razzhivin, A. P., Background subtraction in experimental data arrays illustrated by the example of Raman spectra and fluorescent gel electrophoresis patterns. *Instruments and Experimental Techniques* **2003,** *46* (6), 765-769.
65. Cobas, C., Applications of the Whittaker smoother in NMR spectroscopy. *Magn Reson Chem* **2018,** *56* (12), 1140-1148.
66. Zhang, Z. M.; CHen, S.; Liang, Y. Z., Baseline correction using adaptive iteratively reweighterd penalized least squares. *Analyst* **2010,** *135* (5), 1138-1146.
67. Eilers, P. H. C., A perfect smoother. *Analytical Chemistry* **2003,** *75* (14), 3631-3636.
68. Wise, B. M.; Roginski, R. T., A calibration model maintenance roadmap. *IFAC-PapersOnLine* **2015,** *48* (8), 260-265.
69. Wehry, W. L.; Mamantov, G., Matrix isolation spectroscopy. *Analytical chemistry* **1979,** *51* (6), 643-656.
70. Cho, S.; Chung, H., Investigation of chemometric calibration performace based on difference chemical matrix and signal-to-noise ratio. *Analytical Sciences* **2003,** *19*, 1327-1329.
71. Olivieri, A. C.; Faber, N. M.; Ferré, J.; Boqué, R.; Kalivas, J. H.; Mark, H., Uncertainty estimation and figures of merit for multivariate calibration (IUPAC Technical Report). *Pure and Applied Chemistry* **2006,** *78* (3), 633-661.
72. Allegrini, F.; Olivieri, A. C., IUPAC-consistent approach to the limit of detection in partial least-squares calibration. *Anal Chem* **2014,** *86* (15), 7858-66.

73. Byrne, H. J.; Kneif, P.; Keating, M. E.; Bonnier, F., Spectral pre and post processing for infrared and Raman spectroscpy of biological tissues and cells. *Chem. Soc. Rev.* **2016,** *45*, 1865-1878.
74. Mas, S.; de Juan, A.; Tauler, R.; Olivieri, A. C.; Escandar, G. M., Application of chemometric methods to environmental analysis of organic pollutants: A review. *Talanta* **2010,** *80* (3), 1052-1067.
75. Selectivity in Analytical Chemistry (IUPAC Recommendations 2001). **2002,** *24* (1), 21.
76. McNaught, A. D.; Wilkinson, A., IUPAC. Compendium of chemical terminology. In *The "Gold Book"* [Online] 2nd ed.; Blackwell Scientific Publications: Oxford, 1997. https://goldbook.iupac.org/terms/view/S05564 (accessed 2020-05-18).
77. Liang, Y.-Z.; Kvalheim, O. M.; Manne, R., White, grey and black multicomponent systems: A classification of mixture problems and methods for their quantitative analysis. *Chemometrics and Intelligent Laboratory Systems* **1993,** *18* (3), 235-250.
78. Monakhova, Y. B.; Astakhov, S. A.; Kraskov, A.; Mushtakova, S. P., Independent spectroscopic analysis of complex mixtures. *Chemometr. Intell. Lab.* **2010,** *103*, 108-115.
79. Johansson, M. P.; Kaila, V. R. I.; Sundholm, D., *Ab initio*, density functional theory, and semi-empirical calculations. In *Biomolecular Simulations: Methods and Protocols*, Monticelli, L.; Salonen, E., Eds. Humana Press: Totowa, NJ, 2013; pp 3-27.
80. Gremperline, P. J., Developments in nonlinear multivariate calibration. *Chemometr. Intell. Lab.* **1992,** *15*, 115-126.
81. Bentley, P. M.; Mcdonnell, J. T. E., Wavelet transforms - an introduction. *Electron Commun Eng* **1994,** *6* (4), 175-186.
82. Gendrin, C.; Roggo, Y.; Collet, C., Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review. *Journal of Pharmaceutical and Biomedical Analysis* **2008,** *48* (3), 533-553.
83. Loren, A.; Engelbrektsson, J.; Eliasson, C.; Josefson, M.; Abrahamsson, J.; Johansson, M.; Abrahamsson, K., Internal standard in surface-enhanced Raman spectroscopy. *Anal Chem* **2004,** *76* (24), 7391-5.
84. Stout, F.; Kalivas, J. H., Evaluation of multivariate calibration using a Tikhonov regularization approach and the generalized pair-correlation method with nonlinear data. *Analytical Letters* **2007,** *40* (6), 1227-1251.
85. Wachsmann-Hogiu, S.; Weeks, T.; Huser, T., Chemical analysis *in vivo* and in vitro by Raman spectroscopy--from single cells to humans. *Curr Opin Biotechnol* **2009,** *20* (1), 63-73.
86. Murphy, C. J.; Gole, A. M.; Hunyadi, S. E.; Stone, J. W.; Sicso, P. N.; Alkilany, A.; Kinard, B. E.; Hankins, P., Chemical sensing and imaign with metallic nanorods. *Chemical Communications* **2008**, 544-557.
87. Boerboom, G. P. M. Detection and anlysis of neutrophilic grunulocytes on opsonized gold islets with Raman spectroscopy. Bachelor Assignment, University of Twente, Netherlands, 2013.
88. Schaeberle, M. D.; Levin, I. W.; Lewis, E. N., Biological Vibrational Spectroscopic Imaging. In *Infrared and Raman Spectroscopy pf Biological Materials*, Gremlich, H.-U.; Yan, B., Eds. Marcel Dekker Inc.: New York, 2001; Vol. 24, pp 231-258.
89. Dieing, T.; Hollricher, O.; Toporski, J., *Confocal Raman microscopy*. Springer-Verlag Berlin, 2010.
90. Palonpon, A. F.; Ando, J.; Yamakoshi, H.; Dodo, K.; Sodeoka, M.; Kawata, S.; Fujita, M., Raman and SERS microscopy for molecular imaging of live cells. *Nature Protocols* **2013,** *8* (4), 677-692.
91. Mosier-Boss, P. A., Review of SERS substrates for chemical sensing. *Nanomaterials (Basel)* **2017,** *7* (6), 142-172.
92. Xu, W.; Ling, X.; Xiao, J.; Dresselhaus, M. S.; Kong, J.; Xu, H.; Liu, Z.; Zhang, J., Surface enhanced Raman spectroscopy on a flat graphene surface. *Proceedings of the National Academy of Sciences* **2012,** *109* (24), 9281-9286.
93. Sharma, B.; Frontiera, R. R.; Henry, A.-I.; Ringe, E.; van Duyne, R. P., SERS: Materials, applications, and the future. *Materials Today* **2012,** *15* (1), 16-25.
94. Sharma, B.; Fernanda Cardinal, M.; Kleinman, S. L.; Greeneltch, N. G.; Frontiera, R. R.; Blaber, M. G.; Schatz, G. C.; van Duyne, R. P., High-performance SERS substrates: Advances and challenges. *MRS Bulletin* **2013,** *38* (8), 615-624.
95. Fisk, H.; Westley, C.; Turner, N. J.; Goodacre, R., Achieving optimal SERS through enhanced experimental design. *J Raman Spectrosc* **2016,** *47* (1), 59-66.

96. Boisselier, E.; Astruc, D., Gold nanoparticle in nanomedicine: preparations, imaging, diagnostics, therapies and toxicity. *Chemical Society Reviews* **2009,** *38*, 1759-1782.
97. Vitol, E. A.; Orynbayeva, Z.; Friedman, G.; Gogotsi, Y., Nanoprobes for intracellular and single cell surface-enhance Raman spectroscopy (SERS). *Journal of Raman Spectroscopy* **2012,** *43*, 817-827.
98. Chaplin, T. D.; Clark, R. J. H.; Beech, D. R., Comparison of genuine (1851–1852 AD) and forged or reproduction Hawaiian Missionary stamps using Raman microscopy. *J Raman Spectrosc* **2002,** *33* (6), 424-428.
99. Bell, S. E. J.; Sirimuthu, N. M. S., Quantitative surfance-enhanced Rman spectroscopy. *Chemical Society Reviews* **2008,** *37*, 1012-1024.
100. Murphy, C. J.; Gole, A. M.; Hunyadi, S. E.; Stone, J. W.; Sisco, P. N.; Alkilany, A.; Kinard, B. R.; Hankins, P., Chemical sensing and imaging with metallic nanorods. *Chem Commun* **2008**, 544-557.
101. Jackson, J. B.; Westcott, S. L.; Hirch, L. R.; West, J. L.; Halas, N. J., Controlling the surface enhanced Raman effect via the nanoshell geometry. *Applied Physics letters* **2003,** *82* (2), 257-259.
102. Fraire, J. C.; Perez, L. A.; Coronado, E. A., Cluster size effects in the surface-enhanced Raman scattering response of Ag and Au nanoparticle aggregates: Experimental and theoretical insight. *J Phys Chem C* **2013,** *117* (44), 23090-23107.
103. Tantra, R.; Brown, R. J. C.; Milton, M. J. T., Strategy to improve the reproducibility of colloidal SERS. *J Raman Spectrosc* **2007,** *38* (11), 1469-1479.
104. Pazos-Pérez, N.; Ni, W.; Schweikart, A.; Alvarez-Puebla, R.; Fery, A.; Liz-Marzán, L., Highly uniform SERS substrates formed by wrinkle-confined drying of gold colloids. *Chem Sci* **2010,** *1*, 174-178.
105. Klutse, C. K.; Mayer, A.; Wirrkamper, J.; Cullum, B. M., Applications of Self-Assembled Monolayers in Surface-Enhanced Raman Scattering. *Journal of Nanotechnology* **2012,** *2012*, 10.
106. Nafie, L. A.; Freedman, T. B., Biological and Pharmaceutical Applications of Vibrational Optic Activity. In *Infrared and Raman Spectroscopy of Biolgical Materials*, Gremlich, H.-U.; Yan, B., Eds. Marcel Dekker Inc.: New York, 2001; Vol. 24, pp 15-54.
107. Evans, W. E.; Relling, M. V.; Rodman, J. H.; Crom, W. R.; Boyett, J. M.; Pui, C.-H., Conventional Compared with Individualized Chemotherapy for Childhood Acute Lymphoblastic Leukemia. *New England Journal of Medicine* **1998,** *338* (8), 499-505.
108. Staib, P.; Staltmeier, E.; Neurohr, K.; Cornely, O.; Reiser, M.; Schinköthe, T., Prediction of individual response to chemotherapy in patients with acute myeloid leukaemia using the chemosensitivity index Ci. *British Journal of Haematology* **2005,** *128* (6), 783-791.
109. Yang, Y.-L.; Lin, D.-T.; Chang, S.-K.; Lin, S.-R.; Lin, S.-W.; Chiou, R.-J.; Yen, C.-T.; Lin, K.-H.; Jou, S.-T.; Lu, M.-Y.; Chang, H.-H.; Chang, W.-H.; Lin, K.-S.; Hu, C.-Y., Pharmacogenomic variations in treatment protocols for childhood acute lymphoblastic leukemia. *Pediatric Blood & Cancer* **2010,** *54* (2), 206-211.
110. Byström, M.; Olsson, E. *Barncancerrapporten 2013 - om barncancerfondens satsningar på forskning*; Barncancerfonden: 2013.
111. Frost, B.-M.; Eksborg, S.; Björk, O.; Abrahamsson, J.; Behrendtz, M.; Castor, A.; Forestier, E.; Lönnerholm, G., Pharmacokinetics of Doxorubicin in Children with Acute Lymphoblastic Leukamia: Multi-Institutional Collaborative Study. *Medical and Pediatric Oncology* **2002,** *38*, 329-337.
112. Bro, R., Multivariate calibration: What is in chemometrics for the analytical chemist? *Analytica Chimica Acta* **2003,** *500* (1-2), 185-194.
113. Lorén, A.; Eliasson, C.; Josefson, M.; Murty, K. V. G. K.; Käll, M.; Abrahamsson, J.; Abrahamsson, K., Feasibility of quantitative determination of doxorubicin with surface-enhanced Raman spectroscopy. *Journal of Raman Spectroscopy* **2001,** *32* (11), 971-974.
114. Lorén, A.; Engelbrektsson, J.; Eliasson, C.; Josefson, M.; Abrahamsson, J.; Johansson, M.; Abrahamsson, K., Internal Standard in Surface-Enhanced Raman Spectroscopy. *Analytical Chemistry* **2004,** *76* (24), 7391-7395.
115. Kaczor, A.; Pilarczyk, M., Structural and spatial analysis of carotenoids in a single cell monitored by Raman spectroscopy. In *Optical Spectroscopy and Computational Methods in Biology and Medicine*, Baranska, M., Ed. Springer Netherlands: Dordrecht, 2014; pp 309-326.
116. Willets, K. A., Surface-enahance Raman scattering (SERS) for probing internal cellular structure and dynamics. *Anal Bioanal Chem* **2009,** *394*, 85-94.

117. Clark, D. A.; Henson, M.; Laplant, F.; Šašić, S., Pharmaceutical applications of chemical mapping and imaging. In *Applications of vibrational spectroscopy in pharmaceutical research and development*, Pivonka, D. E.; Chalmers, J. M.; Griffiths, P. R., Eds. John Wiley & Sons: London, 2007.

118. Šašić, S.; Clark, D. A., Defining a strategy for chemical imaging of industrial pharmaceultical samples on Raman line-mapping and global illumination instruments. *Applied Spectroscopy* **2006,** *60* (5), 494-502.

119. Xiao, R.; Zhang, X.; Rong, Z.; Xiu, B.; Yang, X. B.; Wang, C.; Hao, W.; Zhang, Q.; Liu, Z.; Duan, C.; Zhao, K.; Guo, X.; Fan, Y.; Zhao, Y.; Johnson, H.; Huang, Y.; Feng, X.; Xu, X.; Zhang, H.; Wang, S., Non-invasive detection of hepatocellular carcinoma serum metabolic profile through surface-enhanced Raman spectroscopy. *Nanomedicine: Nanotechnology, Biology and Medicine* **2016,** *12* (8), 2475-2484.

120. Gorzsás, A.; Stenlund, H.; Persson, P.; Trygg, J.; Sundberg, B., Cell-specific chemotyping and multivariate imaging by combined FT-IR microspectroscopy and orthogonal projections to latent structures (OPLS) analysis reveals the chemical landscape of secondary xylem. *The Plant Journal* **2011,** *66* (5), 903-914.

121. Gorzsás, A., Chemical imaging of xylem by Raman microscpectroscopy. In *Xylem*, de Lucas, M.; Etchhells, J., Eds. Humana press: New York, 2017.

122. Yue, X.; Li, H.; Tang, J.; Liu, J.; Jiao, J., Rapid and label-free screening of echinococcosis serum profiles through surface-enhanced Raman spectroscopy. *Analytical and Bioanalytical Chemistry* **2020,** *412* (2), 279-288.

123. Sanchez, L.; Pant, S.; Xing, Z.; Mandadi, K.; Kurouski, D., Rapid and noninvasive diagnostics of Huanglongbing and nutrient deficits on citrus trees with a handheld Raman spectrometer. *Analytical and Bioanalytical Chemistry* **2019,** *411* (14), 3125-3133.

124. Kestur, U. S.; Wanapun, D.; Toth, S. J.; Wegiel, L. A.; Simpson, G. J.; Taylor, L. S., Nonlinear optical imaging for sensitive detection of crystals in bulk amorphous powders. *Journal of Pharmaceutical Sciences* **2012,** *101* (11), 4201-4213.

125. Gendrin, C.; Roggo, Y.; Collet, C., Content uniformity of pharmaceutical solid dosage forms by near infrared hyperspectral imaging: A feasibility study. *Talanta* **2007,** *73* (4), 733-741.

126. Jovanovic, N.; Gerich, A.; Bouchard, A.; Jiskoot, W., Near-infrared imaging for studying homogeneity of protein-sugar mixtures. *Pharm Res* **2006,** *23* (9), 2002-13.

127. Draux, F.; Gobinet, C.; Sule-Suso, J.; Manfait, M.; Jeannesson, P.; Sockalingum, G. D., Raman imaging of single living cells: probing effects of non-cytotoxic doses of an anti-cancer drug. *Analyst* **2011,** *136* (13), 2718-25.

128. van Manen, H.-J.; Otto, C., Hybrid confocal Raman fluorescence microscopy on sngle cells using semiconductor quantum dots. *Nano Letters* **2007,** *7* (6), 1631-1636.

129. Chithrani, D. B., Intracellular uptake, transport, and processing of gold nanostructures. *Mol Membr Biol* **2010,** *27* (7), 299-311.

130. Gratton, S. E. A.; Ropp, P. A.; Pohlhaus, P. D.; Luft, J. C.; Madden, V. J.; Napier, M. E.; DeSimone, J. M., The effect of particle design on cellular internalization pathways. *Proceedings of the National Academy of Sciences* **2008,** *105* (33), 11613-11618.

131. Chithrani, B. D.; Ghazani, A. A.; Chan, W. C. W., Determining the Size and Shape Dependence of Gold Nanoparticle Uptake into Mammalian Cells. *Nano Letters* **2006,** *6* (4), 662-668.

132. Kneipp, J.; Kneipp, H.; McLaughlin, M.; Brown, D.; Kneipp, K., *In vivo* molecular probing of cellular compartments with gold nanoparticles and nanoaggregates. *Nano Letters* **2006,** *6* (10), 2225-2231.

133. Tang, H.-W.; Yang, X. B.; Kirkham, J.; Smith, D. A., Chemical probing of single cancer cells with gold nanoaggregates by surface-enhanced Raman scattering. *Applied Spectroscopy* **2008,** *62* (10), 1060-1069.

134. Rejman, J.; Oberle, V.; Zuhorn, I. S.; Hoekstra, D., Size-dependent internalization of particles via the pathways of clathrin- and caveolae-mediated endocytosis. *Biochemical Journal* **2004,** *377* (1), 159-169.

135. Huh, Y. S.; Chung, A. J.; Erickson, D., Surface enhanced Raman spectroscopy and its application to molecular and cellular analysis. *Microfluid Nanofluid* **2009,** *6* (3), 285-297.

136. Cialla, D.; März, A.; Böhme, R.; Theil, F.; Weber, K.; Schmitt, M.; Popp, J., Surface-enhanced Raman spectroscopy (SERS): progress and trends. *Analytical and bioanalytical chemistry* **2012,** *403* (1), 27-54.

137. Otto, A., Surface-enhanced Raman scattering: "Classical" and "Chemical" origins

In *Light Scattering in Solids IV* Cardona, M.; Güntherodt, G., Eds. Springer Berlin Heidelberg: 1984; Vol. 54, pp 289-418.

138. Strehle, K. R.; Cialla, D.; Rösch, P.; Henkel, T.; Köhler, M.; Popp, J., A reproducible surface-enhanced Raman spectroscopy approach. Online SERS measurements in a segmented microfluidic system. *Analytical Chemistry* **2007,** *79* (4), 1542-1547.

139. Dinish, U. S.; Yaw, F. C.; Agarwal, A.; Olivo, M., Development of highly reproducible nanogap SERS substrates: Comparative performance analysis and its application for glucose sensing. *Biosensors and Bioelectronics* **2011,** *26* (5), 1987-1992.

140. Khlebtsov, B. N.; Khanadeev, V. A.; Panfilova, E. V.; Bratashov, D. N.; Khlebtsov, N. G., Gold nanoisland films as reproducible SERS substrates for highly sensitive detection of fungicides. *ACS Applied Materials & Interfaces* **2015,** *7* (12), 6518-6529.

141. Yan, W.; Yang, L.; Chen, J.; Wu, Y.; Wang, P.; Li, Z., *In situ* two-step photoreduced SERS materials for on-chip single-molecule spectroscopy with high reproducibility. *Advanced Materials* **2017,** *29* (36), 1702893.

142. Perales-Rondon, J. V.; Colina, A.; González, M. C.; Escarpa, A., Roughened silver microtubes for reproducible and quantitative SERS using a template-assisted electrosynthesis approach. *Applied Materials Today* **2020,** *20*, 100710.

143. Lorén, A.; Engelbrektsson, J.; Eliasson, C.; Josefson, M.; Abrahamsson, J.; Abrahamsson, K., Self-Assembled Monolayer Coating for Normalization of Surface Enhanced Raman Spectra. *Nano Letters* **2004,** *4* (2), 309-312.

144. Lombardi, J. R.; Birke, R. L.; Lu, T.; Xu, J., Charge-transfer theory of surface enhanced Raman spectroscopy: Herzberg–Teller contributions. *The Journal of Chemical Physics* **1986,** *84* (8), 4174-4180.

145. Schlenoff, J. B.; Ming, L.; Ly, H., Stability and Self-Exchange in Alkanethiol Monolayers. *J Am Chem Soc* **1995,** *177* (50), 12528-12536.

146. Vericat, C.; Vela, M. E.; Benitez, G.; Carro, P.; Salvarezza, R. C., Self-assembled monolayers of thiols and dithiols on gold: new challenges for a well-known system. *Chemical Society Reviews* **2010,** *39*, 1805-1834.

147. Wong, K.; Kwon, K.-Y.; Rao, B. V.; Liu, A.; Bartels, L., Effect of halo substitution on the geometry of arenethiol films on Cu(111). *Journal of the American Chemical Society* **2004,** *126* (25), 7762-7763.

148. Holze, R., The adsorption of thiophenol on gold - a spectroelectrochemical study. *Phys Chem Chem Phys* **2015,** *14*, 21364-21372.

149. Holze, R., Competition of anchoring groups in adsorption on gold electrodes - a comparative spectroelectochemical study of 4-mercaptobenzonitrile and aromatic nitriles. *Journal of Solid State Electrochemistry* **2013,** *14*, 1869-1879.

150. Schkolnik, G.; Salewski, J.; Millo, D.; Zebger, I.; Franzen, S.; Hildebrandt, P., Vibrational start effect of the electric-field reporter 4-mercaptobenzonitrile as a tool for investigating electrostatics at electrode/SAM/solution interfaces. *International Journal of Molecular Sciences* **2012,** *13*, 7466-7482.

151. Gregas, M.; Yan, F.; Scaffidi, J.; Wang, H.-N.; Khoury, C.; Zhang, Y.; Vo-Dinh, T., *Tracking SERS-active nanoprobe intracellular uptake for chemical and biological sensing*. SPIE: 2007; Vol. 6755.

152. Juurlink, B. H. J.; Devon, R. M., Colloidal gold as a permanent marker of cells. *Experientia* **1991,** *47* (1), 75-77.

153. Lunov, O.; Syrovets, T.; Loos, C.; Beil, J.; Delacher, M.; Tron, K.; Nienhaus, G. U.; Musyanovych, A.; Mailänder, V.; Landfester, K.; Simmet, T., Differential uptake of functionalized polystyrene nanoparticles by human macrophages and a monocytic cell line. *ACS Nano* **2011,** *5* (3), 1657-1669.

154. Kettler, K.; Giannakou, C.; de Jong, W. H.; Hendriks, A. J.; Krystek, P., Uptake of silver nanoparticles by monocytic THP-1 cells depends on particle size and presence of serum proteins. *Journal of Nanoparticle Research* **2016,** *18* (9), 286.

155. Srijampa, S.; Buddhisa, S.; Ngernpimai, S.; Sangiamdee, D.; Chompoosor, A.; Tippayawat, P., Effects of gold nanoparticles with different surface charges on cellular internalization and cytokine responses in monocytes. *BioNanoScience* **2019,** *9* (3), 580-586.

156. Chhour, P.; Naha, P. C.; O'Neill, S. M.; Litt, H. I.; Reilly, M. P.; Ferrari, V. A.; Cormode, D. P., Labeling monocytes with gold nanoparticles to track their recruitment in atherosclerosis with computed tomography. *Biomaterials* **2016,** *87*, 93-103.

157. Bratneck, M.; Keul, H. A.; Singh, S.; Czaja, K.; DBornemann, J.; Bockstaller, M.; Moeller, M.; Zwaldo-Klarwasser, G.; Groll, J., Rapid uptake of gold annorods by primary human blood phagocytes and immunomodulatory effects of surface chemistry *American Chemical Society: Nano* **2010,** *4* (6), 3073-3086.
158. Rad, A.; Häggström, M., Development of blood cells. Hematopoiesis_(human)_diagram.svg, Ed. Wikipedia commons, 2014.
159. Ferris, D. P.; Lu, J.; Gothard, C.; Yanes, R.; Thomas, C. R.; Olsen, J.-C.; Stoddart, J. F.; Tamanoi, F.; Zink, J. I., Synthesis of biomolecule-modified mesoporous silica nanoparticles for targeted hydrophobic drug delivery to cancer cells. *Small* **2011,** *7* (13), 1816-1826.
160. Saltan, N.; Kutlu, H. M.; Hür, D.; Işcan, A.; Say, R., Interaction of cancer cells with magnetic nanoparticles modified by methacrylamido-folic acid. *Int J Nanomedicine* **2011,** *6*, 477-484.
161. Hainfeld, J. F.; Smilowitz, H. M.; O'Connor, M. J.; Dilmanian, F. A.; Slatkin, D. N., Gold nanoparticle imaging and radiotherapy of brain tumors in mice. *Nanomedicine* **2013,** *8* (10), 1601-1609.
162. Coulter, J. A.; Jain, S.; Butterworth, K. T.; Taggart, L. E.; Dickson, G. R.; McMahon, S. J.; Hyland, W. B.; Muir, M. F.; Trainor, C.; Hounsell, A. R.; O'Sullivan, J. M.; Schettino, G.; Currell, F. J.; Hirst, D. G.; Prise, K. M., Cell type-dependent uptake, localization, and cytotoxicity of 1.9 nm gold nanoparticles. *Int J Nanomedicine* **2012,** *7*, 2673-2685.
163. Patrick, G. L., *An introduction to medicinal chemistry*. 5 ed.; Oxford University Press: Oxford, 2013.
164. Avdeef, A., Absorption and Drug Development: Solubility, Permeability and Charge State. Second ed.; John Wiley & Sons, Inc.: USA and Canada, 2012. http://books.google.se/books?id=xzq3RtqGSK4C&printsec=frontcover&hl=sv#v=onepage&q&f=false (accessed 22nd of May, 2014).
165. Eliasson, C.; Loren, A.; Murty, K. V.; Josefson, M.; Kall, M.; Abrahamsson, J.; Abrahamsson, K., Multivariate evaluation of doxorubicin surface-enhanced Raman spectra. *Spectrochimica acta. Part A, Molecular and biomolecular spectroscopy* **2001,** *57* (9), 1907-15.
166. Yan, Q.; Priebe, W.; Chaires, J. B.; R., C., Interaction of Doxorubicin and Its Derivatives with DNA: Elucidation by Resonance Raman and Surface-Enhanced Resonance Raman Spectroscopy. *Biospectroscopy* **1997,** *3* (4), 307-316.
167. Lin, H.-H.; Li, Y.-C.; Chang, C.-H.; Liu, C.; Yu, A. L.; Chen, C.-H., Single nuclei Raman spectroscopy for drug evaluation. *Analytical Chemistry* **2012,** *84* (1), 113-120.
168. Huang, H.; Shi, H.; Feng, S. Y.; Chen, W. W.; Yu, Y.; Lin, D.; Chen, R., Confocal Raman spectroscopic analysis of the cytotoxic response to cisplatin in nasopharyngeal carcinoma cells. *Anal Methods-Uk* **2013,** *5* (1), 260-266.
169. Hartmann, K.; Becker-Putsche, M.; Bocklitz, T.; Pachmann, K.; Niendorf, A.; Rösch, P.; Popp, J., A study of Docetaxel-induced effects in MCF-7 cells by means of Raman microspectroscopy. *Analytical and Bioanalytical Chemistry* **2012,** *403* (3), 745-753.
170. Mignolet, A.; Wood, B. R.; Goormaghtigh, E., Intracellular investigation on the differential effects of 4 polyphenols on MCF-7 breat cancer cells by Raman imaging. *Analyst* **2018,** *143*, 258-269.
171. Krafft, C.; Knetschke, T.; Siegner, A.; Funk, R. H. W.; Salzer, R., Mapping of single cells by near infrared Raman microspectroscopy. *Vibrational Spectroscopy* **2003,** *32*, 75-83.
172. Ling, J.; Weitman, S. D.; Miller, M. A.; Moore, R. V.; Bovik, A. C., Direct Raman imaging techniques for study of the subcellular distribution of a drug. *Appl. Opt.* **2002,** *41* (28), 6006-6017.
173. Eliasson, C.; Loren, A.; Engelbrektsson, J.; Josefson, M.; Abrahamsson, J.; Abrahamsson, K., Surface-enhanced Raman scattering imaging of single living lymphocytes with multivariate evaluation. *Spectrochim Acta A Mol Biomol Spectrosc* **2005,** *61* (4), 755-60.
174. Feofanov, A. V.; Grichine, A. I.; Shitova, L. A.; Karmakova, T. A.; Yakubovskaya, R. I.; Egret-Charlier, M.; Vigny, P., Confocal Raman microspectroscopy and imaging study of theraphthal in living cancer cells. *Biophysical Journal* **2000,** *78* (1), 499-512.
175. Westerink, R. H. S.; Ewing, A. G., The PC12 cell as model for neurosecretion. *Acta Physiol (Oxf)* **2008,** *192* (2), 273-285.
176. Pande, S.; Jana, S.; Sinha, A. K.; Sarkar, S.; Basu, M.; Pradhan, M.; Pal, A.; Chowdhury, J.; Pal, T., Dopamine molecules on Aucore-Agshell bimetallic nanocolloids: Fourier transform infrared, Raman, and

surface-enhanced Raman spectroscopy study aided by density functional theory. *J Phys Chem C* **2009,** *133* (6989-7002).

177. Kitayama, S.; Shimada, S.; Xu, H.; Markham, L.; Donovan, D. M.; Uhl, G. R., Dopamine transporter site-directed mutations differentially alter substrate transport and cocaine binding. *Proceedings of the National Academy of Sciences* **1992,** *89* (16), 7782-7785.

178. Salomäki, M.; Marttila, L.; Kivelä, H.; Ouvinen, T.; Lukkari, J., Effects of pH and oxidants on the first steps of polydopamine formation: a thermodynamic approach. *The journal of physical chemistry. B* **2018,** *122* (24), 6314-6327.

179. Dumarey, M.; Galindo-Prieto, B.; Fransson, M.; Josefson, M.; Trygg, J., OPLS methods for the analysis of hyperspectral images - comparison with MCR-ALS. *Journal of Chemometrics* **2014**.

180. Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J., OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classificaiton. *Journal of Chemometrics* **2006,** *20*, 341-351.

181. Wang, J.; Trouillon, R.; Lin, Y.; Svensson, M. I.; Ewing, A., Individually addressable thin-film ultramicroelectrode array for spatial measurements of single vesicle release. *Analytical Chemistry* **2013,** *85* (11), 5600-5608.

182. Josefson, M.; Walsh, A.; Abrahamsson, A., Imaging and identification of marine algal bioactive compounds by surface enhanced Raman spectroscopy (SERS). In *Natural products from marine algae: methods and protocols*, Stengel, D. B.; Connan, S., Eds. Springer: New York, 2015; pp 365-374.

183. Hunter, J. D., Matplotlib: a 2D graphics environment. *Computing in Science and Engineering* 2007, pp 90-95.

184. Josefson, M. Microcomputer systems for real time control and measurements in potentiometric stripping analysis. University of Gothenburg, Gothenburg, 1983.

185. Jang, T. H.; Park, S. C.; Yang, J. H.; Kim, J. Y.; Seok, J. H.; Park, U. S.; Choi, C. W.; Lee, S. R.; Han, J., Cryopreservation and its clinical applications. *Integr Med Res* **2017,** *6* (1), 12-18.

186. Feliu, N.; Hassan, M.; Garcia Rico, E.; Cui, D.; Parak, W.; Alvarez-Puebla, R., SERS quantification and characterization of proteins and other biomolecules. *Langmuir* **2017,** *33* (38), 9711-9730.

187. Bemark, M., Personal communication concerning phagocytosis of lymphocytes. Walsh, A., Ed. 2013.

188. Farhane, Z.; Bonnier, F.; Byrne, H. J., Monitoring doxorubicin cellular uptake and traffiking using in vitro Raman microscopy: short and long exposure time effects on lung cancer cell lines. *Anal Bioanal Chem* **2017,** *409*, 1333-1346.

189. Das, G.; Nicastri, A.; Coluccio, M. L.; Gentile, F.; Candeloro, P.; Cojoc, G.; Liberale, C.; De Angelis, F.; Di Fabrizio, E., FT-IR, Raman, RRS measurements and DFT calculation for doxorubicin. *Microscopy Research and Technique* **2010,** *73* (10), 991-995.

190. Lee, C. J.; Kang, J. S.; Kim, M. S.; Lee, K. P.; Lee, M. S., The study of doxorubicin and its complext with DNA by SERS and UV-resonance Raman spectroscopy. *Bull Korean Chem Soc* **2004,** *25* (8), 1211-1216.

191. An, J. H.; El-Said, W. A.; Yea, C. H.; Kim, T. H.; Choi, J. W., Surface-enhanced Raman scattering of dopamine on self-assembled gold nanoparticles. *Journal of Nanoscience and Nanotechnology* **2011,** *11* (4424-4429).

192. Seasholtz, M. B.; Kowalski, B. R., Qualitative information from multivariate calibration models. *Applied Spectroscopy* **1990,** *44* (8), 1337-1348.

193. Sanchez, E.; Kowalski, B. R., Tensorial calibration I: first order calibration. *Journal of Chemometrics* **1988,** *2*, 247-263.

194. Walsh, A.; Josefson, M.; Abrahamsson, K., Method development of SERS and chemometrics for qualitative studies of *C. officinalis* vanadium bromoperoxidase. *Journal of Chemometrics* **2020,** *13*.

195. Björdal, G. C., Microbial degradation of waterlogged archaeological wood. *J. Cult. Herit.* **2012,** *13* (3, Supplement), S118-S122.

196. Eriksson, K.-E. L.; Blanchette, R. A.; Ander, P. Microbiological and enzymatic degradation of wood components.

197. Pedersen, N.; Gierlinger, N.; Thygesen, L., Bacterial and abiotic decay in waterlogged archaeological Picea abies (L.) Karst studied by confocal Raman imaging and ATR-FTIR spectroscopy. *Holzforschung* **2015,** *69* (1), 103–112.

198. Håfors, B. Conservation of the wood of the Swedish warship Vasa of AD 1628. Evaluation of polyethylene glycol conservation programmes. University of Gothenburg, Gothenburg, Sweden, 2010.
199. Mortensen, M. N.; Egsgaard, H.; Hvilsted, S.; Shashoua, Y.; Glastrup, J., Characterisation of the polyethylene glycol impregnation of the Swedish warship Vasa and one of the Danish Skuldelev Viking ships. *J. Archaeol. Sci.* **2007,** *34* (8), 1211-1218.
200. Hoffman, P., *Conservation of archaelogical hisps and boats personal experiences*. Archetype: London, 2013.
201. Bjurhager, I.; Ljungdahl, J.; Wallström, L.; Gamstedt, E. K.; Berglund, L. A., Towards improved understanding of PEG-impregnated waterlogged archaeological wood: A model study on recent oak. In *Holzforschung*, 2010; Vol. 64, p 243.
202. Hoffmann, P.; Jones, M. A., Structure and degradation process for waterlogged archaeological wood. In *Archaeological Wood*, American Chemical Society: 1989; Vol. 225, pp 35-65.
203. Jeremic, D.; Cooper, P.; Heyd, D., PEG bulking of wood cell walls as affected by moisture content and nature of solvent. *Wood Sci. Technol.* **2006,** *41* (7), 597.
204. Gierlinger, N.; Hansmann, C.; Röder, T.; Sixta, H.; Gindl, W.; Wimmer, R., Comparison of UV and confocal Raman microscopy to measure the melamine–formaldehyde resin content within cell walls of impregnated spruce wood. *Holzforschung* **2005,** *59* (2), 210-213.
205. Christensen, M.; Frosch, M.; Jensen, P.; Schnell, U.; Shashoua, Y.; Nielsen, O. F., Waterlogged archaeological wood—chemical changes by conservation and degradation. *J. Raman Spectrosc.* **2006,** *37* (10), 1171-1178.
206. Rowell, R.; Petersson, R.; Tshabalala, M. A., Cell wall chemistry. In *Handbook of wood chemistry and wood composites*, Rowell, R., Ed. Boca Raton: 2013.
207. Sjöström, E.; Alén, R., *Analytical methods in wood chemistry, pulping, and papermaking*. Springer-Verlag: Berlin; New York, 1999.
208. Gierlinger, N.; Schwanninger, M., The potential of Raman microscopy and Raman imaging in plant research. *J. Spectrosc.* **2007,** *21* (2), 69-89.
209. Larkin, P., *Infrared and Raman spectroscopy : principles and spectral interpretation*. Elsevier: Amsterdam, Netherlands, 2011.
210. Laserna, J. J., *Modern techniques in Raman spectroscopy*. Wiley: Chichester, 1996.
211. Miller, J. N.; Miller, J. C., *Statistics and chemometrics for analytical chemistry*. Pearson Education Limited: Harlow, England, 2005.
212. Lin, S. Y.; Dence, C. W., *Methods in lignin chemistry*. Springer: Berlin, 1992.
213. Grabber, J. H., How do lignin composition, structure, and cross-linking affect degradability? A review of cell wall model studies. *Crop science* **2005,** *v. 45* (no. 3), pp. 820-0-2005 v.45 no.3.
214. Nilsson, T.; Rowell, R., Historical wood – structure and properties. *Journal of Cultural Heritage* **2012,** *13* (3, Supplement), S5-S9.
215. Scheller, H. V.; Ulvskov, P., Hemicelluloses. *Annual Review of Plant Biology* **2010,** *61* (1), 263-289.
216. Björkman, A., Studies on finely divided wood. Part 1. Extraction of lignin with neutral solvents. *Sven. Papperstidn.* **1956,** *59*, 477-485.
217. Holmbom, B., Extractives. In *Analytical methods in wood chemistry, pulping, and papermaking*, Sjöström, E.; Alén, R., Eds. Springer Verlag: Berlin, 1999.
218. Watkins, D.; Nuruddin, M.; Hosur, M.; Tcherbi-Narteh, A.; Jeelani, S., Extraction and characterization of lignin from different biomass resources. *J. Mater. Res. Technol.* **2015,** *4* (1), 26-32.
219. Nisula, L. Wood extractives in conifers: a study of stemwood and knots of industrially important species. Åbo Akademi University, Åbo, Finland, 2018.
220. Agarwal, U. P.; Ralph, S. A., FT-Raman Spectroscopy of Wood: Identifing Contributions of Lignin and Carbohydrate Polymers in the Spectrum of Black Spruce (*Picea mariana*). *Applied Spectroscopy* **1997,** *51* (11), 1648-1655.
221. Aiken, L. S.; West, S. G.; Pitts, S. C.; Baraldi, A. N.; Wurpts, I. C., Multiple linear regression. In *Handbook of Psychology, Second Edition*, Second ed.; Weiner, I.; Schinka, J. A.; Velicer, W. F., Eds. 2012; Vol. 2.
222. Eriksson, L.; Johansson, E.; Wikström, C., Mixture design—design generation, PLS analysis, and model usage. *Chemometr. Intell. Lab.* **1998,** *43* (1), 1-24.

223. Gierlinger, N.; Keplinger, T.; Harrington, M., Imaging of plant cell walls by confocal Raman microscopy. *Nat. Protoc.* **2012,** *7*, 1694.
224. Skaar, C., *Wood-water relations*. Springer-Verlag: Berlin; New York, 1988.
225. Schmidt, U.; Müller, J.; Koenen, J., Confocal Raman imaging of polymeric materials. In *Confocal Raman Miscroscopy*, Toporski, J.; Dieing, T.; Hollricher, O., Eds. Springer Internation Publishing: Cham., 2018.
226. Cebeci-maltaş, D.; Alam, M. A.; Wang, P.; Ben-Amotz, D., Photobleaching profile of Raman peaks and fluorescence background. *European Pharmaceutical Review* **2017,** *22*, 18-21.
227. Bond, J. S.; Atalla, R. H., Raman microprobe investigation of the molecular architechture of loblolly pine tracheids. In *10th International Symposium on Wood and Pulping Chemistry, Main Symposium*, Kami Parupu, G. K., Ed. GA : TAPPI Press: Yokohama, Japan, 1999; pp 96-101.
228. Waggoner, D. C.; Wozniak, A. S.; Cory, R. M.; Hatcher, P. G., The role of reactive oxygen species in the degradation of lignin derived dissolved organic matter. *Geochim Cosmochim Ac* **2017,** *208*, 171-184.
229. Worley, B.; Powers, R., PCA as a practical indicator of OPLS-DA model reliability. *Curr. Metabolomics* **2016,** *4* (2), 97-103.
230. Koenig, J. L.; Angood, A. C., Raman spectra of poly(ethylene glycols) in solution. *J. Polymer Sci. 2 Polymer Phys.* **1970,** *8* (10), 1787-1796.
231. Gierlinger, N.; Luss, S.; König, C.; Konnerth, J.; Eder, M.; Fratzl, P., Cellulose microfibril orientation of Picea abies and its variability at the micron-level determined by Raman imaging. *Journal of experimental botany* **2010,** *61* (2), 587-595.
232. Smith, E.; Dent, G., *Modern Raman spectroscopy : a practical approach*. John Wiley and Sons Ltd.: Southern Gate, Chichester, 2005.
233. Zeise, I.; Heiner, Z.; Holz, S.; Joester, M.; Büttner, C.; Kneipp, J., Raman imaging of plant cell walls in sections of *Cucumis sativus*. *Plants (Basel)* **2018,** *7* (1), 7.
234. Sun, L.; Varanasi, P.; Yang, F.; Loqué, D.; Simmons, B. A.; Singh, S., Rapid determination of syringyl: Guaiacyl ratios using FT-Raman spectroscopy. *Biotechnology and Bioengineering* **2012,** *109* (3), 647-656.
235. Hopke, P. K., Chemometrics applied to environmental systems. *Chemometrics and intelligent laboratory systems* **2015,** *149*, 205-214.
236. Tauler, R., Interpretation of environemtal dta using chemometrics. In *Sample Handling and Trace Analysis of Pollutants; Techniques, Applications and Quality Assurance*, Barceló, D., Ed. Elsevier Science B. V.: 1999.
237. Andreae, M. O.; Atlas, E.; Harris, G. W.; Helas, G.; de Kock, A.; Koppmann, R.; Maenhaut, W.; Manø, S.; Pollock, W. H.; Rudolph, J.; Scharffe, D.; Schebeske, G.; Welling, M., Methyl halide emissions from savanna fires in southern Africa. *Journal of Geophysical Research: Atmospheres* **1996,** *101* (D19), 23603-23613.
238. Quack, B.; Wallace, D. W. R., Air-sea flux of bromoform: Controls, rates, and implications. *Global Biogeochem Cy* **2003,** *17* (1).
239. Keng, F. S. L.; Phang, S. M.; Rahman, N. A.; Leedham Elvidge, E. C.; Malin, G.; Sturges, W. T., The emission of volatile halocarbons by seaweeds and their response towards environmental changes. *Journal of Applied Phycology* **2020,** *32*, 1377-1394.
240. Granfors, A.; Andersson, M.; Chierici, M.; Fransson, A.; Gårdfeldt, K.; Torstensson, A.; Wulff, A.; Abrahamsson, K., Biogenic halocarbons in young Arctic sea ice and frost flowers. *Marine Chemistry* **2013,** *155*, 124-134.
241. Boyd, P. W.; Collins, S.; Dupont, S.; Fabricius, K.; Gattuso, J. P.; Havenhand, J.; Hutchins, D. A.; Riebesell, U.; Rintoul, M. S.; Vichi, M.; Biswas, H.; Ciotti, A.; Gao, K.; Gehlen, M.; Hurd, C. L.; Kurihara, H.; McGraw, C. M.; Navarro, J. M.; Nilsson, G. E.; Passow, U.; Pörtner, H. O., Experimental strategies to assess the biological ramifications of multiple drivers of global ocean change-A review. *Global change biology* **2018,** *24* (6), 2239-2261.
242. Goulitquer, S.; Ritter, A.; Thomas, F.; Ferec, C.; Salaun, J. P.; Potin, P., Release of volatile aldehydes by brown algal kelp Laminaria digitata in response to both biotic and abiotic stress. *ChemBioChem* **2009,** *10*, 977-982.
243. Zou, D.; Gao, K., Temperature response of photosynthetic light- and carbon-use characteristics in the red seaweed *Gracilariopsis lemaneiformis* (Gracilariales, Rhodophyta). *J. Phycol.* **2014,** *50* (2), 366-375.

244. Montzka, S. A.; Reimann, S.; Engel, A.; Kruger, K.; O'Doherty, S. J.; Sturges, W. T., Ozone-depleting substances (ODSs) and related chemicals. In *Scientific assessment of ozone depletion: 2010, global ozone research and monitoring project-report No. 52*, World Meteorological Organization: Geneva, Switzerland, 2011; pp 1-112.
245. Roy, R.; Pratihary, A. K.; Narvenkar, G.; Mochemadkar, S.; Gauns, M. U. In *A study of volatile halocarbons in relation to phytoplankton pigments during a Trichodesmium bloom in the coastal eastern Arabian Sea*, 2011.
246. Yamamoto, H.; Yokouchi, Y.; Otsuki, A.; Itoh, H., Depth profiles of volatile halogenated hydrocarbons in seawater in the Bay of Bengal. *Chemosphere* **2001,** *45* (3), 371-377.
247. Butler, J. H.; King, D. B.; Lobert, J. M.; Montzka, S. A.; Yvon-Lewis, S. A.; Hall, B. D.; Warwick, N. J.; Mondeel, D. J.; Aydin, M.; Elkins, J. W., Oceanic distributions and emissions of short-lived halocarbons. *Global Biogeochem Cy* **2007,** *21* (1).
248. Stemmler, I.; Hense, I.; Quack, B., Marine sources of bromoform in the global open ocean - Global patterns and emissions. *Biogeosciences* **2014,** *12*, 1967-1981.
249. *Scientific assessment of ozone depletion: 2018, global ozone research and monitoring project No. 58*; World Meteorological Organization: 2018.
250. Saiz-Lopez, A.; Plane, J. M. C.; Baker, A. R.; Carpenter, L. J.; von Glasow, R.; Gómez Martín, J. C.; McFiggans, G.; Saunders, R. W., Atmospheric chemistry of iodine. *Chem Rev* **2012,** *112* (3), 1773-1804.
251. Hossaini, R.; Chipperfield, M. P.; Montzka, S. A.; Rap, A.; Dhomse, S.; Feng, W., Efficiency of short-lived halogens at influencing climate through depletion of stratospheric ozone. *Nature Geoscience* **2015,** *8* (3), 186-190.
252. Simpson, W. R.; von Glasow, R.; Riedel, K.; Anderson, P.; Ariya, P.; Bottenheim, J.; Burrows, J.; Carpenter, L. J.; Friess, U.; Goodsite, M. E.; Heard, D.; Hutterli, M.; Jacobi, H. W.; Kaleschke, L.; Neff, B.; Plane, J.; Platt, U.; Richter, A.; Roscoe, H.; Sander, R.; Shepson, P.; Sodeau, J.; Steffen, A.; Wagner, T.; Wolff, E., Halogens and their role in polar boundary-layer ozone depletion. *Atmospheric Chemistry and Physics* **2007,** *7* (16), 4375-4418.
253. von Glasow, R.; von Kuhlmann, R.; Lawrence, M. G.; Platt, U.; Crutzen, P. J., Impact of reactive bromine chemistry in the troposphere. *Atmos. Chem. Phys.* **2004,** *4* (11/12), 2481-2497.
254. Saiz-Lopez, A.; von Glasow, R., Reactive halogen chemistry in the troposphere. *Chemical Society Reviews* **2012,** *41* (19), 6448-6472.
255. Read, K. A.; Mahajan, A. S.; Carpenter, L. J.; Evans, M. J.; Faria, B. V. E.; Heard, D. E.; Hopkins, J. R.; Lee, J. D.; Moller, S. J.; Lewis, A. C.; Mendes, L.; McQuaid, J. B.; Oetjen, H.; Saiz-Lopez, A.; Pilling, M. l. J.; Plane, J. M. C., Extensive halogen-mediated ozone destruction over the tropical Atlantic Ocean. *Nature* **2008,** *453* (7199), 1232-1235.
256. Jacob, D. J., *Introduction to atmospheric chemistry*. Princeton University Press: 1999.
257. Orlikowska, A.; Stolle, C.; Pollehne, F.; Jürgens, K.; Schulz-Bull, D. E., Dynamics of halocarbons in coastal surface waters during short term mesocosm experiments. *Environmental Chemistry* **2015,** *12* (4), 515-525.
258. Küpper, F. C.; Carpenter, L. J.; McFiggans, G. B.; Palmer, C. J.; Waite, T. J.; Boneberg, E.-M.; Woitsch, S.; Weiller, M.; Abela; Grolimund, D.; Potin, P.; Butler, A.; Luther, G. W.; Kroneck, P. M. H.; Meyer-Klaucke, W.; Feiters, M. C., Iodide accumulation provides kelp with an inorganic antioxidant impacting atmospheric chemistry. *Proceedings of the National Academy of Sciences* **2008,** *105* (19), 6954-6958.
259. Brownell, D. K.; Moore, R. M.; Cullen, J. J., Production of methyl halides by *Prochlorococcus* and *Synechococcus*. *Global Biogeochem Cy* **2010,** *24* (2).
260. Johnson, T. L.; Palenik, B.; Brahamsha, B., Characterization of a functional vanadium-dependent bromoperoxidase in the marine cyanobacterium *Synechococcus* Sp. Cc9311(1). *J Phycol* **2011,** *47* (4), 792-801.
261. Karlsson, A.; Auer, N.; Schulz-Bull, D.; Abrahamsson, K., Cyanobacterial blooms in the Baltic — A source of halocarbons. *Marine Chemistry* **2008,** *110* (3), 129-139.
262. Scarratt, M. G.; Moore, R. M., Production of methyl chloride and methyl bromide in laboratory cultures of marine phytoplankton. *Marine Chemistry* **1996,** *54* (3), 263-272.
263. Scarratt, M. G.; Moore, R. M., Production of methyl bromide and methyl chloride in laboratory cultures of marine phytoplankton II. *Marine Chemistry* **1998,** *59* (3), 311-320.
264. Küpper, F. C.; Schweigert, N.; Ar Gall, E.; Legendre, J. M.; Vilter, H.; Kloareg, B., Iodine uptake in Laminariales involves extracellular, haloperoxidase-mediated oxidation of iodide. *Planta* **1998,** *207* (2), 163-171.

265. Gschwend, P. M.; Macfarlane, J. K.; Newman, K. A., Volatile halogenated organic compounds released to seawater from temperate marine macroalgae. *Science* **1985,** *227*, 1033 - 1035.

266. Xu, G.; Wang, B.-G., Independent evolution of six families of halogenating enzymes. *PloS one* **2016,** *11* (5), e0154619-e0154619.

267. Blasiak, L. C.; Drennan, C. L., Structural perspective on enzymatic halogenation. *Acc Chem Res* **2009,** *42* (1), 147-155.

268. Zuo, Z., Why algae release volatile organic compounds—The emission and roles. *Frontiers in Microbiology* **2019,** *10* (491).

269. Crockford, S. J., Evolutionary roots of iodine and thyroid hormones in cell–cell signaling. *Integrative and Comparative Biology* **2009,** *49* (2), 155-166.

270. Morrison, M.; Schonbaum, G. R.; Snell, E. E., Peroxidise-catalysed halogenations. In *Annual Review of Biochemistry*, Annual Reviews Inc.: Palo Alto, USA, 1976.

271. Kuwabara, J. S.; North, W. J., Culturing microscopic stages of *Macrocytis pyrifera* (Phaeophyta) in aquil, a chemically defined medium. . *J. Phycol.* **1980,** *16* (4), 546-549.

272. George, T. F. W.; Ajcharaporn, U. P.; William, M. D., The transformation of iodate to iodide in marine phytoplankton cultures. *Marine Ecology Progress Series* **2002,** *237*, 27-39.

273. de laCuesta, J. L.; Manley, S. L., Iodine assimilation by marine diatoms and other phytoplankton in nitrate-replete conditions. *Limnol Oceanogr* **2009,** *54* (5), 1653-1664.

274. Almeida, M.; Filipe, S.; Humanes, M.; Maia, M. F.; Melo, R.; Severino, N.; da Silva, J. A.; Frausto da Silva, J. J.; Wever, R., Vanadium haloperoxidases from brown algae of the Laminariaceae family. *Phytochemistry* **2001,** *57* (5), 633-42.

275. Manley, S. L., Phytogenesis of halomethanes: A product ofselection or a metabolic accident? *Biogeochemistry* **2002,** *60* (2), 163-180.

276. Toda, H.; Itoh, N., Isolation and characterization of a gene encoding a S-adenosyl-l-methionine-dependent halide/thiol methyltransferase (HTMT) from the marine diatom Phaeodactylum tricornutum: Biogenic mechanism of CH(3)I emissions in oceans. *Phytochemistry* **2011,** *72* (4-5), 337-343.

277. La Barre, S.; Potin, P.; Leblanc, C.; Delage, L., The halogenated metabolism of brown algae (Phaeophyta), its biological importance and its environmental significance. *Marine drugs* **2010,** *8* (4), 988-1010.

278. Ballschmiter, K., Pattern and sources of naturally produced organohalogens in the marine environment: biogenic formation of organohalogens. *Chemosphere* **2003,** *52* (2), 313-324.

279. Punitha, T.; Phang, S. M.; Juan, J. C.; Beardall, J., Environmental Control of Vanadium Haloperoxidases and Halocarbon Emissions in Macroalgae. *Marine biotechnology (New York, N.Y.)* **2018,** *20* (3), 282-303.

280. Neilson, A. H., Biological effects and biosynthesis of brominated metabolites. In *Organic bromine and iodine compounds. The handbook of environemental chemistry*, Neilson, A. H., Ed. Springer: Berling, Heidelberg, 2003; Vol. 3R.

281. Klein, N. J. Marine biogenic halocarbons: potential for heterotrophic bacterial production and seasonality at San Pedro ocean time series. University of Southeren California, Los Angeles, California, 2016.

282. Ohshiro, T.; Nakano, S.; Takahashi, Y.; Suzuki, M.; Izumi, Y., Occurrence of bromoperoxidase in the marine green macro-alga, ulvella lens, and emission of volatile brominated methane by the enzyme. *Phytochemistry* **1999,** *52* (7), 1211-1215.

283. Apel, K.; Hirt, H., Reactive oxygen species: metabolism, oxidative stress, and signal transduction. *Annu Rev Plant Biol* **2004,** *55*, 373-99.

284. Bischof, K.; Rautenberger, R., Seaweed responses to environmental stress: Reactive oxygen and antioxidative strategies. In *Seaweed Biology: Novel Insights into Ecophysiology, Ecology and Utilization*, Wiencke, C.; Bischof, K., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 109-132.

285. Passardi, F.; Bakalovic, N.; Teixeira, F. K.; Margis-Pinheiro, M.; Penel, C.; Dunand, C., Prokaryotic origins of the non-animal peroxidase superfamily and organelle-mediated transmission to eukaryotes. *Genomics* **2007,** *89* (5), 567-79.

286. Wever, R.; van der Horst, M. A., The role of vanadium haloperoxidases in the formation of volatile brominated compounds and their impact on the environment. *Dalton T* **2013,** *42* (33), 11778-11786.

287. Laturnus, F., Marine macroalgae in polar regions as natural sources for volatile organohalogens. *Environmental Science and Pollution Research* **2001,** *8* (2), 103-108.
288. Wever, R., Structure and function of vanadium haloperoxidases. In *Vanadium: biochemical and molecular biological approaches*, Michibata, H., Ed. Springer Netherlands: 2012; pp 95-125.
289. Colin, C.; Leblanc, C.; Michel, G.; Wagner, E.; Leize-Wagner, E.; Van Dorsselaer, A.; Potin, P., Vanadium-dependent iodoperoxidases in *Laminaria digitata,* a novel biochemical function diverging from brown algal bromoperoxidases. *J. Biol. Inorg. Chem.* **2005,** *10* (2), 156-166.
290. Tarakhovskaya, E. R.; Bilova, T. E.; Maslov, Y. I., Hydrogen peroxide content and vanadium-dependent haloperoxidase activity in thalli of six species of Fucales (Phaeophyceae). *Phycologia* **2015,** *54* (4), 417-424.
291. Manley, S. L.; Goodwin, K.; North, W. J., Laboratory production of bromoform, methylene bromide, and methyl-iodide by macroalgae and distribution in nearshore southern California waters. *Limnol Oceanogr* **1992,** *37* (8), 1652-1659.
292. Nightingale, P. D.; Malin, G.; Liss, P. S., Production of chloroform and other low molecular-weight halocarbons by some species of macroalgae. *Limnol Oceanogr* **1995,** *40* (4), 680-689.
293. Cota, G. F.; Sturges, W. T., Biogenic bromine production in the Arctic. *Marine Chemistry* **1997,** *56* (3), 181-192.
294. Laturnus, F., Volatile halocarbons released from Arctic macroalgae. *Marine Chemistry* **1996,** *55* (3), 359-366.
295. Laturnus, F., Release of volatile halogenated organic compounds by unialgal cultures of polar macroalgae. *Chemosphere* **1995,** *31* (6), 3387-3395.
296. Almeida, M. G.; Humanes, M.; Melo, R.; Silva, J. A.; da Silva, J. J.; Wever, R., Purification and characterisation of vanadium haloperoxidases from the brown alga Pelvetia canaliculata. *Phytochemistry* **2000,** *54* (1), 5-11.
297. Wever, R.; Plat, H.; de Boer, E., Isolation procedure and some properties of the bromoperoxidase from the seaweed *Ascophyllum nodosum*. *Biochim Biophys Acta* **1985,** *830* (2), 181-186.
298. Ni, X.; Hager, L. P., Expression of Batis maritima methyl chloride transferase in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **1999,** *96* (7), 3611-3615.
299. Fenical, W., Halogenation in rhodophyta - review. *J. Phycol.* **1975,** *11* (3), 245-259.
300. Carter-Franklin, J. N.; Butler, A., Vanadium bromoperoxidase-catalyzed biosynthesis of halogenated marine natural products. *Journal of the American Chemical Society* **2004,** *126* (46), 15060-15066.
301. Tarakhovskaya, E. R., Mechanisms of bioadhesion of macrophytic algae. *Russ. J. Plant Physiol.* **2014,** *61* (1), 19-25.
302. Hill, V. L.; Manley, S. L., Release of reactive bromine and iodine from diatoms and its possible role in halogen transfer in polar and tropical oceans. *Limnol Oceanogr* **2009,** *54* (3), 812-822.
303. Klein, N. J.; Beck, A. J.; Hutchins, D. A.; Sañudo-Wilhelmy, S. A., Regression modeling of the North East Atlantic Spring Bloom suggests previously unrecognized biological roles for V and Mo. *Front Microbiol* **2013,** *4*, 45.
304. Weyand, M.; Hecht, H.; Kiess, M.; Liaud, M.; Vilter, H.; Schomburg, D., X-ray structure determination of a vanadium-dependent haloperoxidase from *Ascophyllum nodosum* at 2.0 A resolution. *J Mol Biol* **1999,** *293* (3), 595-611.
305. Isupov, M. N.; Dalby, A. R.; Brindley, A. A.; Izumi, Y.; Tanabe, T.; Murshudov, G. N.; Littlechild, J. A., Crystal structure of dodecameric vanadium-dependent bromoperoxidase from the red algae *Corallina officinalis*. *J Mol Biol* **2000,** *299* (4), 1035-49.
306. Littlechild, J.; Garcia-Rodriguez, E., Structural studies on the dodecameric vanadium bromoperoxidase from *Corallina* species. *Coordin Chem Rev* **2003,** *237* (1), 65-76.
307. Butler, A., Vanadium haloperoxidases. *Curr Opin Chem Biol* **1998,** (2), 279-85.
308. Messerschmidt, A.; Prade, L.; Wever, R., Implications for the catalytic mechanism of the vanadium-containing enzyme chloroperoxidase from the fungus Curvularia inaequalis by X-ray structures of the native and peroxide form. *Biological chemistry* **1997,** *378* (3-4), 309-15.

309. Renirie, R.; Charnock, J. M.; Garner, C. D.; Wever, R., Vanadium K-edge XAS studies on the native and peroxo-forms of vanadium chloroperoxidase from Curvularia inaequalis. *J Inorg Biochem* **2010,** *104* (6), 657-64.
310. Christmann, U.; Dau, H.; Haumann, M.; Kiss, E.; Liebisch, P.; Rehder, D.; Santoni, G.; Schulzke, C., Substrate binding to vanadate-dependent bromoperoxidase from Ascophyllum nodosum: a vanadium K-edge XAS approach. *Dalton transactions (Cambridge, England : 2003)* **2004,** (16), 2534-40.
311. Wever, R.; Hemrika, W., Vanadium haloperoxidases. In *Handbook of metalloproteins*, Messerschmidt, A.; Hubert, R., Eds. Wiley: Chichester, 2001; pp 1417-1428.
312. Leblanc, C.; Vilter, H.; Fournier, J. B.; Delage, L.; Potin, P.; Rebuffet, E.; Michel, G.; Solari, P. L.; Feiters, M. C.; Czjzek, M., Vanadium haloperoxidases: From the discovery 30 years ago to X-ray crystallographic and V K-edge absorption spectroscopic studies. *Coordin Chem Rev* **2015,** *301*, 134-146.
313. Butler, A., Mechanistic consideration of hte vanadium haloperoxidases. *Coordin Chem Rev* **1999,** *187*, 17-35.
314. Tanaka, N.; Zulfiquar, H.; Wever, R., Kinetic characterization of the active site mutants Ser42Ala nd Phe397His of vanadium chloroperoxidase from fungus *Curvularia inaequalis*. *Inorganica Chimica Acta* **2003,** *356*, 288-296.
315. Dring, M. J., Stress resistance and disease resistance in seaweeds: the role of reactive oxygen metabolism. In *Advances in Botanical Research*, Academic Press: 2005; Vol. 43, pp 175-207.
316. Lemesheva, V.; Birkemeyer, C.; Garbary, D.; Tarakhovskaya, E. R., Vanadium-dependent haloperoxidase activity and phlorotannin incorporation into the cell wall during early embryogenesis of Fucus vesiculosus (Phaeophyceae). *European Journal of Phycology* **2020**, 1-10.
317. Küpper, F. C.; Carpenter, L. J.; Leblanc, C.; Toyama, C.; Uchida, Y.; Maskrey, B. H.; Robinson, J.; Verhaeghe, E. F.; Malin, G.; Luther, G. W., 3rd; Kroneck, P. M.; Kloareg, B.; Meyer-Klaucke, W.; Muramatsu, Y.; Megson, I. L.; Potin, P.; Feiters, M. C., In vivo speciation studies and antioxidant properties of bromine in Laminaria digitata reinforce the significance of iodine accumulation for kelps. *J Exp Bot* **2013,** *64* (10), 2653-64.
318. Cosse, A.; Potin, P.; Leblanc, C., Patterns of gene expression induced by oligoguluronates reveal conserved and environment-specific molecular defense responses in the browl alga *Laminaria digitata*. *New Phytologist* **2009,** *182*, 239-250.
319. Collen, J.; Ekdahl, A.; Abrahamsson, K.; Pedersen, M., The involvement of hydrogen peroxide in the production of volatile halogenated compounds by *Meristiella gelidium*. *Phytochemistry* **1994,** *36*, 1197-1202.
320. Weinberger, F.; Coquempot, B.; Forner, S.; Morin, P.; Kloareg, B.; Potin, P., Different regulation of haloperoxidation during agar oligosaccharide-activated defence mechanisms in two related red algae, *Gracilaria* sp. and *Gracilaria chilensis*. *J Exp Bot* **2007,** *58* (15-16), 4365-72.
321. Dietz, K. J., Thiol-based peroxidases and ascorbate peroxidases: Why plants rely on multiple peroxidase systems in the photosynthesizing chloroplast? *Molecules and cells* **2016,** *39* (1), 20-5.
322. Manley, S. L.; Barbero, P. E., Physiological constraints on bromoform ($CHBr_3$) production by *Ulva lactuca* (Chlorophyta). *Limnol Oceanogr* **2001,** *46* (6), 1392-1399.
323. Goodwin, K. D.; North, W. J.; Lidstrom, M. E., Production of bromoform and dibromomethane by Giant Kelp: Factors affecting release and comparison to anthropogenic bromine sources. *Limnol Oceanogr* **1997,** *42*, 1725-1734.
324. Harley, C. D. G.; Anderson, K. M.; Demes, K. W.; Jorve, J. P.; Kordas, R. L.; Coyle, T. A.; Graham, M. H., Effects of climate change on global seaweed communities. *J. Phycol.* **2012,** *48* (5), 1064-1078.
325. de Bettignies, T.; Wernberg, T.; Gurgel, C. F. D., Exploring the Influence of Temperature on Aspects of the Reproductive Phenology of Temperate Seaweeds. *Frontiers in Marine Science* **2018,** *5* (218).
326. Colvard, N. B.; Carrington, E.; Helmuth, B., Temperature-dependent photosynthesis in the intertidal alga Fucus gardneri and sensitivity to ongoing climate change. *Journal of Experimental Marine Biology and Ecology* **2014,** *458*, 6-12.
327. Laturnus, F.; Giese, B.; Wiencke, C.; Adams, F. C., Low-molecular-weight organoiodine and organobromine compounds released by polar macroalgae--the influence of abiotic factors. *Fresenius' journal of analytical chemistry* **2000,** *368* (2-3), 297-302.

328. Abrahamsson, K.; Choo, K. S.; Pedersen, M.; Johansson, G.; Snoeijs, P., Effects of temperature on the production of hydrogen peroxide and volatile halocarbons by brackish-water algae. *Phytochemistry* **2003,** *64* (3), 725-34.
329. Mithoo-Singh, P. K.; Keng, F. S.; Phang, S. M.; Leedham Elvidge, E. C.; Sturges, W. T.; Malin, G.; Abd Rahman, N., Halocarbon emissions by selected tropical seaweeds: species-specific and compound-specific responses under changing pH. *PeerJ* **2017,** *5*, e2918.
330. Mtolera, M. S. P.; Collén, J.; Pedersén, M.; Ekdahl, A.; Abrahamsson, K.; Semesi, A. K., Stress-induced production of volatile halogenated organic compounds in *Eucheuma denticulatum* (Rhodophyta) caused by elevated pH and high light intensities. *European Journal of Phycology* **1996,** *31* (1), 89-95.
331. Webb, A. L.; Leedham-Elvidge, E.; Hughes, C.; Hopkins, F. E.; Malin, G.; Bach, L. T.; Schulz, K.; Crawfurd, K.; Brussaard, C. P. D.; Stuhr, A.; Riebesell, U.; Liss, P. S., Effect of ocean acidification and elevated fCO2 on trace gas production by a Baltic Sea summer phytoplankton community. *Biogeosciences* **2016,** *13* (15), 4595-4613.
332. Hopkins, F. E.; Turner, S. M.; Nightingale, P. D.; Steinke, M.; Bakker, D.; Liss, P. S., Ocean acidification and marine trace gas emissions. *Proceedings of the National Academy of Sciences* **2010,** *107* (2), 760-765.
333. Abrahamsson, K., Effect of acidification on VHOC production in mesocosms. 2014.
334. Wong, S.-L.; Chang, J., Salinity and light effects on growth, photosynthesis, and respiration of Grateloupia filicina (Rhodophyta). *Aquaculture* **2000,** *182* (3), 387-395.
335. Tropin, I. V.; Radzinskaya, N. V.; Voskoboinikov, G. M., The influence of salinity on the rate of dark respiration and structure of the cells of brown algae thalli from the Barents Sea littoral. *Biology Bulletin of the Russian Academy of Sciences* **2003,** *30* (1), 40-47.
336. Sung, M. S.; Hsu, Y. T.; Hsu, Y. T.; Wu, T. M.; Lee, T. M., Hypersalinity and hydrogen peroxide upregulation of gene expression of antioxidant enzymes in *Ulva fasciata* against oxidative stress. *Marine biotechnology (New York, N.Y.)* **2009,** *11* (2), 199-209.
337. de Boer, E.; Wever, R., The reaction mechanism of the novel vanadium-bromoperoxidase. A steady-state kinetic analysis. *J Biol Chem* **1988,** *263* (25), 12326-32.
338. Soedjak, H. S.; Walker, J. V.; Butler, A., Inhibition and inactivation of vanadium bromoperoxidase by the substrate hydrogen peroxide and further mechanistic studies. *Biochemistry* **1995,** *34* (39), 12689-12696.
339. Verhaeghe, E. F.; Fraysse, A.; Guerquin-Kern, J.-L.; Wu, T.-D.; Devès, G.; Mioskowski, C.; Leblanc, C.; Ortega, R.; Ambroise, Y.; Potin, P., Microchemical imaging of iodine distribution in the brown alga *Laminaria digitata* suggests a new mechanism for its accumulation. *JBIC Journal of Biological Inorganic Chemistry* **2008,** *13* (2), 257-269.
340. Küpper, F. C.; Leblanc, C.; Meyer-Klaucke, W.; Potin, P.; Feiterns, M. C., Different speciation for bromine in brown and red algae, revealed by in vivo X-ray absorption spectroscopic studies. *J Physcol* **2014,** *50*, 652-664.
341. Plat, H.; Krenn, B. E.; Wever, R., The bromoperoxidase from the lichen *Xanthoria parietina* is a novel vanadium enzyme. *Biochemical Journal* **1987,** *248* (1), 277-279.
342. Kemiska arbetsmiljörisker: Arbetsmiljöverkets föreskrifter om ändring i Arbetsmiljöverkets föreskrifter och allmänna råd (AFS 2011:19) om kemiska arbetsmiljörisker. Arbetsmiljöverket, Ed. 2011.
343. Verhaeghe, E.; Buisson, D.; Zekri, E.; Leblanc, C.; Potin, P.; Ambroise, Y., A colorimetric assay for steady-state analyses of iodo- and bromoperoxidase activities. *Anal Biochem* **2008,** *379* (1), 60-5.
344. Machida, K.; Lee, H.; Uno, T., Resonance Raman spectra of sulfophthalein dyes in aqueous solutions. *J Raman Spectrosc* **1979,** *8* (3), 172-176.
345. Gribble, G. W., The diversity of naturally produced organohalogens. *Chemosphere* **2003,** *52*, 289-297.
346. Bidleman, T. F.; Andersson, A.; Jantunen, L. M.; Kucklick, J. R.; Kylin, H.; Letcher, R. J.; Tysklind, M.; Wong, F., A review of halogenated natural products in Arctic, Subarctic and Nordic ecosystems. *Emerging Contaminants* **2019,** *5*, 89-115.
347. van Pee, K. H., Biosynthesis of halogenated metabolites in bacteria. *Annual Review of Microbiology* **1996,** *50*, 375-399.
348. Goodwin, K. D.; Schaefer, J. K.; Oremland, R. S., Bacterial oxidation of dibromomethane and methyl bromide in natural waters and enrichment cultures. *Appl Environ Microbiol* **1998,** *64* (12), 4629-4636.

349. McAnulla, C.; McDonald, I. R.; Murrell, J. C., Methyl chloride utilising bacteria are ubiquitous in the natural environment. *FEMS Microbiol Lett* **2001,** *201* (2), 151-5.
350. Schäfer, H.; Miller, L. G.; Oremland, R. S.; Murrell, J. C., Bacterial cycling of methyl halides. *Advances in applied microbiology* **2007,** *61*, 307-46.
351. Fujimori, T.; Yoneyama, Y.; Taniai, G.; Kurihara, M.; Tamegai, H.; Hashimoto, S., Methyl halide production by cultures of marine proteobacteria *Erythrobacter* and *Pseudomonas* and isolated bacteria from brackish water. *Limnol Oceanogr* **2012,** *57* (1), 154-162.
352. Fuse, H.; Inoue, H.; Murakami, K.; Takimura, O.; Yamaoka, Y., Production of free and organic iodine by *Roseovarius* spp. *FEMS Microbiology Letters* **2003,** *229* (2), 189-194.
353. Amachi, S.; Kamagata, Y.; Kanagawa, T.; Muramatsu, Y., Bacteria mediate methylation of iodine in marine and terrestrial environments. *Appl Environ Microbiol* **2001,** *67* (6), 2718-22.
354. Liu, Y.; Yvon-Lewis, S. A.; Thornton, D. C. O.; Butler, J. H.; Bianchi, T. S.; Campbell, L.; Hu, L.; Smith, R. W., Spatial and temporal distributions of bromoform and dibromomethane in the Atlantic Ocean and their relationship with photosynthetic biomass. *Journal of Geophysical Research: Oceans* **2013,** *118* (8), 3950-3965.
355. Tanaka, N.; Dumay, V.; Liao, Q.; Lange, A. J.; Wever, R., Bromoperoxidase activity of vanadate-substituted acid phosphatases from *Shigella flexneri* and *Salmonella enterica* ser. typhimurium. *European journal of biochemistry* **2002,** *269* (8), 2162-7.
356. Fournier, J.-B.; Rebuffet, E.; Delage, L.; Grijol, R.; Meslet-Cladière, L.; Rzonca, J.; Potin, P.; Michel, G.; Czjzek, M.; Leblanc, C., The vanadium iodoperoxidase from the marine flavobacteriaceae species *Zobellia galactanivorans* reveals novel molecular and evolutionary features of halide specificity in the vanadium haloperoxidase enzyme family. *Appl Environ Microbiol* **2014,** *80* (24), 7561-7573.
357. Lin, C. Y.; Manley, S. L., Bromoform production from seawater treated with bromoperoxidase. *Limnol Oceanogr* **2012,** *57* (6), 1857-1866.
358. Theiler, R.; Cook, J. C.; Hager, L. P.; Siuda, J. F., Halohydrocarbon synthesis by bromoperoxidase. *Science* **1978,** *202* (4372), 1094-1096.
359. Itoh, N.; Shinya, M., Seasonal evolution of bromomethanes from coralline algae (*Corallinaceae*) and its effect on atmospheric ozone. *Marine Chemistry* **1994,** *45* (1-2), 95-103.
360. Urhahn, T.; Ballschmiter, K., Chemistry of the biosynthesis of halogenated methanes: C1-organohalogens as pre-industrial chemical stressors in the environment? *Chemosphere* **1998,** *37* (6), 1017-1032.
361. Happell, J. D.; Wallace, D., Methyl iodide in the Greenland/Norwegian Seas and the tropical Atlantic Ocean: Evidence for photochemical production. *Geophysical Research Letters* **1996,** *23*, 2105-2108.
362. Moore, R. M., A photochemical source of methyl chloride in saline waters. *Environ Sci Technol* **2008,** *42* (6), 1933-1937.
363. Itoh, N.; Tsujita, M.; Ando, T.; Hisatomi, G.; Higashi, T., Formation and emission of monohalomethanes from marine algae. *Phytochemistry* **1997,** *45* (1), 67-73.
364. Elliott, S.; Rowland, F. S., Nucleophilic substitution rates and solubilities for methyl halides in seawater. *Geophysical Research Letters* **1993,** *20* (11), 1043-1046.
365. Geen, C. E. Selected marine sources and sinks of bromoform and other low molecular weight organobromines. Dalhousie University, 1992.
366. Vogel, T. M.; Criddle, C. S.; McCarty, P. L., ES Critical Reviews: Transformations of halogenated aliphatic compounds. *Environ Sci Technol* **1987,** *21* (8), 722-736.
367. Neidleman, S. L.; Geigert, J., *Biohalogenation: principles, basic roles and applications*. 1986.
368. Beissner, R. S.; Guilford, W. J.; Coates, R. M.; Hager, L. P., Synthesis of brominated heptanones and bromoform by a bromoperoxidase of marine origin. *Biochemistry* **1981,** *20* (13), 3724-31.
369. Hine, J., Carbon dichloride as an intermediate in the basic hydrolysis of chloroform. A mechanism for substitution reactions at a saturated carbon atom1. *Journal of the American Chemical Society* **1950,** *72* (6), 2438-2445.
370. Robinson, E. A., 324. The reaction of dichloromethylene with water and with phenoxide ions (Reimer–Tiemann reaction). *Journal of the Chemical Society (Resumed)* **1961,** (0), 1663-1671.
371. Jones, C. E.; Carpenter, L. J., Chemical destruction of CH3I, C2H5I, 1-C3H7I, and 2-C3H7I in saltwater. *Geophysical Research Letters* **2007,** *34* (13).

372. Carpenter, L. J.; Liss, P. S., On temperate sources of bromoform and other reactive organic bromine gases. *Journal of Geophysical Research: Atmospheres* **2000,** *105* (D16), 20539-20547.

373. Kwok, W. M.; Zhao, C.; Li, Y. L.; Guan, X.; Phillips, D. L., Direct observation of an isopolyhalomethane O-H insertion reaction with water: picosecond time-resolved resonance Raman (ps-TR3) study of the isobromoform reaction with water to produce a CHBr2OH product. *J Chem Phys* **2004,** *120* (7), 3323-32.

374. Li, Y.-L.; Zhao, C.; Kwok, W. M.; Guan, X.; Zuo, P.; Phillips, D. L., Observation of a HI leaving group following ultraviolet photolysis of CH2I2 in water and an ab initio investigation of the O–H insertion/HI elimination reactions of the CH2I–I isopolyhalomethane species with H2O and 2H2O. *The Journal of Chemical Physics* **2003,** *119* (9), 4671-4681.

375. Agarwal, V.; Miles, Z. D.; Winter, J. M.; Eustáquio, A. S.; El Gamal, A. A.; Moore, B. S., Enzymatic Halogenation and Dehalogenation Reactions: Pervasive and Mechanistically Diverse. *Chem Rev* **2017,** *117* (8), 5619-5674.

376. Goodwin, K. D.; Lidstrom, M. E.; Oremland, R. S., Marine bacterial degradation of brominated methanes. *Environ Sci Technol* **1997,** *31* (11), 3188-3192.

377. Ichikawa, K.; Kurihara, M.; Tamegai, H.; Hashimoto, S., Decomposition of brominated organic halogens by cultures of marine proteobacteria: Phaeobacter, Roseobacter, and Rhodobacter. *Marine Chemistry* **2015,** *176*, 133-141.

378. Karlsson, A.; Theorin, M.; Snoeijs, P.; Abrahamsson, K., Volatile halogenated organic compounds in pack ice and snow in the central Arctic Ocean.

379. Allard, A. S.; Neilson, A. H., Degradation and transformation of organic bromine and iodine compounds: comparison with their chlorinated analogues. In *Organic bromine and iodine compounds: the handbook of environmental chemistry*, Neilson, A. H., Ed. Springer: Berlin, 2003; Vol. 3R.

380. Trotsenko, Y. A.; Doronina, N. V., The biology of methylobacteria capable of degrading halomethanes. *Microbiology* **2003,** *72* (2), 121-131.

381. Zhao, S.; Rogers, M. J.; He, J., Microbial reductive dehalogenation of trihalomethanes by a Dehalobacter-containing co-culture. *Applied Microbiology and Biotechnology* **2017,** *101* (13), 5481-5492.

382. van Pée, K. H.; Unversucht, S., Biological dehalogenation and halogenation reactions. *Chemosphere* **2003,** *52* (2), 299-312.

383. Hancock, T. L.; Costello, A. M.; Lidstrom, M. E.; Oremland, R. S., Strain IMB-1, a novel bacterium for the removal of methyl bromide in fumigated agricultural soils. *Appl Environ Microbiol* **1998,** *64* (8), 2899-905.

384. Hines, M. E.; Crill, P. M.; Varner, R. K.; Talbot, R. W.; Shorter, J. H.; Kolb, C. E.; Harriss, R. C., Rapid consumption of low concentrations of methyl bromide by soil bacteria. *Appl Environ Microbiol* **1998,** *64* (5), 1864-1870.

385. Liss, P., Take the shuttle--from marine algae to atmospheric chemistry. *Science* **1999,** *285* (5431), 1217-1218.

386. Ritschard, R. L., Marine algae as a CO_2 sink. *Water, Air, and Soil Pollution* **1992,** *64* (1), 289-303.

387. Abrahamsson, K.; Ekdahl, A.; Collén, J.; Pedersén, M., Formation and distribution of halogenated volatile organics in sea water. In *Naturally-Produced Organohalogens*, Grimvall, A.; de Leer, E. W. B., Eds. Springer Netherlands: Dordrecht, 1995; pp 317-326.

388. Klick, S.; Abrahamsson, K., Biogenic volatile iodated hydrocarbons in the ocean. *Journal of Geophysical Research: Oceans* **1992,** *97* (C8), 12683-12687.

389. Lim, Y. K.; Phang, S. M.; Abdul Rahman, N.; Sturges, W. T.; Malin, G., Halocarbon emissions from marine phytoplankton and climate change. *International Journal of Environmental Science and Technology* **2017,** *14* (6), 1355-1370.

390. Montzka, S.; Reimann, S., Ozone depleting substance (ODSs) and related chemicals. In *Scientific Assessment of Ozone Depletion: 2010, Global Research and Monitoring Project*, WMO (World Meteorological Organization): 2011.

391. de Boer, E.; van Kooyk, Y.; Tromp, M. G. M.; Plat, R.; Wever, R., Bromoperoxidase from *Ascophyllum nodosum*: a novel class of enzymes containing vanadium as a prostheric group? *Biophys. Actra*

1986, *869*, 48-53.
392. Butler, A.; Walker, J. V., Marine Haloperoxidases. *Chem Rev* **1993,** *93* (5), 1937-1944.
393. Vilter, H., Vanadium-dependent haloperoxidases. In *Metal ions in biological systems*, Sigel, H.; Sigel, A., Eds. Marcel Dekker: New York, 1995; Vol. 31, pp 341-344.
394. Almeida, M.; Humanes, M.; Melo, R.; Silva, J. A.; Frausto da Silva, J. J.; Wever, R., Purification and characterisation of vanadium haloperoxidases from the brown alga *Pelveria canaliculata*. *Phytochemistry* **2000,** *54*, 5-11.
395. Carter-Franklin, J. N.; Butler, A., Vanadium bromoperoxidase-catalyzed biosynthesis of halogenated marine natural products. *J Am Chem Soc* **2004,** *126* (46), 15060-6.
396. Verdel, E. F.; Kline, P. C.; Wani, S.; Woods, A. E., Purification and partial characterization of haloperoxidase from fresh water algae *Cladophora glomerata*. *Comp Biochem Physiol B Biochem Mol Biol* **2000,** *125* (2), 179-87.
397. Paul, C.; Pohnert, G., Production and role of volatile halogenated compounds from marine algae. *Natural Product Reports* **2011,** *28* (2), 186-195.
398. Brindley, A. A.; Dalby, A. R.; Isupov, M. N.; Littlechild, J. A., Preliminary X-ray analysis of a new crystal form of the vanadium-dependent bromoperoxidase from *Corallina officinalis*. *Acta Crystallogr D Biol Crystallogr* **1998,** *54* (Pt 3), 454-7.
399. Rush, C.; Willetts, A.; Davies, G.; Dauter, Z.; Watson, H.; Littlechild, J., Purification, crystallisation and preliminary X-ray analysis of the vanadium-dependent haloperoxidase from *Corallina officinalis*. *FEBS Lett* **1995,** *359* (2-3), 244-6.
400. Yu, H.; Whittaker, J. W., Vanadate activation of bromoperoxidase from *Corallina officinalis*. *Biochem Biophys Res Commun* **1989,** *160* (1), 87-92.
401. Xu, L. J.; Zong, C.; Zheng, X. S.; Hu, P.; Feng, J. M.; Ren, B., Label-free detection of native proteins by surface-enhanced Raman spectroscopy using iodide-modified nanoparticles. *Analytical Chemistry* **2014,** *86* (4), 2238-2245.
402. Siddhanta, S.; Narayana, C., Surface Enhanced Raman Spectroscopy of Proteins: Implications for Drug Designing Invited Review Article. *Nanomater Nanotechno* **2012,** *2*, 1-17.
403. Singhal, K.; Kalkan, A. K., Surface-enhanced Raman scattering captures conformational changes of single photoactive yellow protein molecules under photoexcitation. *J Am Chem Soc* **2010,** *132* (2), 429-31.
404. Smulko, J.; Wróbel, M. S.; Barman, I. In *Noise in biological Raman spectroscopy*, 2015 International Conference on Noise and Fluctuations (ICNF), 2-6 June 2015; 2015; pp 1-6.
405. Moskovits, M., Surface-enhanced Spectroscopy. *Rev Mod Phys* **1985,** *57* (3), 783-826.
406. Han, X. X.; Zhao, B.; Ozaki, Y., Surface-enhanced Raman scattering for protein detection. *Anal Bioanal Chem* **2009,** *394*, 1719-1727.
407. Chumanov, G. D.; Efremov, R. G.; Nabiev, I. R., Surface-enhanced Raman spectroscopy of biomolecules. *J Raman Spectrosc* **1990,** *21*, 43-48.
408. Broderick, J. B.; Natan, M. J.; O'Halloran, T. V.; van Duyne, R. P., Evidence for retension of biological activity of a non-heme iron enzyme adsorbed on a silver colloid: a surface-enhanced resonance Raman scattering study. *Biochemistry* **1993,** *32*, 13771-131776.
409. Xiu-Mei, L.; Yan, C.; Yan-Hui, X.; Bin, R.; Zhong-Qun, T., Surface-enhanced Raman spectroscopy: substrate-related issues. *Anal Bioanal Chem* **2009,** *394*, 1729-1745.
410. Ando, J.; Fujita, K.; Smith, N. I.; Kawata, S., Dynamic SERS imaging of cellular transport pathways with endocytosed gold nanoparticles. *Nano Lett* **2011,** *11* (12), 5344-8.
411. Huang, K. C.; Bando, K.; Ando, J.; Smith, N. I.; Fujita, K.; Kawata, S., 3D SERS (surface enhanced Raman scattering) imaging of intracellular pathways. *Methods* **2014,** *68* (2), 348-53.
412. Vitol, E. A.; Orynbayeva, Z.; Friedman, G.; Gogotsi, Y., Nanoprobes for intracellular and single cell surface-enhanced Raman spectroscopy (SERS). *J Raman Spectrosc* **2012,** *43* (7), 817-827.
413. N., M. Y.; Torres, E.; Blazquez, M. L.; Ballester, A.; Gonzalez, F.; Munoz, J. A., Gold(III) biosorption and bioreduction with the brown alga Fucus vesiculosus. *Journal of Hazardous Materials* **2009,** *166*, 612-618.
414. R., H. L.; Vikesland, P. J., Surface-enhanced Raman spectroscopy (SERS) cellular imaging of intracellularly biozynthesized gold nanoparticles. *ACS Sustainable Chem. Eng.* **2014,** *2*, 1599-1608

415. Slater, J. B.; Tadesco, J. M.; Fairchild, R. C.; Lewis, I. R., Raman spectrometry and its adaptation to the industrial environment. In *Handbook of Raman spectroscopy: From the research laboratory to the process line*, Lewis, I. R.; Edwards, G. M., Eds. Marcel Dekker, Inc.: New York, Basel, 2001.
416. Box, G. E. P.; Hunter, J. S.; Gordon, W., *Statistics for experimenters: design, innovation, and discovery*. 2nd ed.; Wiley-Interscience: Hoboken, N.J., 2005.
417. Golub, G. H.; Heath, M.; Wahba, G., Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* **1979,** *21* (2), 215-223.
418. Emmons, E. D.; Fallas, J. C.; Kamisetty, V. K.; Chien, W. M.; Covington, A. M.; Chellappa, R. S.; Gramsch, S. A.; Hemley, R. J.; Chandra, D., High-pressure Raman spectroscopy of tris(hydroxymethyl)aminomethane. *J Phys Chem B* **2010,** *114* (17), 5649-56.
419. Khoma, R. E.; Gel'mbol'dt, V. O.; Shishkin, O. V.; Baumer, V. N.; Ennan, A. A., Synthesis, crystal structure, vibrational spectra, and thermochemical transformations of tris(hydroxymethyl)aminomethane. *Russ J Inorg Chem+* **2014,** *59* (1), 1-6.
420. Brandt, N. N.; Chikishev, A. Y.; Sakodinskaya, I. K., Raman spectroscopy of tris-(hydroxymethyl)aminomethane as a model system for the studies of alpha-chymotrypsin activation by crown ether in organic solvents. *J Mol Struct* **2003,** *648* (3), 177-182.
421. Schroetter, S.; Bougeard, D., The calculated and observed vibrational-spectra of the ordered phase of Tris(hydroxymethyl)aminomethane. *Ber Bunsen Phys Chem* **1987,** *91* (11), 1217-1221.
422. Lund Myhre, C. E.; Christensen, D. H.; Nicolaisen, F. M.; Nielsen, C. J., Spectroscopic study of aqueous H_2SO_4 at different temperatures and compositions: variations in dissociation and optical properties. *J. Phys. Chem. A.* **2003,** *107*, 1979-1991.
423. Myers, G. A.; Wiseman, B.; Cookman, A.; Morgan, Q.; Turner, L. G.; Pearce, J. *Raman spectroscopy - setecting SO_x and NO_x in the precipice sandstone*; 2016.
424. Talari, A. C. S.; Movasaghi, Z.; Rehman, S.; Rehman, I. U., Raman spectroscopy of biological tissues. *Appl Spectrosc Rev* **2014,** *50* (1), 46-111.
425. Ashton, L.; Pudney, P. D. A.; Blanch, E. W.; Yakubov, G. E., Understanding glycoprotein behaviours using Raman and Raman optical activity spectroscopies: Characterising the entanglement induced conformational changes in oligosaccharide chains of mucin. *Adv Colloid Interfac* **2013,** *199*, 66-77.
426. Maiti, N. C.; Apetri, M. M.; Zagorski, M. G.; Carey, P. R.; Anderson, V. E., Raman spectroscopic characterization of secondary structure in natively unfolded proteins: alpha-synuclein. *J Am Chem Soc* **2004,** *126* (8), 2399-408.
427. Rygula, A.; Majzner, K.; Marzec, K. M.; Kaczor, A.; Pilarczyk, M.; Baranska, M., Raman spectroscopy of proteins: a review. *J Raman Spectrosc* **2013,** *44* (8), 1061-1076.
428. Surewicz, W. K.; Mantsch, H. H.; Chapman, D., Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment. *Perspectives in Biochemistry* **1993,** *32* (2), 389–394.
429. Pfender, F.; Ziegler, G. M., Kissing numbers, sphere packings, and some unexpected proofs. *Notices of AMS* **2004,** *51* (8), 873-883.
430. Taha, M.; Lee, M. J., Interactions of TRIS [tris(hydroxymethyl)aminomethane] and related buffers with peptide backbone: thermodynamic characterization. *Phys Chem Chem Phys* **2010,** *12* (39), 12840-50.
431. Littlechild, J.; Rodriguez, E. G.; Isupov, M., Vanadium containing bromoperoxidase - Insights into the enzymatic mechanism using X-ray crystallography. *Journal of Inorganic Biochemistry* **2009,** *103* (4), 617-621.
432. Carter, J. N.; Beatty, K. E.; Simpson, M. T.; Butler, A., Reactivity of recombinant and mutant vanadium bromoperoxidase from the red alga *Corallina officinalis*. *J Inorg Biochem* **2002,** *91* (1), 59-69.
433. Zhu, G. Y.; Zhu, X.; Fan, Q.; Wan, X. L., Raman spectra of amino acids and their aqueous solutions. *Spectrochim Acta A* **2011,** *78* (3), 1187-1195.
434. Sane, S. U.; Cramer, S. M.; Przybycien, T. M., A holistic approach to protein secondary structure characterization using amide I band Raman spectroscopy. *Anal Biochem* **1999,** *269* (2), 255-272.
435. Saptarshi, S. R.; Duschl, A.; Lopata, A. L., Interaction of nanoparticles with proteins: relation to bio-reactivity of the nanoparticle. *J Nanobiotechnology* **2013,** *11*, 26.

436. Vangala, K.; Ameer, F.; Salomon, G.; Le, V.; Lewis, E.; Yu, L. Y.; Liu, D.; Zhang, D. M., Studying protein and gold nanoparticle interaction using organothiols as molecular Probes. *J Phys Chem C* **2012,** *116* (5), 3645-3652.
437. Chi, Z.; Chen, X. G.; Holtz, J. S.; Asher, S. A., UV resonance Raman-selective amide vibrational enhancement: quantitative methodology for determining protein secondary structure. *Biochemistry* **1998,** *37* (9), 2854-64.
438. Cialla, D.; Pollok, S.; Steinbrucker, C.; Weber, K.; Popp, J., SERS-based detection of biomolecules. *Nanophotonics* **2014,** *3* (6), 383-411.
439. Bro, R.; Andersen, C. M., Theory of net analyte signal vectors in inverse regression *Journal of chemometrics* **2003,** *17*, 646-652.
440. Vivó-Truyols, G.; Torres-Lapasió, J. R.; García-Alvarez-Coque, M. C., Net analyte signal as a deconvolution-oriented resolutions criterion in the optimisation of chromatograpic techniques. *Journal of Chromatography A* **2003,** *991*, 47-59.
441. Brown, C. D., Discordance between net analyte signal theory and practical multivariate calibration. *Anal. Chem.* **2004,** *76*, 4364-4373.
442. Bondu, S.; Cocquempot, B.; Deslandes, E.; Morin, P., Effects of salt and light stress on the release of volatile halogenated organic compounds by Solieria chordalis: a laboratory incubation study. *Botanica Marina* **2008,** *51* (6), 485-492.
443. Dummermuth, A. L.; Karsten, U.; Fisch, K. M.; König, G. M.; Wiencke, C., Responses of marine macroalgae to hydrogen-peroxide stress. *Journal of Experimental Marine Biology and Ecology* **2003,** *289* (1), 103-121.
444. Jones, B.; Nachtsheim, C. J., Definitive screening designs with ddded two-level categorical factors. *Journal of Quality Technology* **2013,** *45* (2), 121-129.
445. Jones, B.; Nachtsheim, C. J., Effective design-based model selection for definitive screening designs. *Technometrics* **2017,** *59* (3), 319-329.
446. Ekdahl, A.; Abrahamsson, K., A simple and sensitive method for the determination of volatile halogenated organic compounds in sea water in the amoll−1 to pmoll−1 range. *Analytica Chimica Acta* **1997,** *357* (3), 197-209.
447. Sundström, J.; Collén, J.; Abrahamsson, K.; Pedersén, M., Halocarbon production and in vivo brominating activity of Eucheuma denticulatum. *Phytochemistry* **1996,** *42* (6), 1527-1530.
448. Manley, S. L.; Dastoor, M. N., Methyl iodide (CH3I) production by kelp and associated microbes. *Marine Biology* **1988,** *98* (4), 477-482.
449. Thomas, H.; Bozec, Y.; Elkalay, K.; de Baar, H. J. W., Enhanced Open Ocean Storage of CO₂ from Shelf Sea Pumping. *Science* **2004,** *304* (5673), 1005-1008.
450. Kester, D. R.; Duedall, I. W.; Connors, D. N.; Pytkowicz, R. M., Preparation of artificial seawater. *Limnol Oceanogr* **1967,** *12* (1), 176-179.
451. Bravo-Linares, C.; Mudge, S. M.; Loyola-Sepúlveda, R., Production of volatile organic compounds (VOCs) by temperate macroalgae. The use of solid phase microextraction (SPME) coupled to GC-MS as method of analysis. *Journal of The Chilean Chemical Society* **2010,** *55*, 227-232.
452. Ekdahl, A.; Pedersén, M.; Abrahamsson, K., A study of the diurnal variation of biogenic volatile halocarbons. *Marine Chemistry* **1998,** *63* (1), 1-8.
453. Winter, J. M.; Moore, B. S., Exploring the chemistry and biology of vanadium-dependent haloperoxidases. *J Biol Chem* **2009,** *284* (28), 18577-18581.
454. Collén, J.; Pedersén, M., Production, scavenging and toxicity of hydrogen peroxide in the green seaweed Ulva rigida. *European Journal of Phycology* **1996,** *31* (3), 265-271.
455. Theiler, R.; Cook, J. C.; Hager, L. P.; Siuda, J. F., Halohydrocarbon synthesis by bromoperoxidase. *Science* **1978,** *202* (4372), 1094-6.
456. Pinhassi, J.; Winding, A.; Binnerup, S. J.; Zweifel, U. L.; Riemann, B.; Hagström, Å., Spatial variability in bacterioplankton community composition at the Skagerrak-Kattegat front. *Marine Ecology Progress Series* **2003,** *255*, 1-13.

457. McKnight, D. M.; Aiken, G. R., Sources and age of aquatic humus. In *Aquatic Humic Substances: Ecology and Biogeochemistry*, Hessen, D. O.; Tranvik, L. J., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 1998; pp 9-39.
458. Liu, Y. N.; Thornton, D. C. O.; Bianchi, T. S.; Arnold, W. A.; Shields, M. R.; Chen, J.; Yvon-Lewis, S. A., Dissolved organic matter composition drives the marine production of brominated very short-lived substances. *Environ Sci Technol* **2015,** *49* (6), 3366-3374.
459. Keng, F. S.-L.; Phang, S.-M.; Rahman, N. A.; Leedham, E. C.; Hughes, C.; Robinson, A. D.; Harris, N. R. P.; Pyle, J. A.; Sturges, W. T., Volatile halocarbon emissions by three tropical brown seaweeds under different irradiances. *Journal of Applied Phycology* **2013,** *25* (5), 1377-1386.
460. Palmer, C. J.; Anders, T. L.; Carpenter, L. J.; Küpper, F. f. C.; McFiggans, G. B., Iodine and halocarbon response of *Laminaria digitata* to oxidative stress and links to atmospheric new particle production. *Environmental Chemistry* **2005,** *2* (4), 282-290.
461. Martino, M.; Liss, P. S.; Plane, J. M. C., Wavelength-dependence of the photolysis of diiodomethane in seawater. *Geophysical Research Letters* **2006,** *33* (6).
462. Jones, C. E.; Carpenter, L. J., Solar photolysis of CH2I2, CH2ICl, and CH2IBr in water, saltwater, and seawater. *Environ Sci Technol* **2005,** *39* (16), 6130-6137.
463. Martino, M.; Liss, P. S.; Plane, J. M. C., The photolysis of dihalomethanes in surface seawater. *Environ Sci Technol* **2005,** *39* (18), 7097-7101.
464. Amachi, S., Microbial contribution to global iodine cycling: volatilization, accumulation, reduction, oxidation, and sorption of iodine. *Microbes Environ* **2008,** *23* (4), 269-76.
465. Carpenter, L. J.; Archer, S. D.; Beale, R., Ocean-atmosphere trace gas exchange. *Chemical Society Reviews* **2012,** *41* (19), 6473-6506.
466. Yokouchi, Y.; Osada, K.; Wada, M.; Hasebe, F.; Agama, M.; Murakami, R.; Mukai, H.; Nojiri, Y.; Inuzuka, Y.; Toom-Sauntry, D.; Fraser, P., Global distribution and seasonal concentration change of methyl iodide in the atmosphere. *Journal of Geophysical Research: Atmospheres* **2008,** *113* (D18).
467. Eiler, A.; Johansson, M. P.; Bertilsson, S., Environmental influences on *Vibrio* populations in northern temperate and boreal coastal waters (Baltic and Skagerrak Seas). *Appl Environ Microbiol* **2006,** *72* (9), 6004-6011.
468. Goecke, F.; Labes, A.; Wiese, J.; Imhoff, J. F., Chemical interactions between marine macroalgae and bacteria. *Marine Ecology Progress Series* **2010,** *409*, 267-299.
469. Valcárcel, M., Analytcal chemistry today and tomorrow. In *Analytical Chemistry*, Krull, I. S., Ed. IntechOpen: 2012.
470. Karayannis, M.; Efstathiou, C., Significant steps in the evolution of analytical chemistry - is the today's analytical chemistry only chemistry? *Talanta* **2012,** *102*, 7-15.