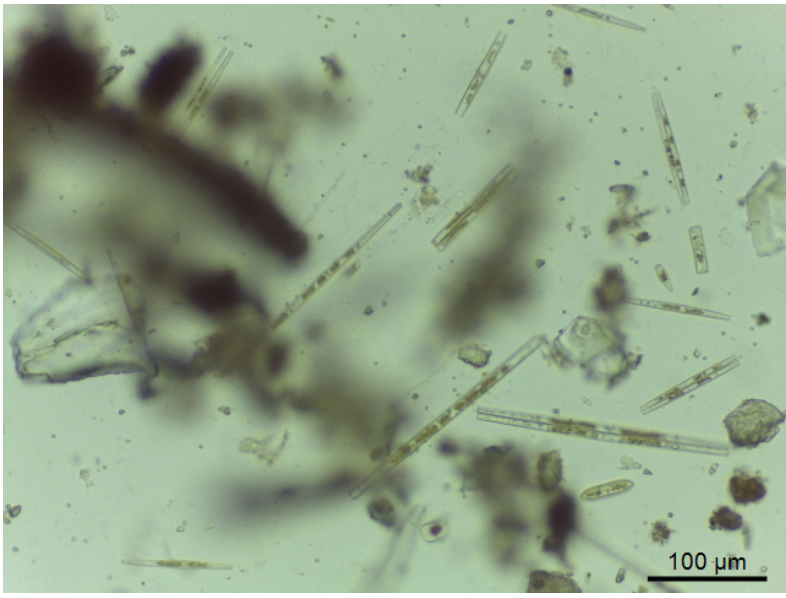


Raman Spectroscopy applied to diatoms for Environmental diagnosis

Raquel Abrunhosa Pinto

Mestrado em Biologia e Gestão da Qualidade da Água

Departamento de Biologia

2020

Orientador

Luís Filipe de Oliva Teles, Professor Auxiliar, FCUP

Coorientadores

António Paulo Carvalho, Professor Auxiliar, FCUP

Laura Guimarães, Investigadora Auxiliar, CIIMAR

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas

pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/__

Agradecimentos

Ao Professor Doutor Luís Filipe de Oliva Teles e à Doutora Laura Guimarães por tudo que me transmitiram, por todo o apoio na resolução de problemas e disponibilidade para me ajudar. A enorme quantidade de conhecimento e polivalência que ambos possuem foi e continua a ser uma inspiração para mim. Muito obrigada.

Ao Professor Doutor António Paulo Carvalho pelo apoio e perfecionismo que foi muito útil tanto no planeamento desta dissertação como nos detalhes da sua execução. Muito obrigada.

Ao Professor Doutor Joaquim Agostinho Moreira e ao Rui Vilarinho pela disponibilidade e partilha de conhecimentos na área da Física de forma tão interessante e simples. Muito obrigada pela colaboração.

Aos meus colegas de laboratório pela forma como me acolheram, pela amizade e por toda a ajuda durante esta etapa. Muito obrigada.

Obrigada a todos os meus amigos que trazem alegria e positividade aos meus dias. Em especial ao Jorge que me acompanhou durante todo o meu percurso e que sempre acreditou em mim, por vezes mais que eu própria. Obrigada por tudo, Amo-te.

Á minha família que me ajudou em tudo para eu chegar até aqui, especialmente aos meus pais por todos os sacrifícios. Obrigada, gosto muito de vocês.

Resumo

A qualidade da água tem vindo a diminuir devido à sua utilização pouco razoável e à poluição causada pela pressão antropogénica. Isto afeta as atividades humanas e a saúde dos ecossistemas. Consequentemente, é extremamente importante desenvolver metodologias mais rápidas e melhoradas para avaliar e monitorizar a qualidade deste importante recurso. As diatomáceas são bioindicadores reconhecidos e têm sido utilizadas como elemento biológico para avaliar a qualidade da água não só na Europa, ao abrigo da Directiva-Quadro da Água, mas também em estudos de monitorização em ambiental em todo o mundo. Os métodos convencionais para avaliar a qualidade da água usando as diatomáceas baseiam-se na identificação taxonómica das espécies e na contagem de valvas. Apesar das diatomáceas possuírem uma frústula siliciosa com padrões intricados e formas que diferem entre espécies, a identificação taxonómica requer bastante experiência, é demorada e as designações e afinidades dos *taxa* mudam frequentemente. A espectroscopia Raman é uma técnica simples e rápida, que não exige preparação das amostras, com potencial para ser utilizada no diagnóstico ambiental e na identificação de diatomáceas. Assim, para apurar a robustez deste pressuposto, realizou-se em primeiro lugar uma revisão crítica da informação existente sobre os recentes avanços da espectroscopia Raman aplicada às diatomáceas. De seguida, foi realizado um estudo experimental que incluiu a obtenção de 790 espetros de diatomáceas pertencentes a 29 espécies recolhidas em três lagos do Parque da Cidade do Porto (Portugal). Com estes dados foi analisada a estrutura de covariância multidimensional entre as variáveis Raman obtidas (matriz das variáveis independentes ou matriz X) e as espécies recolhidas nos três lagos (matriz das variáveis dependentes ou matriz Y) pelo método de regressão de mínimos quadrados parciais (PLS). Os dados foram também modelados com uma rede neural artificial com o objetivo de identificar *taxa* a partir dos dados Raman e diagnosticar os três lagos estudados em função das suas características ambientais.

A revisão dos estudos de Raman em diatomáceas encontrados na literatura mostrou que estes se dedicaram principalmente a investigar a estrutura, localização e conformação de componentes das células de diatomáceas, bem como a sua variação sob diferentes condições ambientais. Desta revisão ressaltou também a escassez de trabalhos relativos à aplicação da espectroscopia Raman a diatomáceas para diagnóstico ambiental e identificação taxonómica. A revisão forneceu ainda assinaturas espetrais e parâmetros Raman importantes para os estudos sobre as diatomáceas do

Parque da Cidade. O trabalho experimental permitiu identificar um total de 14 bandas características do espectro Raman das diatomáceas. A maioria destas bandas Raman, atribuíveis a vibrações moleculares de pigmentos e de componentes da frústula, variaram em função da espécie. O método de PLS permitiu descrever um perfil Raman para cada espécie que pode ser usado para a compreensão do estado fisiológico das diferentes espécies. O método de redes neurais permitiu identificar classes e ordens de diatomáceas com exatidão a variar entre suficiente e excelente (67-89%) dependendo do *taxa*. Globalmente as espécies identificadas com sensibilidade mais elevada foram: *Achananthidium exiguum* (67%), *Fragilaria crotonensis* (67%), *Amphora pediculus* (71%), *Achananthidium minutissimum* (80%) e *Melosira varians* (82%). A aplicação do método ao diagnóstico ambiental permitiu diagnosticar os lagos amostrados no Parque da Cidade do Porto com excelente exatidão (89% a 96%), fornecendo ótimos resultados de classificação. Para além disso, o modelo com exatidão de 89% permitiu a classificação utilizando apenas as variáveis Raman obtidas, sem necessidade de identificação taxonómica das espécies. Tanto quanto foi possível esclarecer, este é o primeiro estudo que fornece evidências empíricas sobre as vantagens e a precisão da espectroscopia Raman aplicada a diatomáceas para a identificação taxonómica o diagnóstico ambiental de ecossistemas de água doce. Os resultados obtidos constituem uma importante base para estudos de avaliação de ecossistemas a uma escala mais vasta, incluindo diferentes localizações geográficas e condições ambientais.

Palavras-Chave: Espetros, Bandas Espetrais, Saúde ambiental, Níveis Taxonómicos, Testes Diagnóstico, Pigmentos, Frústula, Quimiometria

Abstract

Worldwide, water quality in aquatic ecosystems has been decreasing. Factors accounting for this loss of quality are its unreasonable use and pollution caused by human activities. This affects both human activities and the health of aquatic ecosystems. Therefore, it is extremely important to develop fast and improved methodologies for assessing and monitoring the quality of this vital resource. Diatoms are recognized bioindicators and have been used as a biological element for assessing water quality not only in Europe, under the Water Framework Directive, but also in monitoring surveys all over the world. The conventional methods to assess water quality based on diatoms rely on taxonomic identification and valve count. Despite the fact that diatoms possess a siliceous frustule with intricate patterns and shapes differing among species, taxonomic identification requires sound expertise, is time-consuming and *taxa* designations and affinities change frequently. Application of Raman Spectroscopy to diatoms show great potential for environmental and *taxa* identification. To ascertain the robustness of this assumption, a literature review of the information available about the recent advances on the application of Raman Spectroscopy to diatoms was first carried out. An experimental study was then carried out by acquiring 790 Raman spectra of 29 different diatom species collected from three lakes located in Oporto City Park (Portugal). These data were used to analyze the multidimensional covariance structure among the Raman variables obtained (matrix of independent variables or matrix X) and the species collected in the three lakes (matrix of dependent variables or matrix Y) by the method of partial least squares regression (PLS). The data were also modeled with an artificial neural network (ANN) in order to infer about *taxa* identification and diagnose the three studied lakes according to their environmental characteristics.

The results of the literature review revealed that Raman spectroscopy studies were mainly focused on characterizing the structure, location and conformation of diatom cell components, as well as their variation under different laboratory-controlled conditions. This review also highlighted the scarcity of studies on the application of Raman spectroscopy to diatoms for environmental diagnosis and taxonomic identification. The review also provided spectral signatures and important Raman parameters for studies on the diatoms of Oporto City Park. The experimental work made it possible to identify a total of 14 bands typical of diatom Raman spectra. Most of these Raman bands, attributable to molecular vibrations of pigments and frustule components, varied according to the species. The PLS method allowed to describe a Raman profile for each

species, which can be used to improve understanding on the physiological status of the different species. The artificial neural network method made it possible to identify classes and orders of diatoms with an accuracy ranging from sufficient to excellent (67-89%) depending on the *taxa*. Overall, the species identified with the highest sensitivity were: *Achananthidium exiguum* (67%), *Fragilaria crotonensis* (67%), *Amphora pediculus* (71%), *Achananthidium minutissimum* (80%) and *Melosira varians* (82%). Application of this method to the environmental diagnosis made it possible to diagnose the lakes sampled in Oporto City Park with excellent accuracy (89% to 96%). In addition, the model with 89% accuracy allowed classification using only the Raman variables obtained, without the need for taxonomic identification of the species. As far as it was possible to clarify, this is the first study providing empirical evidence on the advantages and precision of Raman spectroscopy applied to diatoms for taxonomic identification and environmental diagnosis of freshwater ecosystems. The results obtained in this work constitute an important basis for future studies carried on at a wider scale, including different geographical locations and environmental conditions, to assess freshwater ecosystems.

Keywords: Spectra, Spectral bands, Environmental Health, Taxonomic Levels, Diagnostic Tests, Pigments, Frustule, Chemometrics

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A – Area

ADMI – *Achanthidium minutissimum*

ADSB – *Achnanthidium straubianum*

AEXG – *Achnanthidium exiguum*

ANN – Artificial Neural Networks

ANOVA – Analysis of variance

APED – *Amphora pediculus*

APTES – Aminopropyltriethoxysilane

ASP – Amnesic Shellfish Poisoning

AuNPs – Gold Nanoparticles

AVEN – *Amphora veneta*

BP – Sample from Buchang deposit in Haikang county of Guangdong province, China

CARs – Coherent Anti-Stokes Raman Spectroscopy

Chl – Chlorophyll

CLIAs – Chemiluminescent Immunoassays

COD – Chemical Oxygen Demand

COVID-19 – Corona virus disease 2019

CPUL – *Ctenophora pulchella*

CS – Common Species

CSTE – *Cyclotella stelligera*

CTUM – *Cymbella tumida*

DA – Domoic Acid

Ddx – Diadinoxanthin

DMS – Dimethyl Sulphide

DO – Dissolved Oxygen

DTT – Ditiotreitol

Dtx – Diatoxanthin

ECVAM – European Centre for the Validation of Alternative Methods

ELISA – Enzymatic-Linked Immunosorbent Assays

EOMI – *Eolimna minima*

EPS – Extra-cellular Polymeric Substances

F – Frequency

FCPs – Fucoxanthin-Chlorophyll Protein Complexes

FCRO – *Fragilaria crotonensis*

FCVA – *Fragilaria vaucheriae*

FT-IR – Fourier-Transformed Infrared Spectroscopy

Fx – Fucoxanthin

GAFF – *Gomphonema affine*

GEXL – *Gomphonema exilissimum*

GLGN – *Gomphonema lagenula*

GMR – Guided Magnetic Resonance

GPAR – *Gomphonema parvulum*

HL – High-light

HPLC – High-Performance Liquid Chromatography

HTS-RS - High-throughput Screening Raman Spectroscopy

IBD – *Indice Biologique Diatomées*

IgG – Immunoglobulin G

IgM – Immunoglobulin M

IPS – *Indice de Poluosensibilité Spécifique*

L - Lake

L1 – Lake 1

L2 – Lake 2

L3 – Lake 3

LFIAs – Lateral Flow Immunoassays

LL – Low-light

LSPs – Low Surface Plasmons

MLP - Multilayer Perceptron

MVAR – *Melosira varians*

NAMP – *Nitzschia amphibia*

NCRY – *Navicula cryptocephala*

NCTE – *Navicula cryptotenella*

NFON – *Nitzschia fonticola*

NGRE – *Navicula gregaria*

NINC – *Nitzschia inconspicua*

NNOT – *Navicula notha*

NPAL – *Nitzschia palea*

NPQ – Non-photochemical quenching

NPs – Nanoparticles

NSBC – *Nitzschia subcapitellata*

PBRE – *Pseudostaurosira brevistriata*

PCA – Principal Components Analysis

PLFR – *Planothidium frequentissimum*

PLS – Partial Least Squares

PLS-DA – Partial Least Squares Discriminant Analysis

PLS-DA – Partial Least Squares Discriminant Analysis

Raman-OLS – Raman Ordinary Least Squares

RCNN - Region based Convolutional Neural Networks

RI – Raman Imaging

RRS – Resonant Raman Spectroscopy

RS – Raman Spectroscopy

RT-PCR – Real-Time Polymerase Chain Reaction

SERS – Surface-Enhanced Raman Spectroscopy

SFD – Saturated Fatty-Acid Dominated

SUFT – Suturated-Unsaturated Fatty-Acid Transformation

TSS – Total Suspended Solids

TTAB – *Tabularia tabulata*

UFD – Unsaturated Fatty-Acid Dominated

USA – United States of America

UULN – *Ulnaria ulna*

UV – Ultra-violet

W – Width

WFD – Water Framework Directive

WQI – Water Quality Index

YP – Sample from Yuanjiawan deposit in Shengxian county of Zhejiang province, China

Chapter 1 - General Introduction

Water quality has been changing due to anthropogenic pressure which implies unreasonable water use and crescent water pollution (UNESCO-WHO-UNEP, 1996). With the decrease in water quality, domestic, industrial, agricultural and recreative activities, fisheries and aquaculture production and the health of ecosystems are affected (Boyd, 2019). Therefore, the creation of fast and improved ways to assess water quality has gained extremely importance (UNESCO-WHO-UNEP, 1996).

Diatoms are well established bioindicators of pollution due to their large geographical distribution (Round *et al.*, 2007) and their fast and differential responses to changes in environmental conditions (Almeida & Gil, 2001; Squires *et al.*, 1979; Vilbaste & Truu, 2003). These siliceous microalgae are biological elements of the aquatic flora obligatory to assess water quality in Europe under the Water Framework Directive (WFD) (European Comission Council, 2000). Apart from Europe, diatoms are also used in ecological surveys all over the world (UNESCO-WHO-UNEP, 1996). Usually, the use of diatoms for water assessment requires an accurate taxonomic identification and valve counting (Pandey *et al.*, 2017; Pinto *et al.*, 2020). Valve abundance is then used for the calculation of autoecological indexes (CEMAGREF, 1982). Diatoms have siliceous frustules differing among species in pattern and shape (Pandey *et al*., 2017). Despite these morphometric differences, taxonomic identification is very challenging (Morin *et al*.,2016). Accurate taxonomic identification requires advanced expertise, is time-consuming and *taxa* designations and affinities are recurrently changing (Morin *et al*., 2016). Additionally, the ecological indexes based on diatoms may be region-specific but are often applied indiscriminately across regions (Morin *et al*., 2016).

In recent years Raman spectroscopy (RS) has come up as a simple, faster and label-free technique that could be applied to environmental diagnosis with diatoms. This method enables the acquisition of spectra of selected organisms with several bands corresponding to the vibrations and conformations of the atomic bonds of each molecule (Smith & Dent, 2019). Despite the wealth of studies about the application of Raman spectroscopy to diatoms produced in the past descades, the use of this approach is poorly explored in literature. A previous work has suggested the interest of vibrational spectroscopy, mostly Fourier-transform infrared spectroscopy (FTIR) (Akkas and Severcan, 2012) for monitoring aquatic systems, but no empirical work is available showing the feasibility of environmental diagnostic with Raman spectroscopy applied to

diatoms. Therefore, the main objective of this work was to investigate if Artificial Neural Networks (ANN) based on RS data from diatoms could be used to diagnose different *taxa* and environmental conditions. To address this research question, four specific goals were established for this study:

1. To gather information on recent advances regarding RS applied to diatoms;
2. To depict the differences in Raman bands acquired on different diatom species;
3. To depict and compare Raman bands obtained from diatoms under different environmental conditions;
4. To test the applicability of ANN based on diatom Raman spectral data for environmental diagnosis and *taxa* diagnosis.

To test the discriminatory ability of ANN, the work was developed with diatoms collected from three lakes with similar hydromorphological conditions and interconnected water circulation. The following three chapters of this dissertation present the results obtained with the work developed. They are organized as follows: Chapter 2 presents the critical review of the literature carried out, structured as a review paper. This paper, entitled "*Advances in Raman Spectroscopy applied to Diatoms (microalgae, Bacillariophyta) with focus on its use for environmental diagnosis of freshwater ecosystems*", was submitted for publication in an international scientific journal (*Water Research*) and presents the information found for diatom RS studies about the composition, structure and conformation of their cell components and toxicological assays performed. Future perspectives about the contribution of studies and methods to freshwater environmental diagnosis are also discussed. Chapter 3, entitled "*A practical technique to identify Diatom taxa: Raman Spectroscopy*" presents the differences in Raman bands found among the species investigated and the use of ANN to predict different taxonomic levels. Chapter 4, is also structure in the form of a paper (to be submitted soon for publication). It is entitled "*Environmental diagnosis with Raman Spectroscopy applied to diatoms*". Here the aim was to present results of Raman spectra acquired in three lakes of the Oporto City Park, compare them and develop ANN models for lake diagnosis with estimation of measures of performance required for diagnostic tests (accuracy and sensitivity). The two final chapters of this dissertation are dedicated to the general discussion, conclusions and future perspectives of the work performed.

References

Almeida, S. F. P., & Gil, M. C. P. (2001). dÉcologie des diatomées d'eau douce de la région centrale du Portugal. *Cryptogamie Algologie, 22*(1), 109-126.

Boyd, C. E. (2019). *Water quality: an introduction*: Springer Nature.

CEMAGREF, M. (1982). Etude des méthodes biologiques d'appréciation quantitative de la qualité des eaux. *Rapport Cemagref QE Lyon-AF Bassin Rhône Méditerranée Corse*.

Coste, M., Boutry, S., Tison-Rosebery, J., & Delmas, F. (2009). Improvements of the Biological Diatom Index (BDI): Description and efficiency of the new version (BDI-2006). *Ecological indicators, 9*(4), 621-650.

Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy, (2000).

Morin, S., Gómez, N., Tornés, E., Licursi, M., & Rosebery, J. (2016). Benthic diatom monitoring and assessment of freshwater environments: standard methods and future challenges. *Aquatic Biofilms*, 111.

Pandey, L. K., Bergey, E. A., Lyu, J., Park, J., Choi, S., Lee, H., . . . Han, T. (2017). The use of diatoms in ecotoxicology and bioassessment: insights, advances and challenges. *Water Research, 118*, 39-58.

Pinto, R., Mortágua, A., Almeida, S. F., Serra, S., & Feio, M. J. (2020). Diatom size plasticity at regional and global scales. *Limnetica, 39*(1), 387-403.

Round, F. E., Crawford, R. M., & Mann, D. G. (2007). *Diatoms: biology and morphology of the genera*: Cambridge university press.

Smith, E., & Dent, G. (2019). *Modern Raman spectroscopy: a practical approach*: John Wiley & Sons.

Squires, L. E., Rushforth, S. R., & Brotherson, J. D. (1979). Algal response to a thermal effluent: study of a power station on the provo river, Utah, USA. *Hydrobiologia, 63*(1), 17-32.

UNESCO-WHO-UNEP. (1996). *Water Quality Assessments*. Cambridge, UK: Chapman & Hall.

Vilbaste, S., & Truu, J. (2003). Distribution of benthic diatoms in relation to environmental variables in lowland streams. *Hydrobiologia, 493*(1-3), 81-93.

Chapter 2 - Advances in Raman Spectroscopy applied to Diatoms (microalgae, Bacillariophyta) with focus on its use for environmental diagnosis of freshwater ecosystems

Raquel Pinto[2], Rui Vilarinho[3], António Paulo Carvalho[2], J. Agostinho Moreira[3], Laura Guimarães[1], Luís Oliva-Teles[2]

[1] CIIMAR/CIMAR - Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Avenida General Norton de Matos, s/n 4450-208 Matosinhos, Portugal
[2] Department of Biology, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n, 4169-007, Porto, Portugal
[3] Department of Physics, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n. 4169-007, Porto, Portugal

Abstract

Diatom species are good pollution bioindicators due to their large distribution, fast response to changes in environmental parameters and different tolerance ranges. These organisms are used in ecological water assessment all over the world. Such assessments commonly rely on taxonomic identification of diatom species-specific shape and frustule ornaments, from which cell counts, species richness and diversity indices can be estimated. The taxonomic identification is, however, time-consuming and requires sound expertise. Additionally, the ecological indexes based on diatoms may be region-specific but are often applied indiscriminately across regions.

Raman spectroscopy is a simpler, faster and label-free technique that can be applied to environmental diagnosis with diatoms. However, this approach has been poorly explored. This work reviews Raman spectroscopy studies involving the structure, location and conformation of diatom cell components and their variation under different conditions. This knowledge provides a strong foundation for the development of environmental protocols using Raman spectroscopy in diatoms. Our work aims at

stimulating further research on the application of Raman spectroscopy as a tool to assess physiological changes and water quality under a changing climate.

Keywords: Raman spectra, SERS, diagnostic tests, pigments, lipids, frustule

1. Introduction

Diatoms are unicellular algae that can be found all over the world in practically all kinds of aquatic environments. Due to their large distribution (Round *et al.*, 2007), morphological diversity (Guiry, 2012), fast response to environmental changes and the existence of species with different tolerance ranges (Almeida & Gil, 2001; Squires *et al.*, 1979; Vilbaste & Truu, 2003), diatoms are recognized bioindicators and have been employed in the assessment of water quality for decades (Blanco *et al.*, 2004; Desrosiers *et al.*, 2013; Dixit *et al.*, 1999; Patrick, 1973). These microalgae are easy to sample, process (Lear *et al.*, 2012) and store (Mendes *et al.*, 2012). Additionally, their taxonomic identification is made through diatom floras based on their siliceous cell wall - the frustule - which has ornaments differing among species (Figure 1; Germain, 1981; Pandey *et al.*, 2017). In Europe, according to the Water Framework Directive (WFD), diatoms are a biologic element of the aquatic flora, of obligatory evaluation for the ecological assessment of water courses in European Union countries (European Comission Council, 2000). Diatoms are also used in routine water monitoring surveys in other parts of the world, such as Canada, USA, Japan, South America and Australia (UNESCO-WHO-UNEP, 1996).

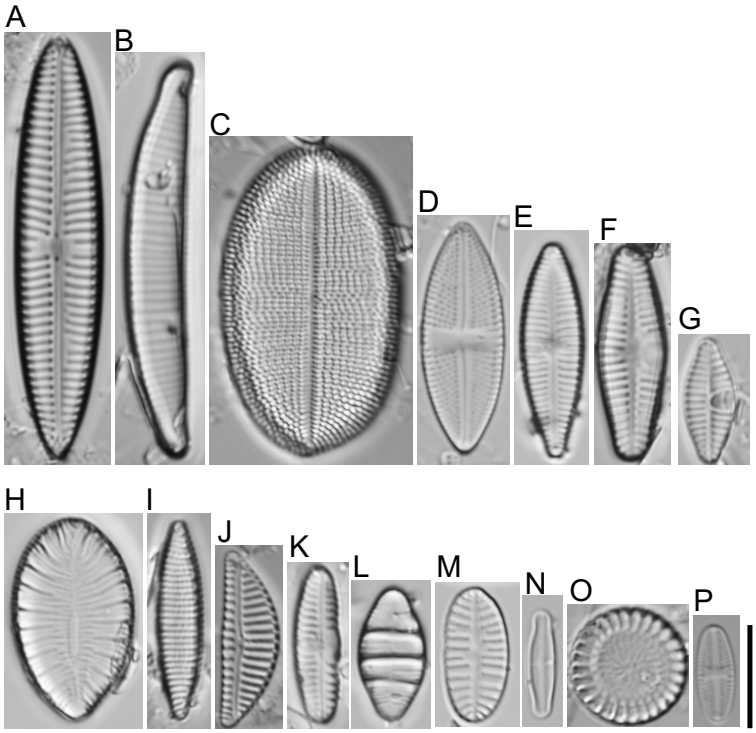


Figure 1 - Examples of the diversity of intricate patterns and shapes characteristic of diatoms. A – *Navicula tripunctata*; B – *Eunotia incisa* var. *incisa*; C – *Cocconeis placentu*la var. *lineata*; D – *Luticola goeppertiana*; E – *Gomphonema saprophilum*; F – *Planothidium lanceolatum*; G – *Planothidium frequentissimum*; H - *Surirella brebissonii*; I – *Nitzschia amphibia*; J – *Encyonema silesia*cum; K – *Reimera sinuata*; L – *Diatoma mesodon;* M - *Karayevia oblongella*; N – *Achnanthidium minutissimum*; O – *Cyclotella meneghiniana*; P – *Eolimna mínima*. Scale bar = 10µm.

The conventional methods for assessing water quality using diatoms are based on taxonomic identification and valve counting (Pandey *et al.*, 2017; Pinto *et al.*, 2020). The species abundance is then used for the calculation of autoecological indexes such as, for example, the IPS - Indice de Polluosensibilité Spécifique (CEMAGREF, 1982) or IBD - Indice Biologique Diatomées (Coste *et al.*, 2009). This approach has some drawbacks, since it is time-consuming, *taxa* are recurrently changing, and taxonomic identification requires many years of expertise. Additionally, ecological indexes are sometimes region specific and consequently can be applied inaccurately (Morin *et al.*, 2016).

Raman spectroscopy (RS) is a promising alternative technique for application to monitoring and diagnostic testing of water with diatoms. In biological analysis, this technique has many advantages in relation to other methods because it is label-free and it requires a minimal or a simpler sample preparation and processing (Heraud *et al.*, 2007). Additionally, water is a low scatterer so its interference can be minimized (Parker, 1983) and the inherent fluorescence of the organic compounds can be minimized using low laser power and Raman imaging (RI) based on chemistry and structure of the compounds of interest. The technique was first described in 1928 and is based on the

scattered radiation, resulting when an incident electromagnetic beam, with an excitation wavelength ranging from near-infrared to UV, is inelastically scattered by molecular vibrations in the sample (Raman & Krishnan, 1928). Figure 2 shows a schematic representation of different types of light scattering. The majority of the photons are elastically scattered with no changes in energy when they interact with the molecules– Rayleigh scattering (Smith & Dent, 2019). Yet, a smaller part ($1:10^6$) of these photons are inelastic scattered (Smith & Dent, 2019). Raman scattering is a type of inelastic light scattering by molecular vibrations, in which the energy of the incident photon can be either larger or lower than the energy of the scattered photon resulting on positive or negative frequency shifts (Stokes or Anti-Stokes Raman scattering, respectively) (Smith & Dent, 2019). These Raman shifts, i.e. the difference between the excitation frequency and the direct frequency, consist on direct measurements of the vibration frequency (Talari *et al.*, 2015). Therefore, the Raman spectrum gives information about molecular structure and composition (Smith & Dent, 2019) of a selected material or individual within a sample. Raman spectroscopy reflects alterations in molecular structure that are consequence of diverse interactions such as hydration, links to membranes or particles, alterations in molecular conformation, etc.

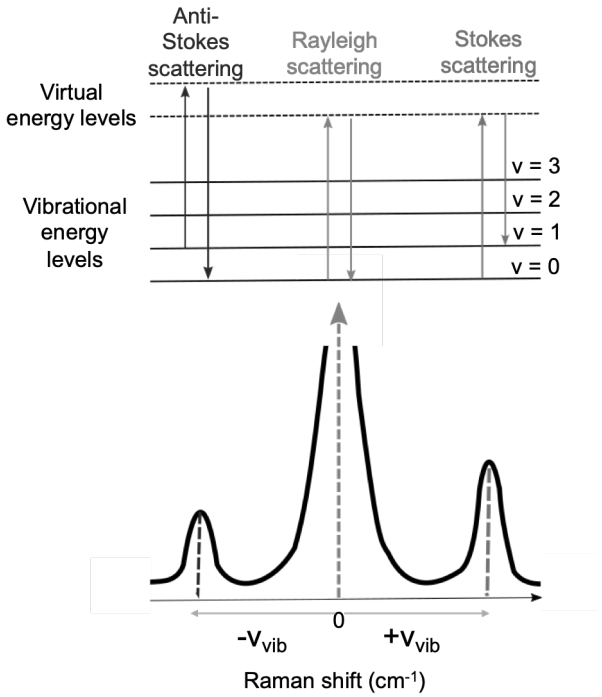


Figure 2 - Energetic transitions and spectral representation of different types of scattered radiation: Rayleigh scattering (elastic scattering), and Raman Stokes and Anti-Stokes scattering (inelastic scattering). v_{vib} – Raman shift, a direct measure of the frequency of excitation

.

In the case of diatoms, the main spectral information obtained through RS are bands assigned to photosynthetic and photoprotective pigments – chlorophylls (Chls) *a* and *c*, and carotenoids such as fucoxanthin (Fx), diadinoxanthin (Ddx) and diatoxanthin (Dtx) (e.g. Alexandre *et al.*, 2014; Premvardhan *et al.*, 2009; Premvardhan *et al.*, 2010). Raman spectroscopy also enables the possibility to obtain spectral information about lipids (e.g. Meksiarun *et al.*, 2015), the frustule (e.g. Yuan *et al.*, 2004), extracellular polymeric substances (EPS) (e.g. Laviale *et al.*, 2019), mucilage (e.g. Chen *et al.*, 2019) and toxins (e.g. Wu *et al.*, 2000).

Studies concerning direct RS applications in diatoms are mainly focused on understanding the diversity of structures and conformations of the main molecules within the cell, the location of molecules within the cell, the optimal abiotic conditions to enhance the production of lipids to be used in biofuel industry and the variation of pigmentation under different light conditions. Raman spectroscopy is frequently complemented by Raman Imaging (RI). Raman imaging consists on taking several Raman spectra covering a certain area and mapping a certain band parameter of interest (intensity or frequency, for example) in that area. With this method, it is possible to locate components within the cell through colour schemes (Stewart, 2012). Usually, the quantitative analyses of Raman data consists on spectral deconvolution using multiple models. For this procedure, the frequency, intensity and width of each band can be determined. Different statistical analysis of the spectral information can be used in order to analyze band variations such as Principal Component Analysis (PCA) or Partial Least Squares Discriminant Analysis (PLS-DA) (Meksiarun *et al.*, 2015; Rüger *et al.*, 2019). Sometimes, application of RS is limited, due to either the low Raman cross-sections of the molecules, which may result in a low scattered signal (Shahbazyan & Stockman, 2013), or the high fluorescence characteristic of some biological samples which may interfere with Raman signal (Byrne *et al.*, 2016). To overcome these problems, some variations of RS can be used, namely the Resonant Raman spectroscopy (RRS) and the Surface-Enhanced Raman Spectroscopy (SERS). In diatoms, RRS is particularly useful to study pigments and consists in choosing the energy of the incident laser close to the energy required for the electronic transition of a target compound. This allows achieving resonance conditions enhancing the intensity of the Raman signal (McCreery, 2005). Surface-enhanced Raman Spectroscopy consists in using a metallic surface or metallic nanoparticles to enhance the Raman signal through plasmonic processes (Baena & Lendl, 2004). It is most frequently used in studies where diatoms are used as biosensors rather than the being the direct organism under investigation. In SERS, diatom frustules

are used as a substrate for the nanometallic surface and act as biosensors to detect the presence of other substances (e.g. Kamińska *et al.*, 2017).

Despite the relatively high number of RS studies applied to diatoms, these have been mostly done under controlled laboratory conditions, whereas field environmental, taxonomic and toxicological studies are lacking. Nevertheless, this represents a promising approach, since Raman bands assigned to diatom cell components can vary with abiotic and physiological conditions. It was previously demonstrated, for instance, that high CO_2 levels in the water can favour the production of fatty acids in the diatom *Thalassiosira pseudonana* (Meksiarun *et al.*, 2014) or that the growth phases in the diatom *Ditylium brightwellii* are accompanied by changes in Raman bands assigned to carotenoids (Rüger *et al.*, 2016). The objective of this review is, thus, to describe the recent advances in RS applied to diatoms and comment on the evidence supporting the use of this technique as a tool for environmental diagnostic tests. Emphasis was given to the qualitative and quantitative analysis of cell constituents and RS application in diatom toxicity assays, both of interest for the diagnostic testing of water quality and the environmental status of aquatic ecosystems.

2. Qualitative and quantitative analysis of cell components

2.1. Pigments

In diatoms, as in other photosynthetic organisms, exist two types of pigments: Chls and carotenoids (Kuczynska *et al.*, 2015). Two types of chlorophylls, represented in figure 3, are present: Chl *a* and Chl *c* (divided in Chl c_1 and Chl c_2). Chlorophylls contain a porphyrin ring constituted by a magnesium coordination complex of cyclic tetrapyrroles and a fifth isocyclic ring (Jeffrey & Vesk, 2005). Chlorophyll *a* (figure 3a) differs from the two types of Chl *c* (figures 3b and 3c) in that it contains a long phytyl chain in the 17^{th} position of the porphyrin ring. Chlorophyll c_2 (figure 3c) differs from Chl c_1 (figure 3b) in that it contains a vinyl group in the 8^{th} position of the porphyrin ring instead of an ethyl group (Wagner & Waidelich, 1986).

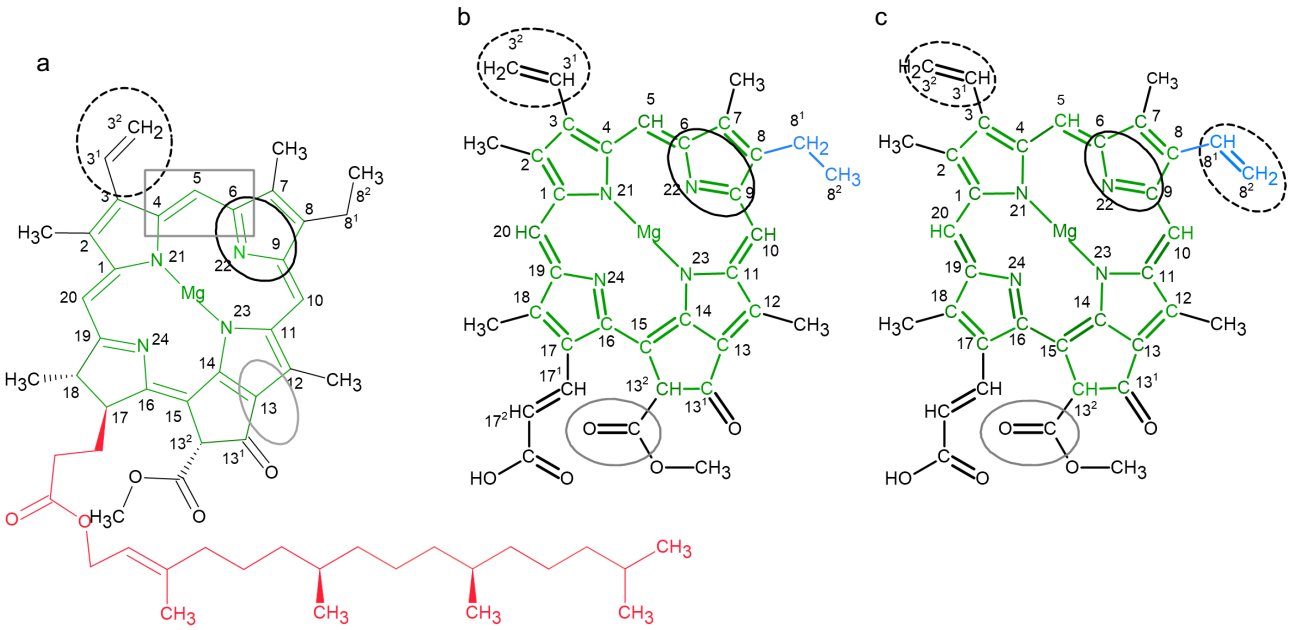


Figure 3 - Molecular structures of Chlorophylls a (a), c1 (b) and c2 (c). The magnesium coordination complex of tetrapyrroles with a fifth isocyclic ring characteristic of the chlorophylls is represented in green; the long phytyl chain that distinguishes Chlorophyll a is represented in red and the ethyl (b) and vinyl (c) groups that distinguish chlorophyll c1 from c2 are represented in blue. The most important groups contributing to Raman spectra are indicated by circles or squares: C-13 keto-carbonyl groups (grey circles), C-N pyrrole modes (black circles), methine bridges (grey squares) and vinyl (dashed circles).

Fucoxanthin-Chlorophyll Protein Complexes (FCPs) are also relevant for the study of diatom pigments. Figure 4l presents RRS spectra of *Cyclotella meneghiniana* FCPs obtained at several laser excitation wavelengths (Premvardhan *et al.*, 2010). These oligomeric or trimeric complexes constitute the main light-harvesting complexes in diatoms (Guglielmi *et al.*, 2005; Lavaud *et al.*, 2002; Lepetit *et al.*, 2010). Through the Raman spectra of FCPs, the characteristic spectral regions of chlorophylls and carotenoids can be identified. Raman spectra of Chls in diatoms were first studied in 1986 in the species *Phaeodactylum tricornutum* and *Gomphonema parvulum* (Wagner & Waidelich, 1986). The Raman spectra of Chls are composed of three main spectral regions (figure 4l): C-N breathing mode region (1300 to 1420 cm^{-1}), the bands assigned to vinyl groups and the respective bonds with porphyrin groups (1580 to 1630 cm^{-1}), and the bands assigned to in-plane stretches or peripheral substituents such as C-13 keto carbonyl (1650 to 1700 cm^{-1}) (Premvardhan *et al.*, 2010). Chlorophylls can be classified in three groups based on their C-13 keto carbonyl groups: weak hydrogen bonds (giving rise to bands located between 1685 and 1705 cm^{-1}), moderate hydrogen bonds (with bands located between 1670 and 1680 cm^{-1}) and strong hydrogen bonds (with bands located between 1650 and 1665 cm^{-1}) (Mimuro *et al.*, 1990).

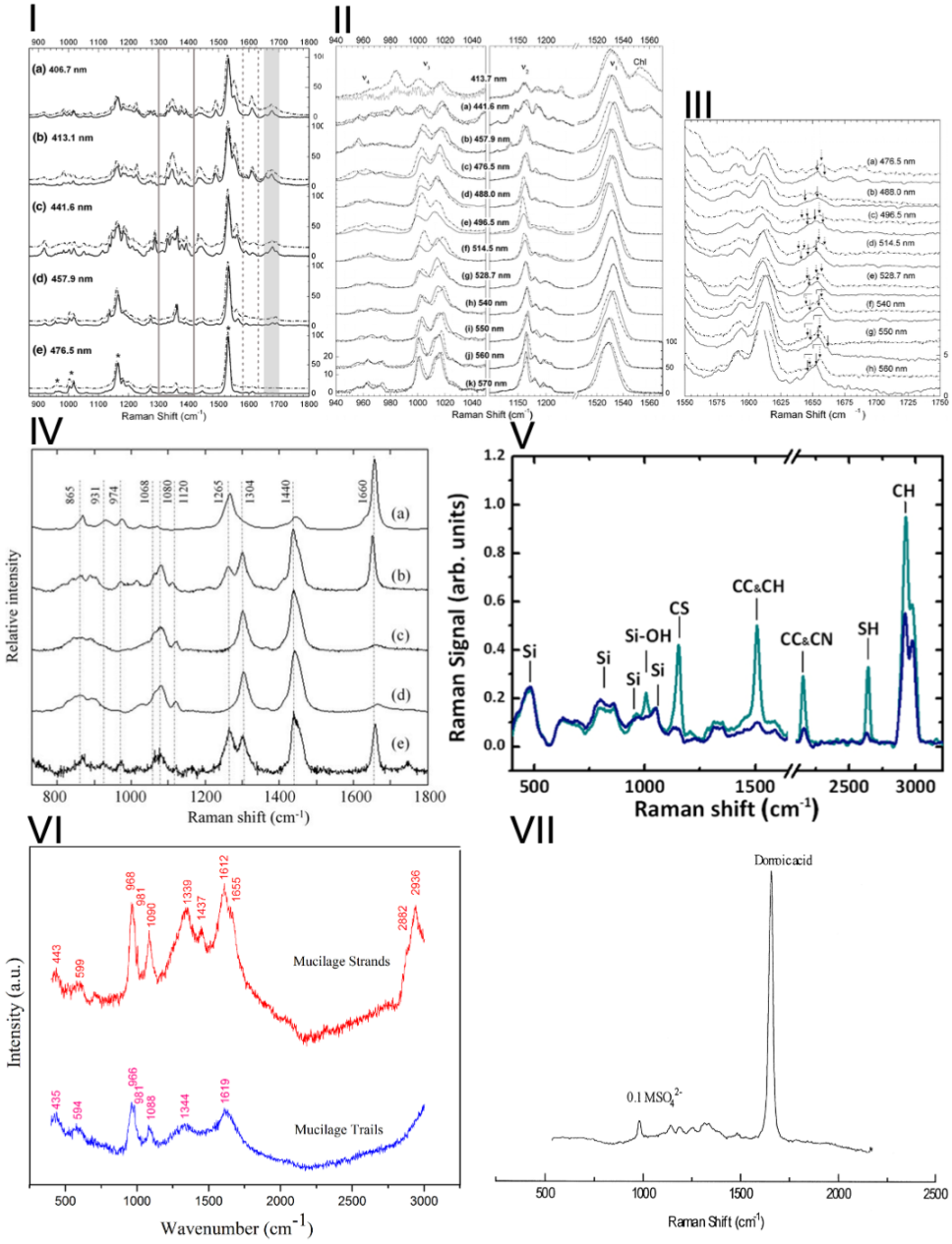


Figure 4 - Raman spectra of the several diatom cell components reviewed. I, II and III- Different Resonant Raman spectral regions of trimeric FCP (solid line) and oligomeric FCP (dash-dot line) on a single scale normalized to the C=C stretch band of Fx (~1530 cm^{-1}) set to 100) for at excitation wavelengths between 406.7 and 570 nm *Adapted from* (Premvardhan *et al.*, 2009; 2010). I - The spectral regions characteristic of chlorophylls are highlighted in grey: C-N breathing mode region (1300 to 1420 cm^{-1}) is marked with a solid squared, the bands assigned to vinyl groups and the respective bonds with porphyrin groups (1580 to 1630 cm^{-1}) is marked with a dashed-dot squared and the modes assigned to in-plane stretches or peripheral substituents such as C-13 keto carbonyl (1650 to 1700 cm^{-1}) is shaded in grey. The main carotenoid bands (υ1, υ2, υ3, υ4) are marked with a "*" in (e). II - Regions υ1 and υ2 of the carotenoid spectra on the right side and regions υ3 and υ4 of carotenoid spectra on the left side. III - Carbonyl stretching region of carotenoid spectra; the peaks and shoulders presumed to arise from carbonyls of different Fx's are indicated with arrows. The two sets of carbonyls from the Fx-red are indicated with horizontal bars for 540-560 excitation wavelengths. IV - (a) Raman spectra of Eicosapentaeonic acid, (b) Palmitoleic acid, (c) Palmitic acid, (d) Mystic acid, and (e) the lipid-body of the diatom. *Adapted from* Meksiarun *et al.* (2015). V - Raman spectra of a single valve of *Coscinodiscus wailesii* over the range 400–3200

cm^{-1} and respective assignments. Blue spectra: mean over 22 spectra. Cyano spectra: mean over 3 spectra where presence of sulfur composites was detected. *Adapted from* De Tommasi *et al.* (2018). VI - Raman spectra of the mucilage trails and mucilage strands, respectively. *Adapted from* Chen *et al. (2019)*. VII - Raman spectrum of domoic acid-producing diatoms of the genus *Pseudo-nitzschia* excited by 251-nm light. *Adapted from* Wu *et al.* (2000).

Regarding carotenoids, the most mentioned in RS studies are Fx, Ddx and Dtx (figure 5). All carotenoids contain a conjugated polyene chain varying in the number of carbon atoms (figures 5a, 5b and 5c). Carotenoids Fx and Ddx have a 5,6 – monoepoxide group (figures 5a and 5b), which is not present in Dtx. Carotenoid Fx contains a unique allenic bond (figure 5a; Kuczynska *et al.*, 2015).

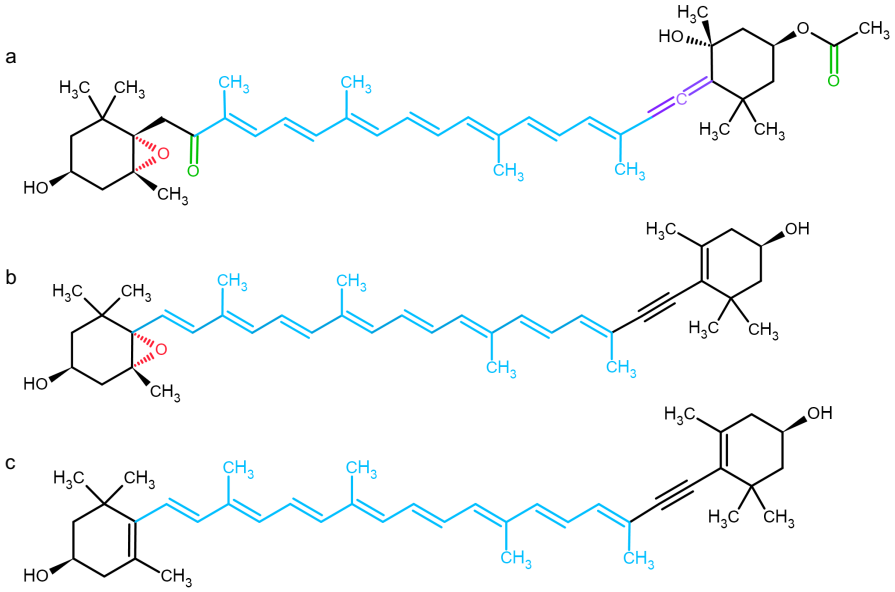


Figure 5 - Molecular structures of Fucoxanthin (a), Diadinoxanthin (b) and Diatoxanthin (c). The conjugated polyene chain, which contributes to υ1 and υ2 regions of Raman spectra, is highlighted in blue. Carbonyl groups of Fucoxanthin (a) are indicated in green. The allenic bond distinguishing Fucoxanthin (a) from the remaining carotenoids is indicated in purple. The 5,6-monoepoxide group present in Fucoxanthin and Diadinoxanthin is highlighted in red (a, b).

Spectral regions characteristic of carotenoids are also shown in figure 4I, namely in the Resonant Raman spectra obtained for the excitation at 476.5 nm: 1) the υ1 region, located between 1500-1570 cm^{-1}, assigned to C=C stretching modes; 2) the υ2 region, located between 1140-1180 cm^{-1}, assigned to C-C stretching modes; 3) the υ3 region, located between 1000-1040 cm^{-1}, assigned to CH_3 in-plain wagging modes; and 4) the υ4 region, located close to 950-980 cm^{-1}, assigned to C-H out-of-plain wagging modes. Alongside the main four regions of carotenoid spectra, the spectral region corresponding to carbonyl groups (around 1700 cm^{-1}) can also be associated with carotenoids (Premvardhan *et al.*, 2009). In carotenoid Raman spectra, the region υ1 is a good marker of the length of the polyene chain, while the region υ4 is for the carotenoids conformation

(Frank *et al.*, 2006; Lutz *et al.*, 1987). The position of the υ1 band is proportional to the length of the polyene conjugated chain (Merlin, 1985), and is evidenced in figure 6.

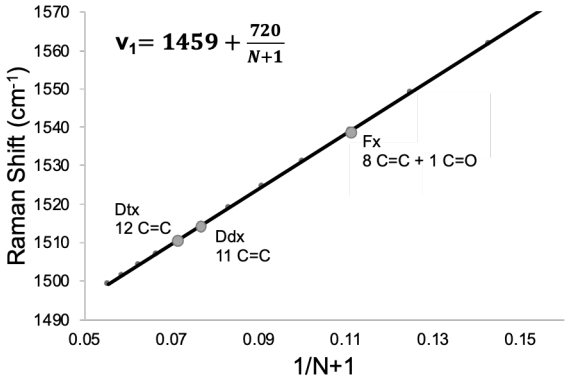


Figure 6 - Relation between the frequency of the carotenoid band (v1) and the number of double bounds in the polyene chain (N), according to the equation derived by Merlin et al. 1965. Fucoxanthin (Fx), Diadinoxanthin (Ddx) and Diatoxanthin (Dtx) are also marked in the graph.

It was originally thought that FCPs stoichiometry was 4:1:1 Fx/Chl *a*/Chl c_2 (Büchel, 2003). However, studies in FCP complexes isolated from the diatom *Cyclotella meneghiniana*, using RRS with other techniques, have evidenced a different ratio. According to the evidence gathered in these studies, the band υ1 of Fx in FCPs is indicative of a polyene chain with about seven double bounds; the low intensity of υ3 and υ4 bands indicates an all-trans and planar conformation in both complexes (Premvardhan *et al.*, 2009, figure 4II and III). Furthermore, it was also shown that differences in all regions of Fx spectra indicate two types of Fx-red in lower energy regions, two types of Fx-blue in higher energy regions and one to two types of Fx-green in intermediate energy regions. Indeed, the wavelengths used for resonant excitation of the various types of Fx, were coincident with the 0-0 and 0-1 of the electroabsorption spectrum (Premvardhan *et al.*, 2009). The two types of Fx-red were detected on both oligomeric and trimeric FCPs when 540 and 550 nm lasers are used as excitation: the presence of a shoulder on the lower energy side of the υ1 band (figure 4IIh; i) and the presence of two carbonyl bands (1647 and 1655 cm^{-1}) that vary in intensity between the two wavelengths (figure 4IIIf; g) were indicative of two types of Fx-red (Premvardhan *et al.*, 2009). The small downshift of the υ1 band and the broadening of the carbonyl band at 514.5 nm (figure 4IIf), have suggested the existence of at least one type of Fx-green (Premvardhan *et al.*, 2009). In trimeric FCP, the carbonyl band 1642 cm^{-1} was associated with one of types of Fx-green (figure 4IIId), whereas the other type of Fx-green was evidenced by the apparent increasing in intensity of the 1645 cm^{-1} band with excitation wavelength of 528.7 nm and 496.5 nm (figure 4IIIc; e; Premvardhan *et al.*, 2009). In

oligomeric FCP, there was also an evidence of two types of Fx-green: one suggested by the shoulder 1647 cm^{-1} at an excitation wavelength of 528.7 nm (figure 4IIIe) and the other one suggested by the increase in full width at half medium of the υ1 band and the decrease of the energy separation of the doublet υ3 at an excitation wavelength at 488 and 496.5 nm (figure 4IId; e; Premvardhan *et al.*, 2009). In oligomeric FCP the lower frequency of the υ1 band (figure 4IIId-g) indicated differences between the molecular conformation of the Fx-green between trimeric and oligomeric FCP (Premvardhan *et al.*, 2009). Also, different frequencies in trimeric FCP together with a lower full width at half medium of the υ1 band with excitation wavelength at 413.7 and 441.6 nm, were found to indicate the presence of an Fx-blue (figure 4IIa; Premvardhan *et al.*, 2009). For Chls, the results are represented in figure 4I. There were two C-13 keto carbonyl bands characteristic of Chl c_2 in Raman spectra of FCP: one with moderate hydrogen bonds (~1675 cm^{-1} in both trimeric and oligomeric FCPs), and another one with very weak hydrogen bonds (~1695 cm^{-1} in trimeric FCP and 1690 cm^{-1} in oligomeric FCP; figure 4Id). Also, there were two bands characteristic of C-N breathing modes in Chl c_2: ~1355 cm^{-1} in oligomeric FCP, ~1360cm^{-1} in trimeric FCP and ~1362 cm^{-1} in both oligomeric and trimeric FCPs (figure 4Id). Altogether, these data pointed out the existence of two types of Chl c_2 differing also between trimeric and oligomeric FCPs. When considering the C-13 keto carbonyl groups in Chl *a* of trimeric FCP, there were three bands to be considered at an excitation wavelength of 406.7 nm: 1654 cm^{-1}, corresponding to strong hydrogen bonds; 1677 cm^{-1}, corresponding to medium hydrogen bonds; and 1688 cm^{-1}, corresponding to weak hydrogen bonds (figure 4Ia; Premvardhan *et al.*, 2010). The broadening and higher intensity of the bands at the incident wavelength of 413.1 nm with higher full width at half maximum indicated the existence of more than one type of Chl *a*: two to three with weak or no hydrogen bonds, three with moderate hydrogen bonds and two to three with strong hydrogen bonds (figure 4Ib). Besides these, the broader and less structured bands between 1650 and 1700 cm^{-1} in oligomeric FCP indicate the existence of other distinguishable types of Chl *a* (figures 4Ia; b; Premvardhan *et al.*, 2010). Furthermore, Chl *a* can be classified into blue (Soret) and red absorbing species depending on the bands that were more enhanced at a wavelength of 406.7 nm in opposition to those that were enhanced at a wavelength of 413.1 and 441.6 nm. The methine-bridge mode of red Chl *a* appears close to 1610 cm^{-1}, at a excitation wavelength of 413.1 and 441.6 nm in both FCPs (figures 4Ib; c), and the blue Chl *a* appears as a shoulder at 1615 cm^{-1}, at 406.7 nm, also in both FCPs (figure 4Ia; Premvardhan *et al.*, 2010).

The remaining carotenoids – Ddx and Dtx – play an important role in the photoprotection of the diatoms from the large light variations caused by caudal changes, tides and their position in the water column (Alexandre *et al.*, 2014). These carotenoids are involved in two photoprotection mechanisms: non-photochemical quenching (NPQ) Chl *a* fluorescence (Grouneva *et al.*, 2008) and the xanthophyll cycle. In the xanthophyll cycle, under high-light (HL) conditions, a de-epoxidation occurs converting Ddx into Dtx (Stransky & Hager, 1970). Diatoxanthin contributes to the dissipation of excess energy and capture of harmful singlet oxygen species (Rüger *et al.*, 2019). Low-light (LL) conditions favour the inverse reaction (Goss *et al.*, 2006; Grouneva *et al.*, 2008).

Variation of Ddx and Dtx with different light conditions was studied in the diatom *Cyclotella meneghiniana* using RS in comparion with High Performance Liquid Chromatography (HPLC) results (Alexandre *et al.*, 2014). For RS recording, cells were concentrated and dropped in a microscope slide. As the chlorophyll content did not change with light conditions in either of the analytical techniques, the chlorophyll band 1555 cm^{-1} was used as a normalization to estimate the carotenoid content. Carotenoid spectra differed the most between LL and HL conditions in the spectral ranges υ1, υ3 and υ4. Under HL, the υ1 region increased in amplitude 1.8 times, at an excitation wavelength of 413.1 nm, and 1.5 times, at an excitation wavelength of 441.6 nm, compared to LL. As the content in carotenoids in the HPLC increased about 1.42 times, the authors concluded the 441.6 nm excitation wavelength was the best for the quantification of carotenoids (due to the similar increase in amplitude). In the HPLC, the increase in carotenoids under HL conditions was due to higher levels of Ddx/Dtx. Since Dtx has a longer polyene conjugated chain than Ddx (Alexandre *et al.*, 2014), increases in content of Dtx relatively to Ddx are accompanied by a red-shift in the frequency of the υ1 band (Merlin, 1985). These red-shifts occur in almost every spectrum under every excitation wavelength registered under HL conditions. In the second part of the study, a normalization to the υ1 region of carotenoid spectra was done to understand the conformation of the carotenoids. Through the analysis of the υ4 region, the authors observed that in cells under HL conditions there was an increase intensity of the 983 cm^{-1} band (compared to LL conditions) probably indicating a twisted Dtx related to a bounded protein. Also, at an excitation wavelength of 496.5 nm, resonant with Ddx, there was a shifting from 980 cm^{-1} to 983 cm^{-1} of the υ4 band and an increase in the band intensity were observed under HL, compared to LL (Alexandre *et al.*, 2014). This was interpreted as an indication of a new, more distorted and complexly protein-bound pool of Ddx under HL conditions. A similar study was carried out using Coherent Anti-

Stokes Raman Scattering (CARS) and RS to investigate carotenoid location and content in the diatoms *Ditylum brightwellii* and *Stephanopyxis turris* (Legesse *et al.*, 2018). Cells were immobilized with poly-L-Lysine onto CaF_2 slides. In general, CARS and RS also detected an increase in carotenoids concentration between LL and HL conditions in both diatoms.

Raman spectroscopy was also used to characterize marennine and marennine-like pigments inside alive diatom specimens and extracted in solution (Gastineau *et al.*, 2012; Gastineau *et al.*, 2016). Marennine is a blue pigment characteristic of the diatom *Haslea ostrearia* and it is observed also in oysters feeding on this species (Gastineau *et al.*, 2014). Other diatoms of the genus *Haslea* have similar blue pigments, though different in structure. Raman detected differences in marennine from *Haslea ostrearia* relative to other marennine-like pigments within the wavenumber range 1240-1420 cm^{-1} (Gastineau *et al.*, 2012; Gastineau *et al.*, 2016). The structure and biosynthesis of marennine are still unknown. Nevertheless, it has been hypothesized that this pigment is greatly produced depending on light conditions, nutrient deficiencies or salt-inducted stress (Gastineau *et al.*, 2014). This pigment can be used as blue dye in the food and cosmetic industries and has anti-inflammatory properties (Carlson et al, 1985; Collin et al., 2001; Bretéché et al., 2002).

The bands associated with the vibration of pigments, species tested, spectral profile according with the excitation wavelength and the respective mode assignments made in the studies above are presented in Table I of the Appendix I. It is important to understand that pigment contents can change not only with light conditions, but also with other environmental or physiological conditions (Kuczynska *et al.*, 2015). More studies are thus required to understand variation in RS bands of diatom pigments in relation to other factors. This will allow developing faster and easier tools for environmental assessment and diagnostic tests. Additionally, some RS studies point out that the pigment bands may differ among *taxa* (Abbas *et al.*, 2011; Wood *et al.*, 2005) and growth phases (Rüger *et al.*, 2016). If so, RS could be used to compensate the inherent constraints in taxonomic identification of diatoms and to assess the influence of stress conditions on their growth phases.

2.2. Lipids

In diatoms, lipids are stored in specific organelles called lipid bodies (Goold *et al.*, 2015; Liu & Benning, 2013; Zienkiewicz *et al.*, 2016). Lipid bodies occupy a large portion of the

cell volume (Supriya *et al.*, 2012), variable among species (Maeda *et al.*, 2017). Enhanced lipid production is usually associated with exposure to environmental stress, such as pH variations, temperature, light intensity, nitrogen, carbon, silica, phosphorous, iron and salinity (Supriya *et al.*, 2012). This occurs because, under stress conditions, lipids act like secondary sources of energy necessary to maintain membrane functions and cell-signalling pathways (Hu *et al.*, 2008).

A RS study of lipid pool variation in relation to different nitrogen sources, *i.e.* sodium nitrate and urea, was performed in diatoms of the genus *Nitzschia*. Measurements were made on cells placed directly under the Raman microscope. After seven days of experiments, the authors detected changes in bands assigned to lipids associated with changes in the lipid pool due to changes in the nitrogen source (Supriya *et al.*, 2012). Similar studies, conjugating RS with chemometric techniques, were performed directly on individuals of the species *Thalassiosira pseudonana* placed on a petri dish with a quartz bottom (Meksiarun *et al*., 2014, 2015) and in purified fatty acids extracts (Meksiarun *et al*., 2015). Raman spectra obtained in this study are represented in figure 4IV. The objective was to determine changes in lipid composition and concentration under different stress conditions. The authors used a ratiometric method based on the ratio of two band itensities: the band 1660 cm^{-1} which corresponded to the number of C=C present in lipids and the band 1440 cm^{-1} assigned to CH_2 bending modes. They discovered that diatom lipids are mostly recorded between palmitic and linoleic acid, suggesting that many of their fatty acid chains contain one to two double bonds. Since diatoms have triglycerides with varied compositions in their fatty acid chains, a conventional ratiometric method would hardly reflect all this variety. To better characterize these molecules, the authors performed a PCA comparing spectral data from pure fatty acid chains and diatom samples. The spectra recorded from diatom samples were mostly related to the spectra of myristic acid (14-carbon, 0-double bond), palmitic acid (16-carbon, 0-double bond), palmitoleic acid (16-carbon, 1-double bond), and eicosapentaenoic acid (20-carbon, 5-double bond). Data from the two other fatty acids tested (linoleic and oleic acids) were not related to the lipid spectra of the diatom. Additionally, Raman ordinary least squares (Raman-OLS) method was used to determine lipid quantities under seven different conditions: control, iron abundance and shortage, carbon dioxide abundance and shortage, and nitrogen abundance and shortage (Meksiarun *et al*., 2015). Lipids were divided in three groups based on the ratio of saturated (myristic and palmitic acids) to unsaturated fatty acid chains (palmitoleic and eicosapentaenoic acids): saturated fatty-acid dominated (SFD), unsaturated fatty-acid

dominated (UFD) and saturated-unsaturated fatty acid transformation (SUFT). The results obtained are shown in Table 1, reflecting changes in lipid groups of diatom populations exposed to the different experimental conditions relative to the control. There was an increase in SFD in all conditions tested. A decrease in UFD was also observed in all conditions tested, although less remarkable in the case of iron shortage. Regarding SUFT, iron shortage induced a SUFT reduction, while iron abundance, carbon dioxide abundance and shortage and nitrogen starvation elicited a SUFT increase (Meksiarun *et al.*, 2015). In another study from the same authors, fatty acid production was found to increase with the increase in carbon dioxide (Meksiarun *et al.*, 2014).

Table 1 - Changes in lipid groups of diatom populations exposed to different experimental conditions in relation to the control indicated by Raman Ordinary Partial Least Squares analysis (adapted (adapted from: Meksiarun et al., 2015). Saturated fatty-acid dominance, SFD; unsaturated fatty-acid dominance, UFD; saturated-unsaturated fatty acid transformation, UFD; experimental stress conditions: shortage (-) and abundance (+) of Fe, CO2 and N; ↑ - increasing; ↓ - decreasing; = - no change.

Conditions	**SFD**	**SUFT**	**UFD**
-Fe	↑	↓	↓
+Fe	↑	↑	↓
-CO_2	↑	↑	↓
+CO_2	↑	↑	↓
-N	↑	↑	↓
+N	↑	=	↓

Although these RS studies were mainly performed to evaluate the optimal conditions for the mass production of unsaturated lipids in the biofuel industry, this knowledge could be used in environmental diagnosis where alterations in lipid bands could be related to stressful aquatic conditions. The bands associated with molecular vibrations of lipids, species tested, spectral profile according with the excitation wavelength and the respective mode assignments in the studies above are presented in Table I of the Appendix I.

2.3. Frustule

The most distinctive feature in diatoms is the frustule. The frustule is a siliceous cell wall that protects the cell against all kinds of stress factors such as mechanical pressure, xenobiotics, environmental changes and grazing (Townley, 2011). Furthermore, it has photonic crystal properties (Fuhrmann *et al.*, 2004). Such properties can play an

important role in light manipulation, since diatoms are often in environments where light is not easily available to perform photosynthesis (De Tommasi, 2016). As in lipids, RS studies concerning the constitution and conformation of the frustule are few but the technique was proven to be very effective for that purpose. Studies involving the RS of the frustule usually require the oxidation of the organic material. As expected, the results showed a general prevalence of bands assigned to siliceous materials in all the studies (Arasuna & Okuno, 2018; Biswas *et al.*, 2018; De Tommasi, 2016; Kammer *et al.*, 2010; Yuan *et al.*, 2004). However, bands attributed to other unknown inorganic compounds were also observed. In addition, bands related to vibrations of some trace organic molecules appeared in the diatoms *Stephanopyxis turris* (Kammer *et al.*, 2010) and *Coscinodiscus wailesii* (Figure 4V, De Tommasi *et al*., 2018). These can possibly be due to the presence of impurities that absorb UV radiation and, consequently, protect the cell (De Tommasi *et al.*, 2018). Interestingly, spectral data from the frustule of the centric diatom *Coscinodiscus wailesii* showed some additional bands assigned to sulphur bonds (De Tommasi, 2016; De Tommasi *et al.*, 2018). These sulphur compounds might be related to the important role of diatoms in the global sulphur cycle. Diatoms produce dimethyl sulphide (DMS) that is emitted to the atmosphere and contributes to cloud condensation and solar radiation scattering. This variation in radiation balances the light reaching the earth, thus influencing phytoplankton growth (Simó, 2001). The sulphur compounds can also act as precursors for the biosynthesis of silica (De Tommasi, 2016; De Tommasi *et al.*, 2018). Details on the bands associated with molecular vibrations of frustule, species tested, spectral profile according with the excitation wavelength and the respective mode assignments are listed in Table I of the Appendix I. From an environmental point of view, frustule silicification is affected by various abiotic factors, such as salinity, heavy metals, temperature, pH, nutrient conditions and light (reviewed in: Su *et al.*, 2018). Some shifts in frequency and differences in the intensity of frustule bands were found in two samples containing different diatom genera (Yuan *et al.*, 2004), which can be elicited by different processes of mineralization. This is another indication that RS is a sound method to overcome limitations imposed by the taxonomic identification of diatoms in traditional ecological assessment. Though, further studies involving its application to explore the effects of abiotic factors in frustules need to be conducted.

2.4. Other components

Apart from the molecules already described, RS was also used to understand the constitution of EPS and the detection of domoic acid (DA) in diatoms. In nature, diatoms are mostly found within biofilms in conjunction with other organisms such as bacteria, protozoa, fungi and other algae. In these biofilms, diatoms are the most abundant group (Battin *et al.*, 2016; Sabater *et al.*, 2007; Salta *et al.*, 2013). Furthermore, they produce the major amount of EPS, which makes up from 50% to 90% of the inorganic mass in biofilms (Nielsen *et al.*, 1997). EPS contributes to cell adhesion, locomotion and colony formation. Reports about the use of RS to investigate EPS characteristics can also be found in the literature. A study about the EPS of *Nitzschia palea*, (Laviale *et al.*, 2019) showed bands mainly assigned to proteins, lipids and carbohydrates. In another study, mucilage bands (released when the diatom is motionless) and mucilage strands (released when the diatom is gliding onto a surface) of the diatom *Navicula sp.* were analysed (Figure 4VI; Chen *et al.*, 2019). As in *N. palea*, bands assigned to the aminoacids phenylalanine (at 594 and 599 cm^{-1}) and tyrosine (at 1612 and 1619 cm^{-1}), as well as to polysaccharides (at 1088 and 1090 cm^{-1}), appeared on both types of mucilage. Additional bands were detected in mucilage strands at wavenumbers 1437 cm^{-1}, 1655 cm^{-1}, 2882 cm^{-1} and 2936 cm^{-1}, corresponding probably to carbohydrate bond vibrations (Chen *et al.*, 2019) (Table I of the Appendix I).

Domoic acid was another substance investigated through Raman spectroscopy. This is a toxin produced by marine diatoms from the genus *Pseudonitzschia.* If ingested by humans (indirectly through contamination of the food-chain), DA causes a pathology named Amnesic shellfish poisoning (ASP) (Todd, 1993). Seafood contamination by DA is usually detected by chemical separation from other interfering substances. This method is time-consuming and, together with bioassays, requires animal sacrifice. Therefore, a faster and non-invasive method, as RS, was needed to detect DA and protect public health. Raman spectra of diatoms with DA showed a strong resonance Raman enhancement at 251 nm of the 1652 cm^{-1} vibrational band, which was assigned to the coupling of the symmetric C=C mode of this compound (Figure 4VII, Wu *et al.*, 2000). In fact, the band 1652 cm^{-1} is more intense in the spectra of toxic algae than in the spectra of non-toxic algae. Moreover, a formula was proposed to determine DA concentration, relating the cross-section, band intensity and frequency of DA with the same parameters and a known concentration of an internal standard (Wu et al., 2000).

The above results highlight the potential of RS to be applied in future studies of EPS in field environmental diagnosis. The EPS constitutes a protective barrier slowing the diffusion of diverse toxic compounds such as metals or even antibiotics (Stewart, 2003).

Spectral data may thus be useful to detect man-made toxicants possibly present in this matrix and better understand their physiological impact on diatoms.

3. Diatoms as an innovative substrate for SERS

As mentioned above, SERS is a technique in which Raman signals are enhanced by coating metals (generally nanoparticles) onto a surface (Baena & Lendl, 2004). This technique can be applied to a flat surface. Better enhancement factors, however, are obtained when it is applied onto a patterned surface. By coating patterned surfaces with nanoparticles (NPs), some of them might be placed inside or near dielectric microcavities and form hybrid photonic-plasmonic modes (Barth *et al.*, 2010; Hu *et al.*, 2011; Schmidt *et al.*, 2012; White *et al.*, 2007; Xu *et al.*, 2012). These modes increase the quality of the local electric field, when coupling of the guided magnetic resonance (GMR) with the local surface plasmons (LSPs) takes place, enhancing RS signals (Cheng & Scherer, 1995). Synthetic patterned surfaces are commercially available. Nevertheless, the processes required to obtain such intricate nanostructures are very expensive (Yang *et al.*, 2014). Diatoms offer a more cost-effective and naturally available alternative, owing to their naturally ornamented frustules (Kong *et al.*, 2016).

Three alternatives are available to coat the frustules of diatoms with NPs. The first is coating by physical processes, such as deposition or immersion (e.g. Chamuah *et al.*, 2017). The second is based on chemical methods. The substrate is immersed in a cleaning solution (RCA clean, 1:1:5 $H_2O_2/NH_4OH/H_2O$) to remove organic contamination and particles, and the sample is rinsed in water and methanol to create hydroxyl groups on the diatoms and the substrate. Diatoms are then immersed in aminopropyltriethoxysilane (APTES) and this compound bounds to the hydroxyl groups on the substrate. The sample is then immersed in a colloidal solution of NPs (Ren *et al.*, 2014). The third alternative is by *in situ* growth of NPs. For this, diatoms are immersed in $SnCl_2$ and HCl to deposit Sn^{2+} onto the frustule. These diatoms are then immersed in $AgNO_3$, promoting the growth of Ag seeds onto the frustule. Finally, the diatoms are immersed in appropriate culture media (Kong *et al.*, 2016).

The use of frustules as substrate has many applications in the detection of all kinds of biomolecules in a variety of fields such as medicine, food safety or environmental quality. For example, in medicine it can be used in immunoassays for the detection of interleukin-8, an inflammatory cytokine playing a role in breast cancer (Kamińska *et al.*, 2017). This method is recognized as offering better results than the traditional Enzyme-Linked

Immunosorbent Assays. In food industry, coated frustules can be used to detect melamine in food, which is illegally added to increase the protein content. Here, SERS detection shows a 3-fold enhancement, compared to common RS. In environmental quality, it can be used for example to detect xylene, a water and air pollutant, offering a simpler and faster alternative to detection by the conventional gas chromatography method (Kong *et al.*, 2016).

4. Toxicity assessment employing Raman spectroscopy

To our knowledge, RS was only used in two toxicity assays with diatoms. In the first, a technique called high-throughput screening Raman spectroscopy (HTS-RS) with automated location algorithms was employed in a chronic toxicity assay to detect the effects of dithiothreitol (DTT) in the diatom *Phaeodactylum tricornutum*. Dithiothreitol inhibits the xanthophyll cycle, which is a mechanism of photoprotection (Rüger *et al.*, 2019). Hence, the assay comprised groups with different combinations of DTT and light conditions: control (adequate light conditions with no DTT treatment), HL^+ (high-light conditions with DTT treatment) and HL^- (high-light conditions with no DTT treatment). Through PLS-LDA, spectra from the different experimental conditions were compared to reference spectra and discriminated. One and a half hours after inoculation, HL^+ was already separated from the other conditions (Rüger *et al.*, 2019). In this group, exposed cells were found to change the most over time, in comparison to the control. According to the authors, this occurred because inhibition of the xanthophyll cycle caused by DTT affected the photoprotection of the cells against the high-light conditions and harmful singlet oxygen species. In the beginning of the experiment, HL^- cells showed a response pattern similar to the control. However, over time they tended to shift towards higher linear discriminant scores probably due to Dtx enhancement elicited by high-light conditions (Rüger *et al.*, 2019).

In another recent study, SERS and RI were used to investigate the incorporation of gold nanoparticles (AuNPs) with different diameters into the diatom *Stephanopyxis turris* (Pytlik *et al.*, 2019). This work is especially relevant because information about the hazardous effects of metal nanoparticles in aquatic organisms is still scarce, despite their increasing use in various technological and medical fields, and potential for environmental contamination. For diatoms in particular, it is not yet clear how nanoparticles can affect the cells. Moreover, the uptake mechanisms are not fully understood and their interpretation is mostly based on information obtained from other

organisms (Navarro *et al.*, 2008). For the purpose of the experiment, RI of the most intense bands assigned to pigments were used to estimate the shape of the cell, whereas SERS spectra of AuNPs were used to locate these compounds within the cells. Briefly, AuNPs with dimensions above 50 nm were most frequently located inside the cell, compared to smaller nanoparticles, suggesting a size-dependent mechanism requiring further investigation (Pytlik *et al.*, 2019).

In the toxicological works above, the cell pigments were major contributors to the Raman spectra detected. Raman parameters (*i.e.* laser wavelength, time acquisition and objective) can also be set to avoid measurement of pigments and their possible masking effects. This allows recording spectra and molecular fingerprints from other cellular components, bring further understanding on other effects of environmental toxicants on diatoms. Compared to the conventional endpoints, RS-based methods are usually a more expeditious technique to couple with exposure bioassays, providing a wealth of data useful to characterize diatom responses to environmental stress. Raman spectroscopy has also been employed to analyse other types of algae in order to characterize their growth phases under laboratory simulated environmental conditions (He et al., 2018). Based on a review of a large variety of algae (including diatoms) studies employing mostly Fourier-transform infrared spectroscopy, (Akkas & Severcan, 2012) suggested previously that vibrational spectroscopy analysis could be used in the screening and monitoring of aquatic systems. Nevertheless, there is a lack of field studies concerning the application of RS on diatoms in diagnose tests of environmental quality. To develop a protocol for a diagnose test with RS on diatoms it is necessary to define adequate apparatus parameters. Firstly, it is important to choose the suitable laser wavelength, which depends on the cell component under analysis. Figure 7 presents laser wavelengths in relation to cell components, retrieved from the studies reviewed in this work. It is noticeable that RS studies about diatom pigments and frustule have mainly used blue and green lasers. Violet and infrared lasers have also been employed in the study of pigments, as well as red and infrared lasers in frustules. Lipids have been evaluated using only infrared lasers, while DA Raman signals have been enhanced only with UV laser. The molecular components of EPS have been investigated using UV and green lasers.

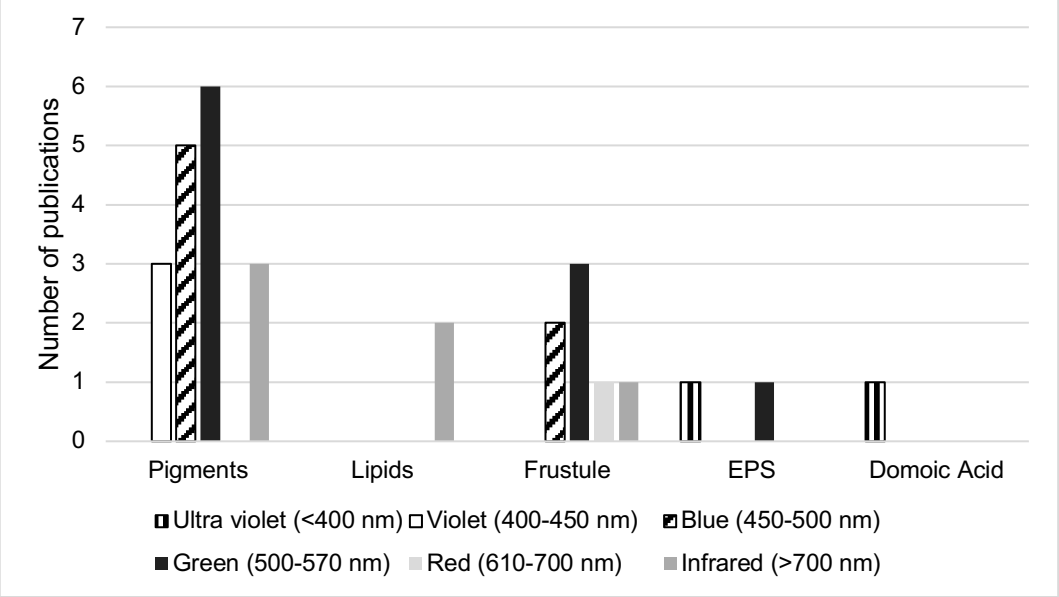


Figure 7 - Laser wavelengths used to investigate different diatom components (pigments, lipids, frustule, EPS and Domoic Acid) by Raman Spectroscopy. Data was retrieved from the literature reviewed in this work. In articles employing more than one type of laser in different light ranges, each light range was considered as a distinct study.

Additionally, due to a resonance phenomenon, different laser wavelengths enhance the measurement of different types of pigments, according to their maximum absorption (McCreery, 2005). The laser wavelengths enhancing each type of pigment are described in Table 2. Apart from laser wavelength, the selection of the microscope objective, acquisition time, laser power and spectral range are also important. The selection of the objective depends on the detail and size of the component to be analysed. Other relevant parameters are the time of acquisition and the laser power. Larger acquisition times and higher laser power can cause alterations within the cell such as pigment rearrangements (Barletta *et al.*, 2015). Hence, preliminary experiments should be conducted to confirm the acquisition conditions providing the most adequate measurements in relation to the testing hypotheses. Regarding the spectral range, according to the studies reviewed (Table I of the Appendix I), pigment bands appear from 340 to ~1700 cm^{-1}, lipid bands from ~865 to 1800 cm^{-1}, frustule bands from 373 to 3000 cm^{-1}, EPS bands from ~594 to 3000 cm^{-1}, and the DA band corresponds to 1652 cm^{-1}.

Table 2 - Resonant Raman spectroscopy (RRS) laser wavelength for each type of diatom pigment.

Pigment type	Pigment subtype	RRS wavelength (nm)	Reference
Chlorophyll a		406.7, 413.1, 441.6	(Premvardhan *et al.*, 2010; Wagner & Waidelich, 1986)
Chlorophyll c	Chlorophyll c_1	441.6	(Wagner & Waidelich, 1986)

	Chlorophyll c_2	441.6, 457.9, 476.5	(Premvardhan *et al.*, 2010; Wagner & Waidelich, 1986)
Fucoxanthin	Fx Red	496.5-570	(Premvardhan *et al.*, 2009)
	Fx Blue	413.7 and 446.1	(Premvardhan *et al.*, 2009)
	Fx Green	488 and 514	(Premvardhan *et al.*, 2009)
Diadinoxanthin		~490 nm	(Alexandre *et al.*, 2014)
Diatoxanthin		~510-515 nm	(Alexandre *et al.*, 2014)

The selection of the correct Raman methodology, and subsequent analysis of data, are also important. RI can be useful to detect, for example, toxic contaminants inside the cells as was done for gold nanoparticles (Pytlik *et al.*, 2019), or to detect compounds in EPS before they reach the cells. Statistical methods such as PCA and PLS-DA are useful to analyze Raman data in diatoms under different environmental conditions (Meksiarun *et al.*, 2015; Rüger *et al.*, 2019). Although RS is a good technique to detect physiological and molecular changes in diatoms in relation to some environmental stress, RS bands of frustule and pigments can vary also depending on the diatom genera (Abbas *et al.*, 2011; Wood *et al.*, 2005; Yuan *et al.*, 2004). Thus, this hypothesis should also be investigated in order to improve some of the drawbacks of traditional taxonomic identification currently used for environmental diagnosis.

5. Conclusion and future perspectives

In this work we reviewed RS studies concerning the structure, location and conformation of diatoms and their cell components. Over the last decade great progress has been made about the application of RS to diatom research. Diatoms have high biodiversity, give a significant contribution to aquatic primary production and are sensitive to nutrients as well as inorganic and organic contaminants. Owing to this, they are recommended for assessing the ecological status of water bodies worldwide. Nevertheless, the need for their taxonomic identification, which is time consuming and requires relevant expertise, is limiting their use in the monitoring and risk assessment of lakes, rivers and streams. Raman spectroscopy is a faster and easier technique, with a great potential to overcome these constraints and widen the use of diatoms in environmental diagnosis. The studies reviewed in this work provide a solid foundation supporting the application of RS in diatoms for environmental research. The information gathered about wavelengths,

wavenumbers and parameter setting in relation to cell components provides a basis allowing investigating the effects of toxicants at individual, population and community levels. The technique should be used, among others, to obtain molecular fingerprints for species and other diatom taxa, as well as for aquatic ecosystems with similar or different characteristics. The samples can be directly used to record Raman spectra, favouring its rapid and straightforward application in both laboratory and field studies. The spectra obtained can be further related to exposure or environmental stress (*i.e.* chemical contamination, nutrient levels, water physico-chemical properties). Once this relationship is established, RS can then be solely used in routine testing of aquatic ecosystems. On the other hand, toxicological laboratory studies based on RS will enlighten our understanding and knowledge about diatom responses to toxicants, helping interpretation of field results. Investigation focusing on RS of different cell components, under different experimental conditions, will also bring crucial physiological data. Namely about pigments and their production systems, their lipids, the extracellular polymeric substances segregated and their extrinsic embedded substances. This constitutes an added value to detect sub-lethal effects elicited by contaminants and evaluate the health status of cells. It can also lead to identification of sensitive early-warning RS biomarkers of exposure and effect. These are particularly useful for routine monitoring and the follow-up of mitigation actions established for the recovering of water bodies exhibiting poor chemical or ecological status. Studies using RS in diatoms under carbon dioxide abundance also open new perspectives to climate change investigation, a topic in much need for urgent research. Actually, the impacts of water acidification on diatoms, for instance, have yet to be further investigated and understood. Finally, RS data are amenable to analysis by both sensitive and robust chemometrics methods, providing significant testing results (including sensitivity analysis) for decision-making towards mitigation and management actions in the target ecosystems. Overall, a myriad of possibilities and experiments are opened to exploration by Raman spectroscopy to enhance its application to the environmental diagnostic testing of water bodies.

Acknowledgements

The authors would like to thank the EU and FCT/UEFISCDI/FORMAS for funding, in the frame of the collaborative international consortium REWATER, financed under the ERA-NET Cofund WaterWorks2015 (Water JPI). This research was also supported by

national funds through FCT (Portuguese Foundation for the Science and Technology) within the scope of UIDB/04423/2020 and UIDP/04423/2020.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abbas, A., Josefson, M., & Abrahamsson, K. (2011). Characterization and mapping of carotenoids in the algae Dunaliella and Phaeodactylum using Raman and target orthogonal partial least squares. *Chemometrics and Intelligent Laboratory Systems, 107*(1), 174-177.

Akkas, S. B., & Severcan, F. (2012). Diagnosis and Screening of Aquatic Environments by Vibrational Spectroscopy. *Vibrational Spectroscopy in Diagnosis and Screening, 6*, 321.

Alexandre, M. T., Gundermann, K., Pascal, A. A., van Grondelle, R., Buchel, C., & Robert, B. (2014). Probing the carotenoid content of intact Cyclotella cells by resonance Raman spectroscopy. *Photosynthesis research, 119*(3), 273-281.

Almeida, S. F. P., & Gil, M. C. P. (2001). dÉcologie des diatomées d'eau douce de la région centrale du Portugal. *Cryptogamie Algologie, 22*(1), 109-126.

Arasuna, A., & Okuno, M. (2018). Structural change of the frustule of diatom by thermal treatment. *Geoscience Letters, 5*(1), 1.

Baena, J. R., & Lendl, B. (2004). Raman spectroscopy in chemical bioanalysis. *Current opinion in chemical biology, 8*(5), 534-539.

Barletta, R. E., Krause, J. W., Goodie, T., & El Sabae, H. (2015). The direct measurement of intracellular pigments in phytoplankton using resonance Raman spectroscopy. *Marine Chemistry, 176*, 164-173.

Barth, M., Schietinger, S., Fischer, S., Becker, J., Nusse, N., Aichele, T., . . . Benson, O. (2010). Nanoassembled plasmonic-photonic hybrid cavity for tailored light-matter coupling. *Nano letters, 10*(3), 891-895.

Battin, T. J., Besemer, K., Bengtsson, M. M., Romani, A. M., & Packmann, A. I. (2016). The ecology and biogeochemistry of stream biofilms. *Nature Reviews Microbiology, 14*(4), 251.

Biswas, R. K., Khan, P., Mukherjee, S., Mukhopadhyay, A. K., Ghosh, J., & Muraleedharan, K. (2018). Study of short range structure of amorphous Silica from PDF using Ag radiation in laboratory XRD system, RAMAN and NEXAFS. *Journal of Non-Crystalline Solids, 488*, 1-9.

Blanco, S., Ector, L., & Bécares, E. (2004). Epiphytic diatoms as water quality indicators in Spanish shallow lakes. *Vie et Milieu, 54*(2-3), 71-80.

Büchel, C. (2003). Fucoxanthin-chlorophyll proteins in diatoms: 18 and 19 kDa subunits assemble into different oligomeric states. *Biochemistry, 42*(44), 13027-13034.

Byrne, H. J., Knief, P., Keating, M. E., & Bonnier, F. (2016). Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chemical Society Reviews, 45*(7), 1865-1878.

CEMAGREF, M. (1982). Etude des méthodes biologiques d'appréciation quantitative de la qualité des eaux. *Rapport Cemagref QE Lyon-AF Bassin Rhône Méditerranée Corse*.

Chamuah, N., Chetia, L., Zahan, N., Dutta, S., Ahmed, G. A., & Nath, P. (2017). A naturally occurring diatom frustule as a SERS substrate for the detection and quantification of chemicals. *Journal of Physics D: Applied Physics, 50*(17), 175103.

Chen, L., Weng, D., Du, C., Wang, J., & Cao, S. (2019). Contribution of frustules and mucilage trails to the mobility of diatom Navicula sp. *Scientific reports, 9*(1), 1-12.

Cheng, C.-C., & Scherer, A. (1995). Fabrication of photonic band-gap crystals. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena, 13*(6), 2696-2700.

Coste, M., Boutry, S., Tison-Rosebery, J., & Delmas, F. (2009). Improvements of the Biological Diatom Index (BDI): Description and efficiency of the new version (BDI-2006). *Ecological indicators, 9*(4), 621-650.

Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy, (2000).

De Tommasi, E. (2016). Light manipulation by single cells: the case of diatoms. *Journal of Spectroscopy, 2016*.

De Tommasi, E., Congestri, R., Dardano, P., De Luca, A. C., Managò, S., Rea, I., & De Stefano, M. (2018). UV-shielding and wavelength conversion by centric diatom nanopatterned frustules. *Scientific reports, 8*(1), 1-14.

Desrosiers, C., Leflaive, J., Eulin, A., & Ten-Hage, L. (2013). Bioindicators in marine waters: benthic diatoms as a tool to assess water quality from eutrophic to oligotrophic coastal ecosystems. *Ecological indicators, 32*, 25-34.

Dixit, S. S., Smol, J. P., Charles, D. F., Hughes, R. M., Paulsen, S. G., & Collins, G. B. (1999). Assessing water quality changes in the lakes of the northeastern United States using sediment diatoms. *Canadian Journal of Fisheries and Aquatic Sciences, 56*(1), 131-152.

Frank, H. A., Young, A., Britton, G., & Cogdell, R. J. (2006). *The photochemistry of carotenoids* (Vol. 8): Springer Science & Business Media.

Fuhrmann, T., Landwehr, S., El Rharbi-Kucki, M., & Sumper, M. (2004). Diatoms as living photonic crystals. *Applied Physics B, 78*(3), 257-260.

Gastineau, R., Davidovich, N. A., Bardeau, J.-F., Caruso, A., Leignel, V., Hardivillier, Y., . . . Gaudin, P. (2012). Haslea karadagensis (Bacillariophyta): a second blue diatom, recorded from the Black Sea and producing a novel blue pigment. *European Journal of Phycology, 47*(4), 469-479.

Gastineau, R., Hansen, G., Davidovich, N. A., Davidovich, O., Bardeau, J.-F., Kaczmarska, I., . . . Jacquette, B. (2016). A new blue-pigmented hasleoid diatom, Haslea provincialis, from the Mediterranean Sea. *European Journal of Phycology, 51*(2), 156-170.

Gastineau, R., Turcotte, F., Pouvreau, J.-B., Morançais, M., Fleurence, J., Windarto, E., . . . Babin, M. (2014). Marennine, promising blue pigments from a widespread Haslea diatom species complex. *Marine drugs, 12*(6), 3161-3189.

Germain, H. (1981). *Flore Des Diatomees: Diatomophycees: Eaux Douces Et Saumatres Du Massif Armoricain Et Des Contrees Voisines D'Europe Occidentale*: Societe Nouvelle des Editions Boubee.

Goold, H., Beisson, F., Peltier, G., & Li-Beisson, Y. (2015). Microalgal lipid droplets: composition, diversity, biogenesis and functions. *Plant cell reports, 34*(4), 545-555.

Goss, R., Ann Pinto, E., Wilhelm, C., & Richter, M. (2006). The importance of a highly active and DeltapH-regulated diatoxanthin epoxidase for the regulation of the PS II antenna function in diadinoxanthin cycle containing algae. *Journal of Plant Physiology, 163*(10), 1008-1021. doi:10.1016/j.jplph.2005.09.008

Grouneva, I., Jakob, T., Wilhelm, C., & Goss, R. (2008). A new multicomponent NPQ mechanism in the diatom Cyclotella meneghiniana. *Plant Cell Physiol, 49*(8), 1217-1225.

Guglielmi, G., Lavaud, J., Rousseau, B., Etienne, A. L., Houmard, J., & Ruban, A. V. (2005). The light-harvesting antenna of the diatom Phaeodactylum tricornutum: Evidence for a diadinoxanthin-binding subcomplex. *The FEBS journal, 272*(17), 4339-4348.

Guiry, M. D. (2012). How many species of algae are there? *Journal of phycology, 48*(5), 1057-1063.

Heraud, P., Wood, B. R., Beardall, J., & McNaughton, D. (2007). Probing the Influence of the Environment on Microalgae Using Infrared and Raman Spectroscopy. In *New Approaches in Biomedical Spectroscopy* (Vol. 963, pp. 85-106): American Chemical Society.

Hu, M., Fattal, D., Li, J., Li, X., Li, Z., & Williams, R. S. (2011). Optical properties of sub-wavelength dielectric gratings and their application for surface-enhanced Raman scattering. *Applied Physics A, 105*(2), 261-266.

Hu, Q., Sommerfeld, M., Jarvis, E., Ghirardi, M., Posewitz, M., Seibert, M., & Darzins, A. (2008). Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. *The plant journal, 54*(4), 621-639.

Jeffrey, S. W., & Vesk, M. (2005). Introduction to marine phytoplankton and their pigments signatures. *Phytoplankton Pigments in Oceanography-Guidelines to Modern Methods*, 37-84.

Kamińska, A., Sprynskyy, M., Winkler, K., & Szymborski, T. (2017). Ultrasensitive SERS immunoassay based on diatom biosilica for detection of interleukins in blood plasma. *Analytical and bioanalytical chemistry, 409*(27), 6337-6347.

Kammer, M., Hedrich, R., Ehrlich, H., Popp, J., Brunner, E., & Krafft, C. (2010). Spatially resolved determination of the structure and composition of diatom cell walls by Raman and FTIR imaging. *Analytical and bioanalytical chemistry, 398*(1), 509-517.

Kong, X., Squire, K., Li, E., LeDuff, P., Rorrer, G. L., Tang, S., . . . Wang, A. X. (2016). Chemical and biological sensing using diatom photonic crystal biosilica with in-situ growth plasmonic nanoparticles. *IEEE transactions on nanobioscience, 15*(8), 828-834.

Kuczynska, P., Jemiola-Rzeminska, M., & Strzalka, K. (2015). Photosynthetic pigments in diatoms. *Marine drugs, 13*(9), 5847-5881.

Lavaud, J., Rousseau, B., van Gorkom, H. J., & Etienne, A. L. (2002). Influence of the diadinoxanthin pool size on photoprotection in the marine planktonic diatom Phaeodactylum tricornutum. *Plant Physiol, 129*(3), 1398-1406.

Laviale, M., Beaussart, A., Allen, J., Quilès, F., & El-Kirat-Chatel, S. (2019). Probing the Adhesion of the Common Freshwater Diatom Nitzschia palea at Nanoscale. *ACS applied materials & interfaces, 11*(51), 48574-48582.

Lear, G., Dopheide, A., Ancion, P. Y., Roberts, K., Washington, V., Smith, J., & Lewis, G. (2012). Biofilms in freshwater: their importance for the maintenance and monitoring of freshwater health. *Microbial Biofilms: Current Research and Applications*, 129-151.

Legesse, F. B., Ruger, J., Meyer, T., Krafft, C., Schmitt, M., & Popp, J. (2018). Investigation of Microalgal Carotenoid Content Using Coherent Anti-Stokes Raman Scattering (CARS) Microscopy and Spontaneous Raman Spectroscopy. *Chemphyschem, 19*(9), 1048-1055.

Lepetit, B., Volke, D., Gilbert, M., Wilhelm, C., & Goss, R. (2010). Evidence for the existence of one antenna-associated, lipid-dissolved and two protein-bound pools of diadinoxanthin cycle pigments in diatoms. *Plant physiology, 154*(4), 1905-1920.

Liu, B., & Benning, C. (2013). Lipid metabolism in microalgae distinguishes itself. *Current opinion in biotechnology, 24*(2), 300-309.

Lutz, M., Szponarski, W., Berger, G., Robert, B., & Neumann, J.-M. (1987). The stereoisomerism of bacterial, reaction-center-bound carotenoids revisited: an electronic absorption, resonance Raman and 1H-NMR study. *Biochimica et Biophysica Acta (BBA)-Bioenergetics, 894*(3), 423-433.

Maeda, Y., Nojima, D., Yoshino, T., & Tanaka, T. (2017). Structure and properties of oil bodies in diatoms. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 372*(1728), 20160408. doi:10.1098/rstb.2016.0408

McCreery, R. L. (2005). *Raman spectroscopy for chemical analysis* (Vol. 225): John Wiley & Sons.

Meksiarun, P., Spegazzini, N., Matsui, H., Matsuda, Y., & Sato, H. (2014). Raman Spectroscopy for Monitoring CO2 Effects on Fatty Acid Synthesis of Microalgal Marine Diatom Thalassiosira pseudonana. *Advanced Science, Engineering and Medicine, 6*(8), 873-875.

Meksiarun, P., Spegazzini, N., Matsui, H., Nakajima, K., Matsuda, Y., & Sato, H. (2015). In vivo study of lipid accumulation in the microalgae marine diatom Thalassiosira pseudonana using Raman spectroscopy. *Applied Spectroscopy, 69*(1), 45-51.

Mendes, T., Almeida, S. F., & Feio, M. J. (2012). Assessment of rivers using diatoms: effect of substrate and evaluation method. *Fundamental and Applied Limnology/Archiv für Hydrobiologie, 179*(4), 267-279.

Merlin, J. C. (1985). Resonance Raman spectroscopy of carotenoids and carotenoid-containing systems. *Pure and Applied Chemistry, 57*(5), 785-792.

Mimuro, M., Katoh, T., & Kawai, H. (1990). Spatial arrangement of pigments and their interaction in the fucoxanthin-chlorophyll ac protein assembly (FCPA) isolated from the brown alga Dictyota dichotoma. Analysis by means of polarized spectroscopy. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1015*(3), 450-456.

Morin, S., Gómez, N., Tornés, E., Licursi, M., & Rosebery, J. (2016). Benthic diatom monitoring and assessment of freshwater environments: standard methods and future challenges. *Aquatic Biofilms*, 111.

Navarro, E., Baun, A., Behra, R., Hartmann, N. B., Filser, J., Miao, A.-J., . . . Sigg, L. (2008). Environmental behavior and ecotoxicity of engineered nanoparticles to algae, plants, and fungi. *Ecotoxicology, 17*(5), 372-386.

Nielsen, P. H., Jahn, A., & Palmgren, R. (1997). Conceptual model for production and composition of exopolymers in biofilms. *Water Science and Technology, 36*(1), 11.

Pandey, L. K., Bergey, E. A., Lyu, J., Park, J., Choi, S., Lee, H., . . . Han, T. (2017). The use of diatoms in ecotoxicology and bioassessment: insights, advances and challenges. *Water Research, 118*, 39-58.

Parker, F. S. (1983). *Applications of infrared, Raman, and resonance Raman spectroscopy in biochemistry*: Springer Science & Business Media.

Patrick, R. (1973). Use of algae, especially diatoms, in the assessment of water quality. In *Biological methods for the assessment of water quality*: ASTM International.

Pinto, R., Mortágua, A., Almeida, S. F., Serra, S., & Feio, M. J. (2020). Diatom size plasticity at regional and global scales. *Limnetica, 39*(1), 387-403.

Premvardhan, L., Bordes, L., Beer, A., Buchel, C., & Robert, B. (2009). Carotenoid structures and environments in trimeric and oligomeric fucoxanthin chlorophyll a/c2 proteins from resonance Raman spectroscopy. *J Phys Chem B, 113*(37), 12565-12574.

Premvardhan, L., Robert, B., Beer, A., & Büchel, C. (2010). Pigment organization in fucoxanthin chlorophyll a/c2 proteins (FCP) based on resonance Raman

spectroscopy and sequence analysis. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1797*(9), 1647-1656.

Pytlik, N., Klemmed, B., Machill, S., Eychmüller, A., & Brunner, E. (2019). In vivo uptake of gold nanoparticles by the diatom Stephanopyxis turris. *Algal research, 39*, 101447.

Raman, C. V., & Krishnan, K. S. (1928). A new type of secondary radiation. *Nature, 121*(3048), 501-502.

Ren, F., Campbell, J., Rorrer, G. L., & Wang, A. X. (2014). Surface-enhanced Raman spectroscopy sensors from nanobiosilica with self-assembled plasmonic nanoparticles. *IEEE Journal of Selected Topics in Quantum Electronics, 20*(3), 127-132.

Round, F. E., Crawford, R. M., & Mann, D. G. (2007). *Diatoms: biology and morphology of the genera*: Cambridge university press.

Rüger, J., Mondol, A. S., Schie, I. W., Popp, J., & Krafft, C. (2019). High-throughput screening Raman microspectroscopy for assessment of drug-induced changes in diatom cells. *Analyst, 144*(15), 4488-4492.

Rüger, J., Unger, N., Schie, I. W., Brunner, E., Popp, J., & Krafft, C. (2016). Assessment of growth phases of the diatom Ditylum brightwellii by FT-IR and Raman spectroscopy. *Algal research, 19*, 246-252.

Sabater, S., Guasch, H., Ricart, M., Romaní, A., Vidal, G., Klünder, C., & Schmitt-Jansen, M. (2007). Monitoring the effect of chemicals on biological communities. The biofilm as an interface. *Analytical and bioanalytical chemistry, 387*(4), 1425-1434.

Salta, M., Wharton, J. A., Blache, Y., Stokes, K. R., & Briand, J. F. (2013). Marine biofilms on artificial surfaces: structure and dynamics. *Environmental microbiology, 15*(11), 2879-2893.

Schmidt, M. A., Lei, D. Y., Wondraczek, L., Nazabal, V., & Maier, S. A. (2012). Hybrid nanoparticle–microcavity-based plasmonic nanosensors with improved detection resolution and extended remote-sensing ability. *Nature communications, 3*(1), 1-8.

Shahbazyan, T. V., & Stockman, M. I. (2013). *Plasmonics: theory and applications*: Springer.

Simó, R. (2001). Production of atmospheric sulfur by oceanic plankton: biogeochemical, ecological and evolutionary links. *Trends in Ecology & Evolution, 16*(6), 287-294.

Smith, E., & Dent, G. (2019). *Modern Raman spectroscopy: a practical approach*: John Wiley & Sons.

Squires, L. E., Rushforth, S. R., & Brotherson, J. D. (1979). Algal response to a thermal effluent: study of a power station on the provo river, Utah, USA. *Hydrobiologia, 63*(1), 17-32.

Stewart, P. S. (2003). Diffusion in biofilms. *Journal of bacteriology, 185*(5), 1485-1491.

Stewart, S., Priore, R. J., Nelson, M. P., & Treado, P. J. (2012). Raman Imaging. *Annual Review of Analytical Chemistry, 5*(1), 337-360.

Stransky, H., & Hager, A. (1970). [The carotenoid pattern and the occurrence of the light-induced xanthophyll cycle in various classes of algae. IV. Cyanophyceae and Rhodophyceae]. *Arch Mikrobiol, 72*(1), 84-96.

Su, Y., Lundholm, N., & Ellegaard, M. (2018). Effects of abiotic factors on the nanostructure of diatom frustules—ranges and variability. *Applied microbiology and biotechnology, 102*(14), 5889-5899.

Supriya, G., Asulabha, K., & Ramachandra, T. (2012). *Use of Raman microspectroscopy to detect changes in lipid pools of microalgae*. Paper presented at the National Conference on Conservation and Management of Wetland Ecosystems, Kerala, India.

Talari, A. C. S., Movasaghi, Z., Rehman, S., & Rehman, I. U. (2015). Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews, 50*(1), 46-111.

Todd, E. C. (1993). Domoic acid and amnesic shellfish poisoning-a review. *Journal of food protection, 56*(1), 69-83.

Townley, H. E. (2011). Diatom frustules: physical, optical, and biotechnological applications. In *The Diatom World* (pp. 273-289): Springer.

UNESCO-WHO-UNEP. (1996). *Water Quality Assessments*. Cambridge, UK: Chapman & Hall.

Vilbaste, S., & Truu, J. (2003). Distribution of benthic diatoms in relation to environmental variables in lowland streams. *Hydrobiologia, 493*(1-3), 81-93.

Wagner, W. D., & Waidelich, W. (1986). Selective Observation of Chlorophyll c in Whole Cells of Diatoms by Resonant Raman Spectroscopy. *Applied Spectroscopy, 40*(2), 191-196.

White, I. M., Gohring, J., & Fan, X. (2007). SERS-based detection in an optofluidic ring resonator platform. *Optics express, 15*(25), 17433-17442.

Wood, B. R., Heraud, P., Stojkovic, S., Morrison, D., Beardall, J., & McNaughton, D. (2005). A portable Raman acoustic levitation spectroscopic system for the identification and environmental monitoring of algal cells. *Analytical chemistry, 77*(15), 4955-4961.

Wu, Q., Nelson, W., Treubig, J., Brown, P., Hargraves, P., Kirs, M., . . . Hanlon, E. (2000). UV resonance Raman detection and quantitation of domoic acid in phytoplankton. *Analytical chemistry, 72*(7), 1666-1671.

Xu, X., Hasan, D., Wang, L., Chakravarty, S., Chen, R. T., Fan, D., & Wang, A. X. (2012). Guided-mode-resonance-coupled plasmonic-active SiO2 nanotubes for surface enhanced Raman spectroscopy. *Applied physics letters, 100*(19), 191114.

Yang, J., Ren, F., Le, Z., Campbell, J., Rorrer, G. L., & Wang, A. X. (2014). *Surface-enhanced raman scattering immuno-assay using diatom frustules.* Paper presented at the 2014 Conference on Lasers and Electro-Optics (CLEO)-Laser Science to Photonic Applications.

Yuan, P., He, H. P., Wu, D. Q., Wang, D. Q., & Chen, L. J. (2004). Characterization of diatomaceous silica by Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 60*(12), 2941-2945.

Zienkiewicz, K., Du, Z.-Y., Ma, W., Vollheyde, K., & Benning, C. (2016). Stress-induced neutral lipid biosynthesis in microalgae—molecular, cellular and physiological insights. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1861*(9), 1269-1281.

Chapter 3 - A practical technique to identify Diatom *taxa*: Raman Spectroscopy

Raquel Pinto[2], Rui Vilarinho[3], António Paulo Carvalho[2], J. Agostinho Moreira[3], Laura Guimarães[1], Luís Oliva-Teles[2]

[1] CIIMAR/CIMAR - Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Avenida General Norton de Matos, s/n 4450-208 Matosinhos, Portugal
[2] Department of Biology, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n, 4169-007, Porto, Portugal
[3] Department of Physics, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n. 4169-007, Porto, Portugal

Abstract

Diatoms are recommended to assess the ecological status of freshwater ecosystems and have been used in water monitoring programs all over the world. The use of diatoms for water assessment requires an accurate taxonomic identification for calculation of autoecological indexes. Although taxonomic identification is very practical due to the fact that diatoms siliceous frustules have ornaments and shapes that vary among species, this methodology have also some drawbacks, such as the time and expertise required for the identification. To overcome these drawbacks Raman spectroscopy was applied to diatoms from lakes located in Northern Portugal. The Raman spectra obtained from 29 species, 15 genus, 12 families, 9 orders and 4 subclasses were analysed using chemometric methods. Data were used to analyze the multidimensional covariance structure among the Raman variables obtained (matrix of independent variables or matrix X) and the species collected in the three lakes (matrix of dependent variables or matrix Y) by a Partial Least Squares regression (PLS). Furthermore, a method was developed to streamline interpretation of results based on a high number of significant components calculated by the PLS regression. Subsequently an Artificial Neural Network (ANN) was used for *taxa* identification from Raman data. The PLS interpretation produced a Raman profile for each species useful to further characterize the

physiological status of the species. Artificial Neuronal Network models were useful to identify various *taxa* with high accuracy. High sensitivity was found for the identification of *Achananthidium exiguum* (67%), *Fragilaria crotonensis* (67%), *Amphora pediculus* (71%), *Achananthidium minutissimum* (80%) and *Melosira varians* (82%).

Keywords: Spectra, Artificial Neuronal Networks, Taxonomic levels, Pigments, Frustule, Lipids

Introduction

Human Activities have negatively impacted aquatic ecosystems in the recent decades due to the increase of water pollution and unreasonable use of resources (UNESCO-WHO-UNEP, 1996). Therefore, it was necessary to create aquatic monitoring programs in order to assess water contamination and design adequate management and protection measures that could help improving or maintaining water quality (UNESCO-WHO-UNEP, 1996). An example is Water Framework Directive (WFD), created in Europe, where biological elements complemented with hydromorphological and physical-chemical parameters are obligatorily used to assess ecological water quality in the member states (European Comission Council, 2000). Diatoms are part of the biological quality elements due to the fact that they are abundant in practically all kinds of aquatic systems (Round *et al.*, 2007), have fast and differential responses to changes in environmental parameters (Almeida & Gil, 2001; Squires *et al.*, 1979; Vilbaste & Truu, 2003) and are easy to sample (Lear *et al.*, 2012) and preserve (Mendes *et al.*, 2012). Since they are such good bioindicators they are frequently used to assess water quality not only under WFD but also in other world countries, namely, Canada, USA, Japan, South America and Australia (UNESCO-WHO-UNEP, 1996).

Usually, protocols involving diatoms as indicators of water quality, require an accurate taxonomic identification and counting of a fixed number of valves (Pandey *et al.*, 2017; Pinto *et al.*, 2020). The species abundance is then used to calculate autoecological indexes such as *"Indice of Polluossensibilité Specifique"* (IPS; CEMAGREF, 1982) that is used to calculate water ecological quality taking to account reference values without anthropogenic pressure for a given water course (INAG, 2008). Despite the fact that diatoms have an intricate siliceous outer wall differing in shape and patterns among species (Germain, 1981; Pandey *et al.*, 2017), the taxonomic identification is often challenging (Morin *et al.*, 2016). This process is frequently time-consuming and requires

some expertise (Morin *et al.*, 2016). Additionally, taxonomic levels and generic affinities between *taxa* change frequently owing to the result of new research on their morphometric characteristics and genetic affinities (Potapova & Hamilton, 2007). For example, the species *Achnanthidium minutissimum*, largely distributed worldwide, is frequently considered a species complex, due to its smaller size, which makes it difficult to distinguish this species. This species complex, initially considered *Achnantes minutissima,* also diverged in several similar species making even more difficult to clarify the concept of *A. minutissimum* (Potapova & Hamilton, 2007).

Raman spectroscopy (RS) is a promising technique that could ease the constraints inherent to diatom taxonomic identification based on morphology. In biological analysis, this technique has many advantages in relation to other methods because it is label-free, water interference can be minimized (Parker, 1983) and it requires no or minimal preparation and processing of the samples to be analysed (Heraud *et al.*, 2007). Studies applying RS to diatoms do not usually focus on taxonomic differences. Instead, they are mainly centered on understanding the conformation, location and variation with abiotic factors for a variety of cell components such as pigments (Alexandre *et al.*, 2014; Premvardhan *et al.*, 2009; Premvardhan *et al.*, 2010), siliceous frustule (Yuan *et al.*, 2004), lipids (Meksiarun *et al.*, 2015), extracellular polymeric substances (EPS) (Laviale *et al.*, 2019), mucilage (Chen *et al.*, 2019) and toxins (Wu *et al.*, 2000). For example, bands characteristic of carotenoids in diatom *Cyclotella meneghiniana* vary with light conditions in the different carotenoids produced and their conformation (Alexandre *et al.*, 2014). Also, in the diatom *Thalassiosira pseudonana* bands associated with fatty acids vary with high carbon dioxide levels due to the enhancement of their production *(*Meksiarun *et al., 2015)*. Raman Spectroscopy was also used in toxicological assays: in *Phaeodactylum tricornutum* to discriminate diothreitol effects between high light and low light conditions (Rüger *et al.*, 2019) and in *Stephanopyxis turris* to study the mechanisms of incorporation of gold nanoparticles in the cell (Pytlik *et al.*, 2019). Moreover, while RS applied to diatoms has been used to purposes other *taxa* identification*,* some authors refer that spectral bands can vary with *taxa* (Abbas *et al*., 2011; Wood *et al*., 2005; Yuan *et al*., 2004). In contrast, ANN have been largely used to predict *taxa* not only in diatoms (Pedraza *et al.*, 2018), but also in other types of organisms such as other algae and phytoplankton *(*Boddy *et al*., 2000; Medlin, 2012*)*, macroinvertebrates (Joutsijoki *et al.*, 2014), fungi (Naumann, 2009) and even plants (Wolski & Kruk, 2020). These studies, however, do not rely in Raman spectral data. They rely on photographs of the specimens (Joutsijoki *et al.*, 2014; Pedraza *et al.*, 2018), cytometric flux data *(*Boddy *et al*., 2000;

Medlin, 2012) and Fourier-Transformed Infrared (FT-IR) spectrometric data (Naumann, 2009). A recent work using Region based Convolutional Neural Networks (RCNN) based on diatom images belonging to ten diatom species enabled diatom species prediction with 13-68% of overall sensitivity (Pedraza *et al.*, 2018). Hence, given the advances and advantages shown by these two methods, in the present work it was hypothesized if after spectral deconvolution, diatom RS, together with ANN could be used to develop a methodology for correct diatom taxonomic identification.

Thus, the main objective of this work was to develop a diagnose test for taxonomic levels (species, genus, family, order) using diatom RS. In parallel, differences between Raman bands belonging to different diatom species were also explored. For this purpose, 790 spectra were acquired in 29 diatom *taxa* sampled in lakes of a city park in Northern Portugal. The data were analyzed using Partial Least Squares (PLS) and with ANN to obtain classification models discriminating the various taxonomic levels.

2. Materials and Methods

2.1. Study Area

Collection of diatom samples took place in the three main lakes of Oporto City Park represented in figure 8. This urban park is located at limit between Oporto and Matosinhos cities (North of Portugal). It consists on a recreative area of about 83 ha. Water recirculates between the lakes and it is used to irrigate the extensive forested area composed by many tree and shrub species (Morais, 2009). The fauna of this park is mainly composed by numerous species of native and non-native birds and fish. In relation to phytoplankton, some species of cyanobacteria were detected in these lakes: *Microcystis* sp., *M.aeruginosa* and *Planktothrix* sp. *Cylindrospermopsis racirborskii, Planktothrix agardhi* (Morais, 2009; Matos, 2014).

Figure 8 - Air photograph of Oporto City Park where the sampled lakes are located, retrieved from Google Earth version 7.3.2.5776. The park is delimited by a dark line and sampling points are marked with a dark blue icon and the respective designation: Lake 1 (41.1678357º; -8.6737829º), Lake 2 (41.1676818º; -8.6778465º) and Lake 3 (41.1690561º; -8.6835319º).

2.2. Sampling and storing procedure

Diatom samples were collected according to the protocol described by the Portuguese Water Institute (INAG, 2008) with few adaptations taking into account the availability of suitable substrates on sampling sites. A toothbrush was used to scrap natural and artificial substrates over an area of 100 cm^2 for each lake. The substrates were rocks, wood, sediment, bricks or underwater plastic tubes. Previous work has shown collection from different substrates does not influence on the sample quality for diatom analysis (Mendes *et al.*, 2012). When toothbrush sampling was not feasible, a similar area of biofilm was pippeted from the substrate surface. The biofilm collected was then resuspended in lake water contained in a laboratory tray. The sampled biofilm was then transferred into ten flasks for each lake: one 60 mL capacity dark glass flask containing the biofilm preserved in 33% formaldehyde and nine 120 mL capacity plastic flasks. The biofilm samples were temporarily stored in a thermal box for transport. In the laboratory the plastic flasks were stored at -80ºC until further analyzis.

2.3. Sample processing and Diatom taxonomic identification

For taxonomic identification, samples belonging to each lake were oxidized with 10 mL of nitric acid with some crystals of potassium dichromate for 24 hours, as described in the protocol of the Water Portuguese Institute (INAG, 2008). Then, the oxidants present in the samples were removed by successive centrifugation, followed by supernatant discharge and ensuing resuspension in distilled water. This process had to be repeated for several times, ensuring all the oxidants in the samples were removed. Centrifugations were done at 1500 rpm at room temperature, in a Kubota 2420 Centrifuge (Kubota Corporation, Japan). After the cleaning process, the turbidity in the samples was decreased by dilution in water. The dilution also helped achieving lower number of cell frustules in optical field, ideal for the taxonomic identification. The ideal dilution was confirmed under light microscope (Zeiss Primo Star, 40x, N.A.= 0.65). Permanent slides were obtained by mounting with *Naphrax* ® (Brunel Microscopes, Ltd.). The identification of diatoms was done using a light microscope (Zeiss Primo Star, 100x, N.A.=1.25) with the aid of supporting diatom floras (Germain, 1981). For each sample, 400 valves were observed and used for identification.

2.4. Raman Spectroscopy

Biofilm samples were defrosted at 4ºC and dropped onto microscope slides that were dried at room temperature. The aim of this step was to preventing valve movement, owing to water evaporation, during the RS acquisition. The Raman readings were done with an InVia™ Qontor® confocal Raman microscope (Renishaw, United Kingdom) assembled with a Leica DM2700 microscope (Ernst Leitz GmbH, Germany) and a 50x objective used for focusing the laser beam onto each diatom in the sample. A Cobolt 04-01 Series Samba™ (Hübner Photonics, Germany) incident laser was employed. The laser was set to 532 nm and 0.1 mW. The spectra acquisition time used was 10s and 3 accumulations. Eighteen spectra were obtained per diatom species with significant abundance (>1%) in each lake. Some *taxa* previously detected in the oxidized samples could not be found in the non-oxidized samples, though they had met the abundance criteria. No Raman readings could thus be done for these *taxa*. For the remaining *taxa* the readings were done in the cell region located between the central area and the apex, including the chloroplast; the raphe area was excluded from the readings. The software WiRE™ 5.2. (Renishaw Inc., UK) was used to acquire the Raman Spectra. Furthermore,

the area, width and frequency of the bands were estimated in the software Igor Pro™ with with its oscillatory function (Wavemetrics Inc., 1998).

2.5. Data Analysis

The first step of data analysis was to normalize the areas of the bands detected. This is common step in the analysis allowing to correct for fluctuations of intensity in the spectra obtained. To achieve the most appropriate normalization, data correlations for the three Raman variables (area, width and frequency) were done. These correlations were firstly investigated in the whole raw dataset and led to the identification of area of the Raman band 1526 cm^{-1} as the variable bearing more significant correlations with the remaining areas. After the normalization by this band, such correlations were reduced or disappeared. This was in agreement with findings of previous authors applying Raman spectroscopy to diatom investigation who also employed this band for normalization (Alexandre *et al.*, 2014; Premvardhan *et al.*, 2009). The correlations found are represented in Tables IA, IB, IC and ID of Appendix II.

A PLS was then performed to describe the studied species and identify the band components explaining the highest covariability in the dataset. In order to further investigate patterns in Raman variables and to infer about relationships in the dataset, a cluster analysis were done on the *x loadings* and the *y loadings* obtained from the PLS.

For *taxa* diagnosis, classification models were derived using ANN with supervised learning. For this, Multilayer Perceptron (MLP) was used as network architecture. Operations were organized into a three-dimension network; namely input layer, hidden layer and output layer as described previously by Bishop (1995). Through this procedure, each neuron performs a weighted sum of its inputs and passes it through a transfer function to produce an output (Bishop, 1995). To achieve this, the data was randomly subdivided into a training series for preliminary adjustment of the model to the data, a testing series serving to calibrate the model and a validation series used by the ANN to validate the model (Drobatz, 2009; Langlotz, 2003). The ANN was performed taking into account various diatom taxonomic levels: species, genus, family, order and subclass as categorical target and Raman variables as continuous input. The class was not considered as categorical input due to the large differences in the distribution of the data. Finally, the ANN method was evaluated for its classification performance by common measures employed in diagnostic tests; accuracy and sensitivity rates. All statistical

analysis was done with the software Statsoft Statistica™ 64 (Statsoft Software Inc., 2014).

3. Results

3.1. Diatom Taxonomic Identification

In total, 45 species were counted in all the three sampled lakes. Of all the 45 species, 29 species showing >1% abundance in at least one lake were used for RS measurements. Diatom valve count and percentages per lake and in total are represented in Table II of Appendix II; species measured in the RS analysis are also highlighted. Figure 9 presents light microscope photographs of the two most abundant species in each lake.

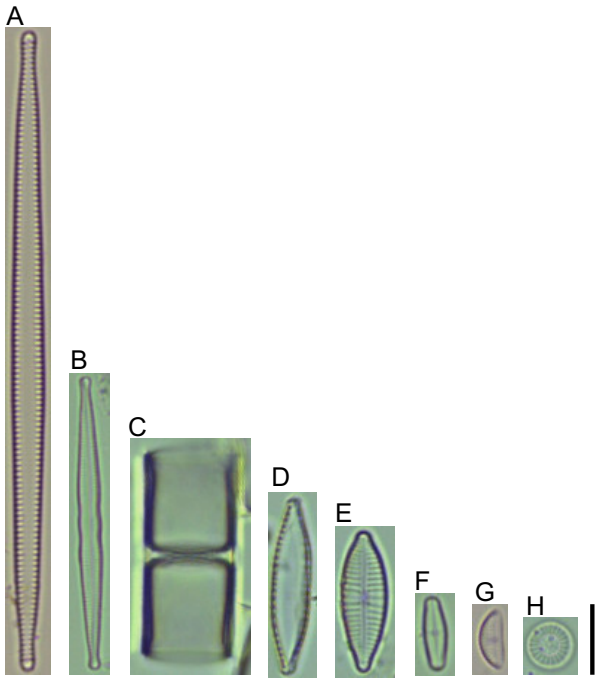


Figure 9 - Most abundant species in the three lakes of the city park of Oporto. A – *Tabularia tabulata* abundant in lakes 2 and 3; B – *Fragilaria crotonensis* abundant in lake 1; C – *Melosira varians* (colony in connective view) abundant in lake 1; D – *Nitzschia palea* abundant in lake 2; E – *Gomphonema parvulum* abundant in lake 1; F – *Achnanthidium minutissimum* abundant in lake 2; G– *Amphora pediculus* abundant in lake 3; H – *Cyclotella stelligera* abundant in lake 3. Scale bar=10 µm.

In Lake 1, the three most abundant species were *Gomphonema parvulum* (Fig. 9E), *Melosira varians* (Fig. 9C) and *Fragilaria crotonensis* (Fig. 9B). In Lake 2 the three most abundant species were *Tabularia tabulata* (Fig. 9A), *Achnanthidium minutissimum* (Fig.

9F) and *Nitzschia palea* (Fig. 9D). In Lake 3, *Tabularia tabulata* was also very abundant, however, the most abundant species was *Amphora pediculus* (Fig. 9G). In this lake, *Cyclotella stelligera* (Fig. 9H) was also very abundant.

3.2. Species characterization

The 29 species used in the RS measurements are included in 15 genus, 12 families, 9 orders, 4 subclasses and 3 classes. In total, 790 Raman spectra were measured. The amount of Raman spectra collected for each diatom genus, family, order, subclass and class is represented in Table 3. A total of 14 Raman bands were identified in diatom Raman spectra: around 867 cm^{-1}, 920 cm^{-1}, 963 cm^{-1}, 1013 cm^{-1}, 1160 cm^{-1}, 1180 cm^{-1}, 1198 cm^{-1}, 1270 cm^{-1}, 1315 cm^{-1}, 1390 cm^{-1}, 1445 cm^{-1}, 1526 cm^{-1}, 1606 cm^{-1} and 1656 cm^{-1}. Examples of Raman spectra obtained for different species are presented in Figure 10.

Table 3 - Amount of Raman spectra collected for each diatom genus, family, order, subclass and class. The numbers in brackets represent the amount of spectra collected for each *taxa*. A total of 790 Raman spectra were acquired.

Genus	**Family**	**Order**	**Subclass**	**Class**
Achnanthidium (72)	*Achnanthidiaceae* (80)	*Cocconeidales* (80)	*Bacillariophycidae* (556)	Bacillariophyceae (718)
Planothidium (8)				
Amphora (54)	*Ctenulaceae* (54)	*Thalassiophysales* (54)		
Cymbella (18)	*Cymbelaceae* (18)	*Cymbellales* (126)		
Gomphonema (108)	*Gomphonemataceae* (108)			
Nitzschia (162)	*Bacillariaceae* (162)	*Bacillariales* (162)		
Navicula (126)	*Naviculaceae* (126)	*Naviculales* (134)		
Eolimna (8)	*Sellaphoraceae* (8)			
Fragilaria (54)	*Fragilariaceae* (54)	*Fragillariales* (72)	*Fragilariophycidae* (162)	
Pseudostaurosira (18)	*Staurosiraceae* (18)			
Ctenophora (36)	*Ulnariaceae* (90)	*Licmophorales* (90)		
Tabularia (36)				
Ulnaria (18)				

Melosira (54)	*Melosiraceae* (54)	*Melosirales* (54)	*Melosirophycidae* (54)	*Coscinodiscophyceae* (54)
Cyclotella (18)	*Stephanodiscaceae* (18)	*Stephanodiscales* (18)	*Thalassiosirophycidae* (18)	*Mediophyceae* (18)

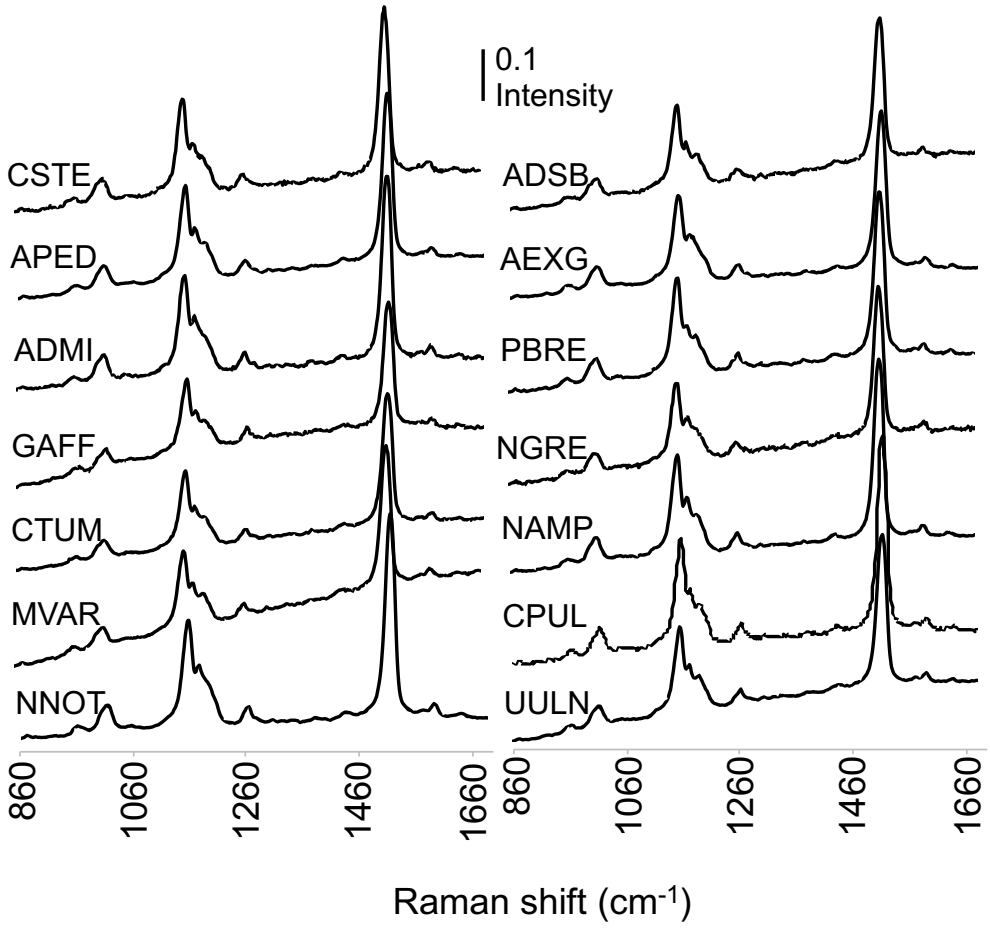


Figure 10 - Examples of Raman spectra recorded in various species: *Cyclotella stelligera* (CSTE), *Amphora pediculus* (APED), *Achnanthidium minutissimum* (ADMI); *Gomphonema affine* (GAFF); *Cymbella tumida* (CTUM); *Melosira varians* (MVAR); *Navicula notha* (NNOT); *Achnanthidium straubianum* (ADSB); *Achnanthidium exiguum* (AEXG); *Pseudostaurosira bevistriata* (PBRE); *Nitzschia gregaria* (NGRE); *Nitzschia amphibia* (NAMP); *Ctenophora puclchella* (CPUL); *Ulnaria ulna* (UULN).

A PLS regression was done with the Raman data obtained to depict the species response profiles. This is a chemometric method useful to model multiple response variables that may be related. The PLS regression calculated six significant components. Eleven variables were found to have higher contribution to these components. In Figure 11, Raman variables are ordered according to their importance to the components calculated by the PLS regression. The eleven important variables are highlighted in red. These Raman variables are related to six different bands. Table 4 summarizes the bands

and their known molecular assignment. Most of these bands are assigned to pigments, *i.e*. carotenoids and chloropyll *a*. Band 1270 cm^{-1} was assigned to proteins, lipids or nucleic acids. Bands 1160, 1180 and 1198 cm^{-1} can also be assigned to sulfur compounds in the frustule.

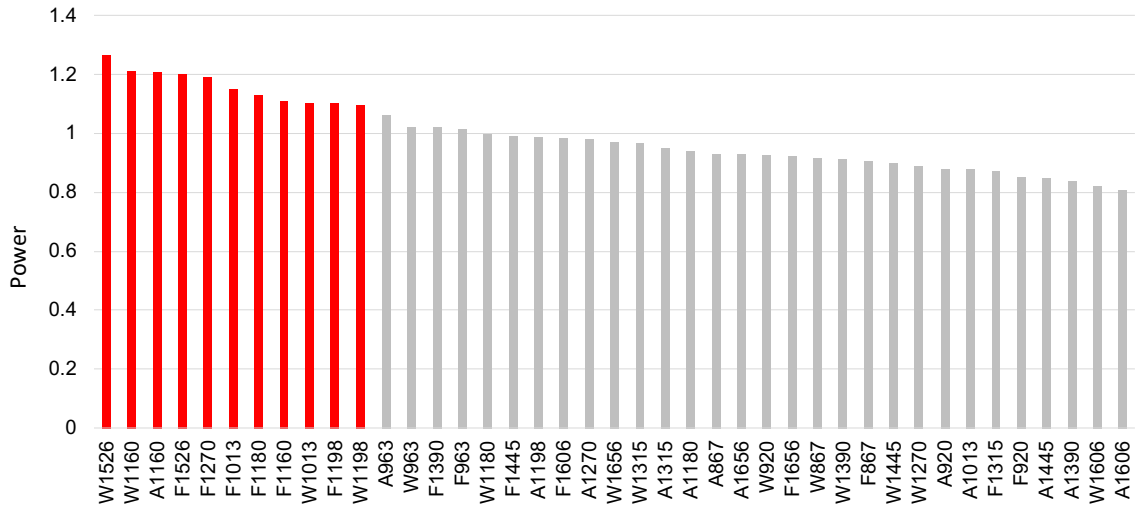


Figure 11 - Most important Raman variables explaining the combined variance in the components calculated by the Partial Least Squares regression. The most important variables are highlighted in red: Width (W) of the bands 1526, 1160, 1013 and 1198 cm^{-1}, Area (A) of the band 1160 cm^{-1} and Frequency (F) of the bands 1526, 1270, 1013, 1180, 1160 and 1198 cm^{-1}.

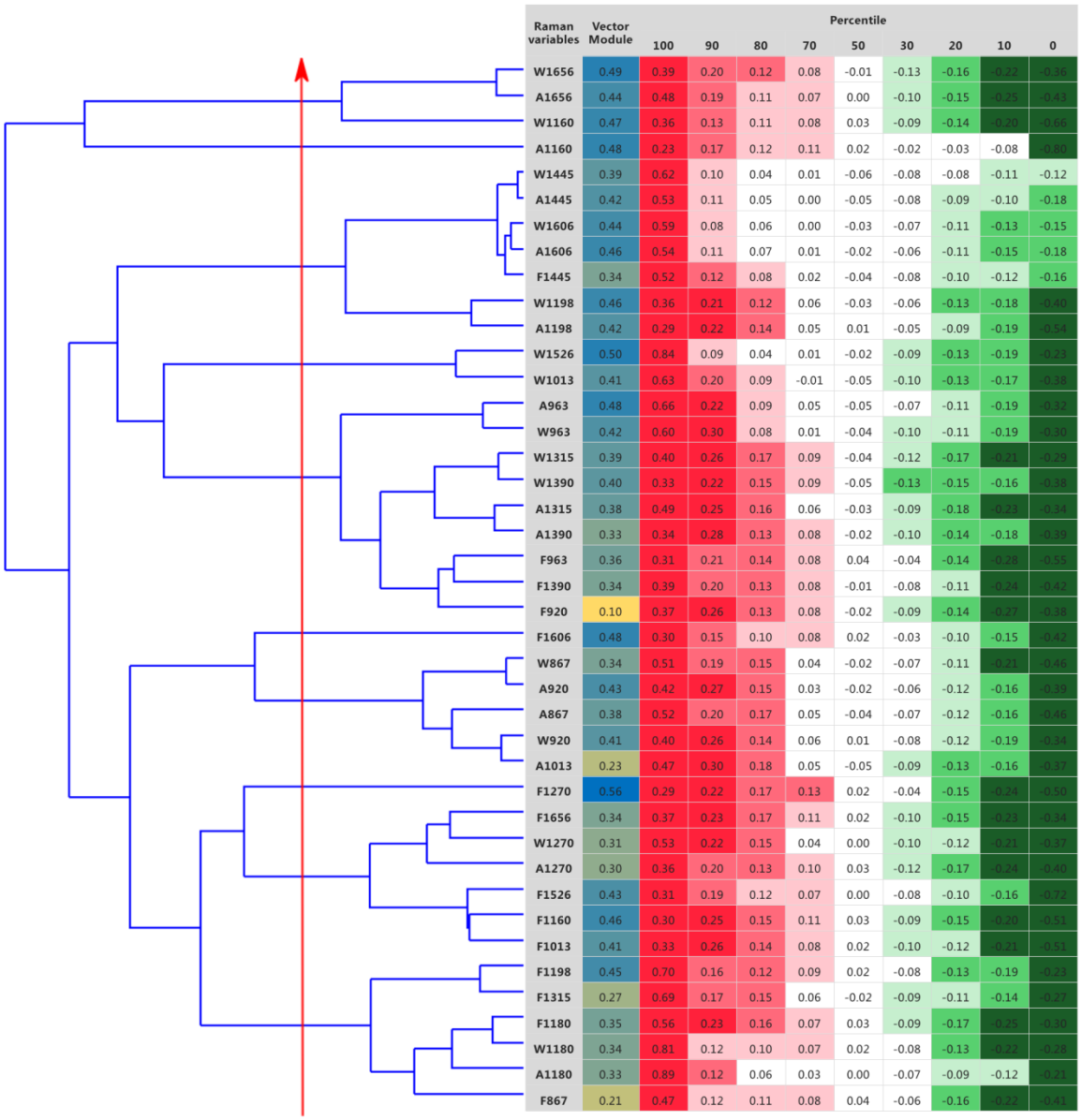
Table 4 - Most important Raman bands calculated by the Partial Least Squares regression, their mode assignments and the respective reference.

Band (cm^{-1})	Assignments	Reference
1013	CH3 in-plane wagging modes from Carotenoids	(Premvardhan *et al.*, 2009; Ruger et al., 2016)
1160	C=S (Frustule)	(De Tommasi, 2016; De Tommasi *et al.*, 2018)
	C-C stretching modes from Carotenoids	(Premvardhan *et al.*, 2009; Ruger et al., 2016)
1180	C=S (Frustule)	(De Tommasi, 2016; De Tommasi *et al.*, 2018)
	C-H deformational modes from Carotenoids	(Premvardhan *et al.*, 2009)
1198	C=S (frustule)	(De Tommasi, 2016; De Tommasi *et al.*, 2018)
	N-C stretching modes from Chl a	(Ruger *et al.*, 2016)

1270	Amide III; =CH bend (lipids); T,A	(Notingher, 2007)
1526	C=C streching modes from Carotenoids	(Premvardhan *et al.*, 2009; Alexandre *et al.*, 2014; Ruger *et al.*, 2016)

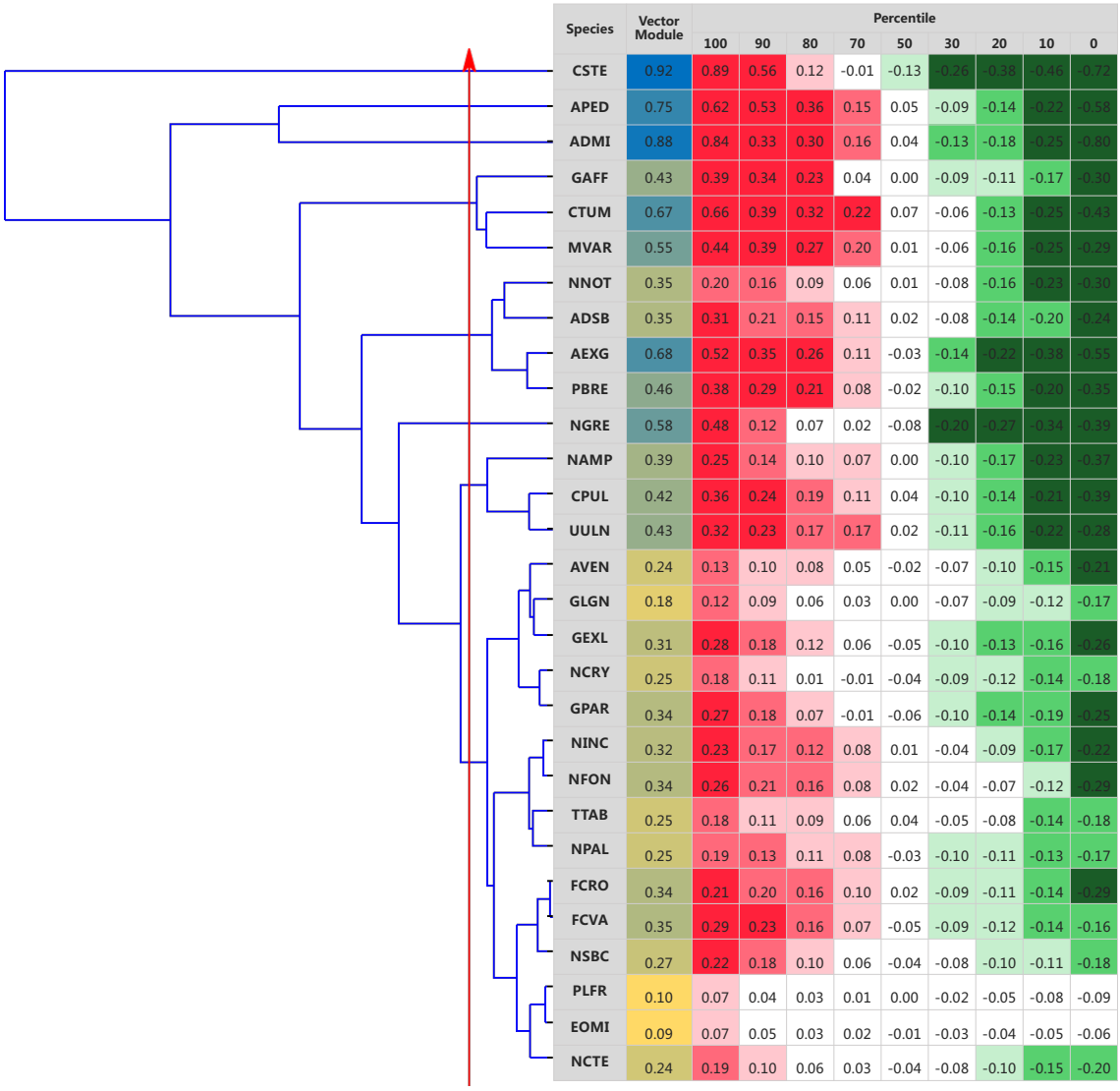
To aid the interpretation of the results, an integrated measure of relation between the species and the Raman variables was derived. This was done by calculating the weight of each Raman variable on each species in the hyperspace formed by the six significant components derived by the PLS (Figures 12 to 15). This weight can be obtained from the projection of species vectors (y_i) on the Raman variables (x_i). The vector modules, i.e. the distance from the origin to a given Raman variable or a species in the hyperspace defined by the six significant components, were also calculated. Figure 12 presents the percentiles for the weights of each Raman variable on all species. From its observation it is clear that all Raman variables contributed to species descrimination or profile, as showed by the high weight values found for all variables indicated by the extreme percentiles (0-10th percentile for negative values and 90-100th percentile). Furthermore, the weights obtained allowed to clearly identify the species best discriminated by the model (Figure 13). Globally two main groups of species are visible in the figure; one with vector module values ranging from 0.43 to 0.92 representing the species better characterized by the Raman variables and a second group (less coloured) with lower vector modules. The first group is composed by the species *Cyclotella stelligera* (CSTE), *Amphora pediculus* (APED), *Achnanthidium minutissimum* (ADMI), *Gomphonema affine* (GAFF), *Cymbella tumida* (CTUM), *Melosira varians* (MVAR), *Navicula notha* (NNOT), *Achnanthidium straubianum* (ADSB), *Achnanthidium exiguum* (AEXG), *Pseudostaurosira brevistriata* (PBRE), *Nitzschia gregaria* (NGRE), *Nitzschia amphibia* (NAMP), *Ctenophora pulchella* (CPUL) and *Ulnaria ulna* (UULN). The second group is composed by the species *Amphora veneta* (AVEN); *Gomphonema lagenula* (GLGN); *Gomphonema exilissimum* (GEXL); *Navicula cryptocephala* (NCRY), *Gomphonema parvulum* (GPAR), *Nitzschia inconspicua* (NINC), *Nitzschia fonticola* (NFON), *Tabularia tabulata* (TTAB), *Nitzschia palea* (NPAL), *Fragilaria crotonensis* (FCRO), *Fragilaria vaucheriae* (FCVA), *Nitzschia subcapitellata* (NSBC), *Planothidium frequentissimum* (PLFR), *Eolimna minima* (EOMI) and *Navicula cryptotenella* (NCTE). Figures 14 and 15 present the weight of each Raman variable on each species for these two groups of species. Globally, each species exhibits its own Raman profile. Moreover, the highest weights were found for *Cyclotella stelligera* (CSTE) (W1160, F1526, F1198, F1315, F1180, W1180, A1180), *Amphora pediculus* (APED) (A1160, W1445, W1606, A1606, W1013), *Achnanthidium minutissimum* (ADMI) (A1160, A1198, W1526, W1013,

W1270), *Cymbella tumida* (CTUM) (A963, W963), and *Achnanthidium exiguum* (AEXG) (F963).



Raman variables	Vector Module	Percentile 100	90	80	70	50	30	20	10	0
W1656	0.49	0.39	0.20	0.12	0.08	-0.01	-0.13	-0.16	-0.22	-0.36
A1656	0.44	0.48	0.19	0.11	0.07	0.00	-0.10	-0.15	-0.25	-0.43
W1160	0.47	0.36	0.13	0.11	0.08	0.03	-0.09	-0.14	-0.20	-0.66
A1160	0.48	0.23	0.17	0.12	0.11	0.02	-0.02	-0.03	-0.08	-0.80
W1445	0.39	0.62	0.10	0.04	0.01	-0.06	-0.08	-0.08	-0.11	-0.12
A1445	0.42	0.53	0.11	0.05	0.00	-0.05	-0.08	-0.09	-0.10	-0.18
W1606	0.44	0.59	0.08	0.06	0.00	-0.03	-0.07	-0.11	-0.13	-0.15
A1606	0.46	0.54	0.11	0.07	0.01	-0.02	-0.06	-0.11	-0.15	-0.18
F1445	0.34	0.52	0.12	0.08	0.02	-0.04	-0.08	-0.10	-0.12	-0.16
W1198	0.46	0.36	0.21	0.12	0.06	-0.03	-0.06	-0.13	-0.18	-0.40
A1198	0.42	0.29	0.22	0.14	0.05	0.01	-0.05	-0.09	-0.19	-0.54
W1526	0.50	0.84	0.09	0.04	0.01	-0.02	-0.09	-0.13	-0.19	-0.23
W1013	0.41	0.63	0.20	0.09	-0.01	-0.05	-0.10	-0.13	-0.17	-0.38
A963	0.48	0.66	0.22	0.09	0.05	-0.05	-0.07	-0.11	-0.19	-0.32
W963	0.42	0.60	0.30	0.08	0.01	-0.04	-0.10	-0.11	-0.19	-0.30
W1315	0.39	0.40	0.26	0.17	0.09	-0.04	-0.12	-0.17	-0.21	-0.29
W1390	0.40	0.33	0.22	0.15	0.09	-0.05	-0.13	-0.15	-0.16	-0.38
A1315	0.38	0.49	0.25	0.16	0.06	-0.03	-0.09	-0.18	-0.23	-0.34
A1390	0.33	0.34	0.28	0.13	0.08	-0.02	-0.10	-0.14	-0.18	-0.39
F963	0.36	0.31	0.21	0.14	0.08	0.04	-0.04	-0.14	-0.28	-0.55
F1390	0.34	0.39	0.20	0.13	0.08	-0.01	-0.08	-0.11	-0.24	-0.42
F920	0.10	0.37	0.26	0.13	0.08	-0.02	-0.09	-0.14	-0.27	-0.38
F1606	0.48	0.30	0.15	0.10	0.08	0.02	-0.03	-0.10	-0.15	-0.42
W867	0.34	0.51	0.19	0.15	0.04	-0.02	-0.07	-0.11	-0.21	-0.46
A920	0.43	0.42	0.27	0.15	0.03	-0.02	-0.06	-0.12	-0.16	-0.39
A867	0.38	0.52	0.20	0.17	0.05	-0.04	-0.07	-0.12	-0.16	-0.46
W920	0.41	0.40	0.26	0.14	0.06	0.01	-0.08	-0.12	-0.19	-0.34
A1013	0.23	0.47	0.30	0.18	0.05	-0.05	-0.09	-0.13	-0.16	-0.37
F1270	0.56	0.29	0.22	0.17	0.13	0.02	-0.04	-0.15	-0.24	-0.50
F1656	0.34	0.37	0.23	0.17	0.11	0.02	-0.10	-0.15	-0.23	-0.34
W1270	0.31	0.53	0.22	0.15	0.04	0.00	-0.10	-0.12	-0.21	-0.37
A1270	0.30	0.36	0.20	0.13	0.10	0.03	-0.12	-0.17	-0.24	-0.40
F1526	0.43	0.31	0.19	0.12	0.07	0.00	-0.08	-0.10	-0.16	-0.72
F1160	0.46	0.30	0.25	0.15	0.11	0.03	-0.09	-0.15	-0.20	-0.51
F1013	0.41	0.33	0.26	0.14	0.08	0.02	-0.10	-0.12	-0.21	-0.51
F1198	0.45	0.70	0.16	0.12	0.09	0.02	-0.08	-0.13	-0.19	-0.23
F1315	0.27	0.69	0.17	0.15	0.06	-0.02	-0.09	-0.11	-0.14	-0.27
F1180	0.35	0.56	0.23	0.16	0.07	0.03	-0.09	-0.17	-0.25	-0.30
W1180	0.34	0.81	0.12	0.10	0.07	0.02	-0.08	-0.13	-0.22	-0.28
A1180	0.33	0.89	0.12	0.06	0.03	0.00	-0.07	-0.09	-0.12	-0.21
F867	0.21	0.47	0.12	0.11	0.08	0.04	-0.06	-0.16	-0.22	-0.41

Figure 12 - Vector module and percentiles for the weights of each Raman variable over all species calculated from the projection of each given species (y_i) over each Raman variable (x_i) in the hyperspace defined by the six components returned by the Partial Least Squares regression. The Raman variables are represented as width (W), area (A) and frequency (F) of the spectral bands



Species	Vector Module	Percentile								
		100	90	80	70	50	30	20	10	0
CSTE	0.92	0.89	0.56	0.12	-0.01	-0.13	-0.26	-0.38	-0.46	-0.72
APED	0.75	0.62	0.53	0.36	0.15	0.05	-0.09	-0.14	-0.22	-0.58
ADMI	0.88	0.84	0.33	0.30	0.16	0.04	-0.13	-0.18	-0.25	-0.80
GAFF	0.43	0.39	0.34	0.23	0.04	0.00	-0.09	-0.11	-0.17	-0.30
CTUM	0.67	0.66	0.39	0.32	0.22	0.07	-0.06	-0.13	-0.25	-0.43
MVAR	0.55	0.44	0.39	0.27	0.20	0.01	-0.06	-0.16	-0.25	-0.29
NNOT	0.35	0.20	0.16	0.09	0.06	0.01	-0.08	-0.16	-0.23	-0.30
ADSB	0.35	0.31	0.21	0.15	0.11	0.02	-0.08	-0.14	-0.20	-0.24
AEXG	0.68	0.52	0.35	0.26	0.11	-0.03	-0.14	-0.22	-0.38	-0.55
PBRE	0.46	0.38	0.29	0.21	0.08	-0.02	-0.10	-0.15	-0.20	-0.35
NGRE	0.58	0.48	0.12	0.07	0.02	-0.08	-0.20	-0.27	-0.34	-0.39
NAMP	0.39	0.25	0.14	0.10	0.07	0.00	-0.10	-0.17	-0.23	-0.37
CPUL	0.42	0.36	0.24	0.19	0.11	0.04	-0.10	-0.14	-0.21	-0.39
UULN	0.43	0.32	0.23	0.17	0.17	0.02	-0.11	-0.16	-0.22	-0.28
AVEN	0.24	0.13	0.10	0.08	0.05	-0.02	-0.07	-0.10	-0.15	-0.21
GLGN	0.18	0.12	0.09	0.06	0.03	0.00	-0.07	-0.09	-0.12	-0.17
GEXL	0.31	0.28	0.18	0.12	0.06	-0.05	-0.10	-0.13	-0.16	-0.26
NCRY	0.25	0.18	0.11	0.01	-0.01	-0.04	-0.09	-0.12	-0.14	-0.18
GPAR	0.34	0.27	0.18	0.07	-0.01	-0.06	-0.10	-0.14	-0.19	-0.25
NINC	0.32	0.23	0.17	0.12	0.08	0.01	-0.04	-0.09	-0.17	-0.22
NFON	0.34	0.26	0.21	0.16	0.08	0.02	-0.04	-0.07	-0.12	-0.29
TTAB	0.25	0.18	0.11	0.09	0.06	0.04	-0.05	-0.08	-0.14	-0.18
NPAL	0.25	0.19	0.13	0.11	0.08	-0.03	-0.10	-0.11	-0.13	-0.17
FCRO	0.34	0.21	0.20	0.16	0.10	0.02	-0.09	-0.11	-0.14	-0.29
FCVA	0.35	0.29	0.23	0.16	0.07	-0.05	-0.09	-0.12	-0.14	-0.16
NSBC	0.27	0.22	0.18	0.10	0.06	-0.04	-0.08	-0.10	-0.11	-0.18
PLFR	0.10	0.07	0.04	0.03	0.01	0.00	-0.02	-0.05	-0.08	-0.09
EOMI	0.09	0.07	0.05	0.03	0.02	-0.01	-0.03	-0.04	-0.05	-0.06
NCTE	0.24	0.19	0.10	0.06	0.03	-0.04	-0.08	-0.10	-0.15	-0.20

Figure 13 - Figure 1 - Vector module and percentiles for the weights of each species over all Raman variables calculated from the projection of each given species (y_i) over each Raman variable (x_i) in the hyperspacedefined by the six components returned by the Partial Least Squares regression. Species are *Cyclotella stelligera* (CSTE), *Amphora pediculus* (APED), *Achnanthidium minutissimum* (ADMI), *Gomphonema affine* (GAFF), *Cymbella tumida* (CTUM), *Melosira varians* (MVAR), *Navicula notha* (NNOT), *Achnanthidium straubianum* (ADSB), *Achnanthidium exiguum* (AEXG), *Pseudostaurosira brevistriata* (PBRE), *Nitzschia gregaria* (NGRE), *Nitzschia amphibia* (NAMP), *Ctenophora pulchella* (CPUL), *Ulnaria ulna* (UULN), *Amphora veneta* (AVEN); *Gomphonema lagenula* (GLGN); *Gomphonema exilissimum* (GEXL); *Navicula cryptocephala* (NCRY), *Gomphonema parvulum* (GPAR), *Nitzschia inconspicua* (NINC), *Nitzschia fonticola* (NFON), *Tabularia tabulata* (TTAB), *Nitzschia palea* (NPAL), *Fragilaria crotonensis* (FCRO), *Fragilaria vaucheriae* (FCVA), *Nitzschia subcapitellata* (NSBC), *Planothidium frequentissimum* (PLFR), *Eolimna minima* (EOMI) and *Navicula cryptotenella* (NCTE).

Raman variables	Species													
	CSTE	APED	ADMI	GAFF	CTUM	MVAR	NNOT	ADSB	AEXG	PBRE	NGRE	NAMP	CPUL	UULN
W1656	-0.29	0.02	0.32	0.05	-0.36	-0.20	0.02	-0.21	-0.14	-0.15	0.39	0.17	-0.18	-0.11
A1656	-0.19	0.00	0.14	0.00	-0.43	-0.29	0.05	-0.24	-0.16	-0.15	0.48	0.25	-0.11	-0.03
W1160	-0.66	-0.19	-0.15	0.00	-0.26	0.05	0.11	-0.15	0.11	0.07	0.36	0.06	-0.08	0.03
A1160	0.12	-0.58	-0.80	0.02	-0.06	0.10	0.02	-0.13	-0.03	-0.03	0.23	0.12	0.05	0.17
W1445	-0.11	0.62	0.04	0.09	0.22	-0.06	-0.08	-0.07	-0.12	-0.07	-0.05	0.02	0.06	0.14
A1445	-0.01	0.53	-0.18	0.09	0.24	-0.04	-0.06	-0.08	-0.09	-0.06	0.00	0.04	0.08	0.17
W1606	-0.13	0.59	-0.07	0.00	0.08	-0.15	0.01	-0.07	-0.03	-0.01	0.07	0.07	0.11	0.17
A1606	-0.04	0.54	-0.18	-0.01	0.07	-0.17	0.03	-0.07	0.01	0.01	0.12	0.08	0.10	0.17
F1445	0.09	0.52	-0.15	0.03	0.13	-0.16	-0.04	-0.08	-0.12	-0.08	0.04	0.10	0.12	0.19
W1198	-0.40	0.36	-0.25	-0.17	-0.02	-0.03	0.19	0.09	0.35	0.29	0.05	-0.06	0.18	0.18
A1198	-0.26	0.10	-0.54	-0.17	-0.01	0.01	0.15	0.06	0.22	0.21	0.03	0.02	0.27	0.29
W1526	-0.22	0.46	0.84	0.00	0.01	-0.06	-0.01	0.11	0.05	0.04	-0.17	-0.14	-0.12	-0.23
W1013	-0.38	0.59	0.63	-0.05	0.01	-0.05	0.09	0.12	0.26	0.18	-0.08	-0.17	-0.09	-0.17
A963	0.13	0.19	-0.15	0.32	0.66	0.39	-0.22	0.00	-0.05	-0.06	-0.32	-0.23	-0.14	-0.05
W963	0.06	0.38	0.16	0.32	0.60	0.29	-0.23	-0.02	-0.10	-0.10	-0.30	-0.21	-0.19	-0.11
W1315	-0.16	-0.29	-0.13	0.38	0.35	0.40	-0.23	-0.20	-0.21	-0.20	-0.01	-0.09	-0.29	-0.11
W1390	-0.16	0.01	0.33	0.30	0.19	0.20	-0.15	-0.14	-0.09	-0.14	0.07	-0.11	-0.38	-0.28
A1315	0.04	-0.31	-0.34	0.38	0.49	0.44	-0.26	-0.16	-0.22	-0.20	-0.12	-0.10	-0.21	-0.04
A1390	-0.06	-0.14	0.08	0.34	0.32	0.33	-0.16	-0.12	-0.01	-0.09	0.02	-0.17	-0.39	-0.28
F963	-0.01	0.23	0.17	0.23	0.31	0.07	-0.30	-0.14	-0.55	-0.35	-0.27	0.09	0.07	0.17
F1390	-0.04	0.15	0.11	0.39	0.39	0.20	-0.30	-0.23	-0.42	-0.34	-0.07	0.00	-0.21	-0.03
F920	-0.38	0.09	0.01	0.37	0.35	0.27	-0.27	-0.24	-0.38	-0.28	-0.06	0.00	-0.15	0.06
F1606	-0.12	-0.42	0.30	-0.11	-0.16	0.10	0.06	0.15	0.11	0.10	-0.15	-0.10	-0.03	-0.15
W867	-0.46	0.05	-0.25	-0.09	0.14	0.26	0.18	0.17	0.51	0.38	-0.09	-0.25	0.04	0.02
A920	-0.17	-0.12	-0.03	0.03	0.37	0.42	0.01	0.23	0.33	0.25	-0.39	-0.34	-0.03	-0.10
A867	-0.18	-0.10	-0.46	-0.10	0.16	0.26	0.18	0.18	0.52	0.38	-0.08	-0.23	0.05	0.02
W920	-0.27	-0.12	0.12	0.04	0.30	0.40	0.02	0.21	0.34	0.25	-0.34	-0.34	-0.10	-0.17
A1013	-0.05	0.27	0.04	0.01	0.41	0.29	0.07	0.25	0.47	0.33	-0.31	-0.37	-0.08	-0.16
F1270	0.12	-0.03	0.18	0.01	0.07	-0.04	-0.20	-0.01	-0.50	-0.27	-0.35	0.14	0.24	0.23
F1656	-0.29	0.12	0.21	-0.23	0.02	0.09	0.16	0.31	0.37	0.34	-0.34	-0.21	0.19	0.02
W1270	-0.25	0.13	0.53	-0.20	-0.13	-0.09	0.01	0.18	-0.10	0.03	-0.37	0.00	0.26	0.11
A1270	-0.40	-0.22	-0.24	-0.24	-0.19	0.01	0.11	0.10	0.03	0.13	-0.14	0.07	0.36	0.32
F1526	-0.72	-0.09	0.31	-0.08	-0.15	0.08	-0.01	0.02	-0.18	-0.02	-0.20	0.04	0.20	0.19
F1160	-0.51	-0.20	0.10	0.00	-0.08	0.07	-0.13	-0.08	-0.43	-0.20	-0.20	0.15	0.24	0.30
F1013	-0.51	0.13	0.33	-0.10	-0.11	-0.06	-0.06	0.02	-0.31	-0.10	-0.27	0.11	0.30	0.28
F1198	0.70	0.08	0.10	-0.08	0.02	-0.23	-0.08	0.09	-0.23	-0.14	-0.21	0.10	0.16	0.05
F1315	0.69	0.00	-0.09	-0.12	-0.07	-0.27	-0.04	0.05	-0.23	-0.14	-0.10	0.17	0.21	0.12
F1180	0.56	0.03	0.16	-0.30	-0.25	-0.28	0.20	0.25	0.34	0.23	-0.01	-0.04	0.06	-0.16
W1180	0.81	0.12	0.04	-0.10	-0.06	-0.25	0.09	0.11	0.21	0.08	0.12	-0.02	-0.11	-0.24
A1180	0.89	-0.06	-0.03	0.00	0.14	-0.07	-0.03	0.12	0.07	-0.01	-0.09	-0.07	-0.10	-0.21
F867	0.22	-0.22	0.47	-0.15	-0.41	-0.25	0.08	0.05	-0.01	-0.02	0.12	0.08	-0.08	-0.22

Figure 14 - Weights obtained for each combination of species and Raman variable calculated from the projection of each given species (yi) over each Raman variable in the hyperspace defined by the six components returned by the Partial

Least Squares (PLS) regression. Raman variables are represented as width (W), area (A) and frequency (F) of the spectral bands. The species better characterised by the PLS model are presented. Species legend as in Figure 13.

Raman variables	Species														
	AVEN	GLGN	GEXL	NCRY	GPAR	NINC	NFON	TTAB	NPAL	FCRO	FCVA	NSBC	PLFR	EOMI	NCTE
W1656	0.11	0.11	0.18	0.14	0.26	-0.11	-0.01	-0.03	0.11	-0.27	-0.14	-0.10	-0.01	0.00	0.00
A1656	0.09	0.09	0.13	0.18	0.27	-0.09	-0.02	0.00	0.09	-0.29	-0.15	-0.09	0.01	0.01	0.03
W1160	0.08	0.01	0.06	0.11	0.23	0.12	0.08	0.09	0.15	-0.24	-0.13	-0.10	-0.01	-0.03	-0.10
A1160	-0.03	0.00	-0.02	0.12	0.12	0.23	0.10	0.18	0.12	-0.03	-0.02	0.03	0.01	0.00	0.07
W1445	-0.07	-0.08	-0.08	-0.09	-0.06	-0.04	0.02	0.05	-0.11	-0.10	-0.11	-0.07	-0.03	-0.02	-0.06
A1445	-0.08	-0.09	-0.10	-0.07	-0.05	0.01	0.03	0.09	-0.11	-0.10	-0.12	-0.07	-0.02	-0.02	-0.05
W1606	-0.06	-0.10	-0.12	-0.04	-0.03	-0.03	-0.02	0.06	-0.12	-0.14	-0.13	-0.07	0.00	-0.02	-0.06
A1606	-0.05	-0.10	-0.13	-0.01	-0.01	-0.02	-0.04	0.07	-0.13	-0.14	-0.16	-0.08	0.00	-0.02	-0.05
F1445	-0.09	-0.08	-0.12	-0.04	-0.06	-0.03	-0.01	0.07	-0.13	-0.10	-0.10	-0.04	0.00	-0.01	-0.01
W1198	-0.02	-0.17	-0.23	-0.02	-0.06	0.08	-0.05	0.06	-0.11	-0.07	-0.10	-0.05	0.03	-0.04	-0.15
A1198	-0.09	-0.16	-0.26	-0.01	-0.09	0.18	0.02	0.14	-0.07	0.00	-0.02	0.03	0.04	-0.02	-0.07
W1526	0.07	0.03	0.08	-0.09	-0.05	-0.22	-0.09	-0.18	-0.06	0.00	0.01	-0.05	-0.02	-0.01	-0.08
W1013	0.09	-0.04	0.00	-0.08	-0.03	-0.17	-0.12	-0.14	-0.10	-0.07	-0.09	-0.11	-0.01	-0.04	-0.16
A963	-0.05	-0.03	0.05	-0.18	-0.09	0.07	0.11	0.06	-0.03	0.07	-0.05	-0.09	-0.07	-0.05	-0.08
W963	-0.02	-0.01	0.10	-0.18	-0.07	-0.03	0.07	0.00	-0.04	0.02	-0.07	-0.12	-0.08	-0.04	-0.09
W1315	0.03	0.09	0.24	-0.04	0.15	0.12	0.21	0.09	0.19	-0.09	-0.07	-0.12	-0.08	-0.03	-0.04
W1390	0.13	0.11	0.28	-0.01	0.19	-0.07	0.06	-0.05	0.12	-0.16	-0.14	-0.18	-0.07	-0.04	-0.07
A1315	-0.02	0.06	0.18	-0.07	0.07	0.17	0.22	0.12	0.15	0.00	-0.03	-0.08	-0.08	-0.03	-0.01
A1390	0.13	0.09	0.27	-0.02	0.18	0.01	0.07	-0.01	0.13	-0.11	-0.15	-0.18	-0.08	-0.05	-0.08
F963	-0.17	0.04	0.04	-0.15	-0.13	0.02	0.21	0.08	0.02	0.04	0.12	0.06	-0.05	0.03	0.08
F1390	-0.03	0.08	0.20	-0.09	0.07	0.00	0.18	0.06	0.09	-0.11	-0.07	-0.10	-0.09	-0.01	0.00
F920	-0.06	0.05	0.16	-0.10	0.07	0.10	0.25	0.12	0.14	-0.14	-0.05	-0.10	-0.08	-0.02	-0.04
F1606	0.05	0.07	0.06	0.00	-0.03	0.00	-0.01	-0.10	0.08	0.16	0.18	0.10	0.02	0.02	0.02
W867	0.03	-0.14	-0.15	-0.06	-0.05	0.15	-0.02	0.04	-0.04	0.02	-0.06	-0.07	0.00	-0.06	-0.20
A920	0.00	-0.07	-0.05	-0.15	-0.15	0.11	0.03	-0.02	-0.01	0.21	0.09	0.00	-0.02	-0.04	-0.12
A867	0.03	-0.14	-0.16	-0.03	-0.06	0.17	-0.04	0.05	-0.05	0.08	-0.05	-0.04	0.02	-0.06	-0.15
W920	0.05	-0.04	0.01	-0.13	-0.10	0.08	0.02	-0.05	0.02	0.16	0.06	-0.03	-0.03	-0.05	-0.14
A1013	0.05	-0.12	-0.08	-0.14	-0.14	0.01	-0.09	-0.07	-0.12	0.12	-0.05	-0.09	-0.02	-0.06	-0.18
F1270	-0.21	0.04	-0.06	-0.12	-0.23	0.02	0.16	0.03	0.00	0.20	0.29	0.22	0.00	0.07	0.16
F1656	-0.03	-0.11	-0.20	-0.12	-0.23	0.02	-0.07	-0.08	-0.10	0.21	0.16	0.10	0.03	-0.01	-0.10
W1270	-0.12	-0.01	-0.13	-0.12	-0.25	-0.05	0.02	-0.08	-0.05	0.21	0.29	0.20	0.03	0.05	0.04
A1270	-0.15	-0.09	-0.24	-0.03	-0.17	0.19	0.10	0.10	0.02	0.14	0.22	0.19	0.05	0.03	0.02
F1526	-0.11	0.00	-0.06	-0.09	-0.11	0.10	0.18	0.04	0.10	0.07	0.23	0.14	0.00	0.03	-0.01
F1160	-0.18	0.03	-0.05	-0.09	-0.12	0.15	0.26	0.11	0.13	0.08	0.27	0.18	-0.01	0.05	0.08
F1013	-0.18	-0.02	-0.13	-0.12	-0.20	0.05	0.16	0.04	0.02	0.10	0.26	0.18	0.01	0.05	0.04
F1198	-0.10	0.02	-0.07	-0.04	-0.19	-0.13	-0.08	-0.07	-0.13	0.20	0.16	0.16	0.03	0.06	0.17
F1315	-0.11	0.01	-0.10	0.01	-0.16	-0.08	-0.07	-0.02	-0.12	0.17	0.14	0.17	0.05	0.06	0.19
F1180	0.08	-0.02	-0.10	0.07	-0.09	-0.18	-0.29	-0.18	-0.17	0.15	0.04	0.07	0.07	0.02	0.05
W1180	0.11	0.01	0.02	0.09	0.02	-0.22	-0.28	-0.15	-0.16	0.04	-0.12	-0.04	0.03	0.01	0.06
A1180	0.05	0.03	0.04	0.02	-0.06	-0.15	-0.18	-0.12	-0.11	0.16	0.00	0.02	0.01	0.02	0.10
F867	0.10	0.12	0.12	0.11	0.07	-0.19	-0.14	-0.17	0.03	0.03	0.07	0.06	0.04	0.04	0.10

Figure 15 - Weights obtained for each combination of species and Raman variable calculated from the projection of each given species (yi) over each Raman variable in the hyperspace defined by the six components returned by the Partial Least Squares (PLS) regression. Raman variables are represented as width (W), area (A) and frequency (F) of the spectral bands. Species less well characterized by the PLS model are presented. Species legend as in Figure 13.

3.4. *Taxa* identification using Raman data

The ANN models generated to predict the different diatom taxonomic levels using Raman variables are presented in Table III of Appendix II. The models with the best accuracy in the test series for each taxonomical level are represented in Table 5.

Table 5 - Categorical target, continuous input variables and data set accuracy of the Artificial Neuronal Network (ANN) models with the highest validation accuracy in the test series. The network architecture used was Multilayer Perceptron (MLP).

Categorical target	**Species**	**Genus**	**Family**	**Order**	**Subclass**
Continuous input	Width Frequency	All	All	All	Width Frequency A1526NN
Train accuracy (%)	50.9	70.1	74.0	84.2	78.3
Test accuracy (%)	32.6	52.6	54.9	58.3	78.9
Validation accuracy (%)	33.7	52.0	52.6	53.1	76.0

The ANN methodology was more accurate in predicting the diatom subclass with a validation accuracy of 76.0%. However, this level encompassed only four subclasses, and more than half of the individuals studied belonged to subclass *Bacillariophycidae (*556 diatoms, see Table 3). Due to this uneven distribution of the data between the subclasses, the second model with the highest validation accuracy (Order, 53.1%) was also considered for further analysis. The model targeting the species *taxa* was the less accurate in the classification with a validation accuracy of 33.7%. However, this model was also detailed because species is the elementary level in diatom taxonomic identification. Table 6 indicates the prediction accuracy for these models diatoms *taxa* with a prediction accuracy >65% are indicated in bold. Sensitivity was also calculated and is presented in Tables IV of Appendix II.

Table 6 - Accuracy (Ac.) per *taxa* of each classification model derived by the Artificial Neuronal Network (ANN) using Raman variables as continuous input variables. Diatoms species, orders and subclasses with a prediction accuracy > 65% are indicated in bold. The accuracy classes, as proposed by the European Centre for the Validation of Alternative Methods (Winter *et al.*, 2008) are indicated by asterisks: * sufficient accuracy (65-74%); ** good accuracy (75-84%), *** excellent accuracy >85%.

Subclass	Ac. (%)	Order	Ac. (%)	Species	Ac. (%)
Bacillariophycidae	**89*** **	*Bacillariales*	56	*Nitzschia amphibia*	0

				Nitzschia fonticola	0
				Nitzschia inconspicua	25
				Nitzschia palea	20
				Nitzschia subcapitellata	25
		Cocconeidales	63	***Achnanthidium exiguum***	**67***
				Achnanthidium minutissimum	**80****
				Achnanthidium straubianum	0
				Planothidium frequentissimum	0
		Cymbellales	63	*Cymbella tumida*	0
				Gomphonema affine	50
				Gomphonema exilissimum	40
				Gomphonema lagenula	0
				Gomphonema parvulum	21
		Naviculales	59	*Navicula cryptocephala*	56
				Navicula cryptotenella	25
				Navicula gregaria	11
				Navicula notha	25
				Eolimna mínima	0
		Thalassiophysales	42	***Amphora pediculus***	**71***
				Amphora veneta	40
Fragilariophycidae	44	*Fragilariales*	47	***Fragilaria crotonensis***	**67***
				Fragilaria vaucheriae	0
				Pseudostaurosira brevistriata	50
		Licmophorales	32	*Ctenophora pulchella*	22
				Tabularia tabulata	17
				Ulnaria Ulna	29
Melosirophycidae	45	*Melosirales*	64	***Melosira varians***	**82***
Thalassiosirophycidae	**75****	*Stephanodiscales*	25	*Cyclotella stelligera*	50

Considering the overall models described in Table 5, only the subclass was predicted with a good performance. The remaining models were predicted with less than sufficient performance. However, Table 6 shows that some diatom Species, orders and Subclasses were predicted with a higher performance than others. *Bacillariophycidae* was the best predicted subclass with an excellent accuracy (89%). *Thalassiosirophycidae* had also a high accuracy (75%) reaching a good overall performance. All the orders were predicted with insufficient accuracy. Nevertheless, the best accuracy was obtained for the prediction of the order *Melosirales.* Furthermore, *A.*

minutissimum and *M. varians* were predicted with good accuracy and *A. exiguum, A. pediculus*, *Fragilaria crotonensis* were predicted with sufficient accuracy.

4. Discussion

All the individuals measured in this study showed Raman spectra composed by fourteen bands: around 867 cm^{-1}, 920 cm^{-1}, 963 cm^{-1}, 1013 cm^{-1}, 1160 cm^{-1}, 1180 cm^{-1}, 1198 cm^{-1}, 1270 cm^{-1}, 1315 cm^{-1}, 1390 cm^{-1}, 1445 cm^{-1}, 1526 cm^{-1}, 1606 cm^{-1} and 1656 cm^{-1}. Similar spectra were obtained for *Thalassiosira pseudonana* and *Ditylum brightwellii* using an incident laser with 750 nm wavelength, 30 mW potency and 30 and 2 seconds of acquisition time, respectively (Meksiarun *et al.*, 2015; Rüger *et al.*, 2016). The species *Cylindrotheca closterium* also produced a similar spectrum with a laser wavelength similar to the one used in this study (532 nm), a potency of 0-0.1 mW and an acquisition time of one second (Pinzaru *et al.*, 2016). Of all the fourteen bands detected, six bands as accounting for the most important Raman variables in the Partial Least Squares regression: 1013, 1160, 1180, 1198, 1270 and 1526 cm^{-1}. The bands 1013, 1160, 1180 and 1526 cm^{-1} can be assigned to C-CH3 in plane rocking modes (Premvardhan *et al.*, 2009; Rüger *et al.*, 2016; C-C (Premvardhan *et al.*, 2009; Rüger *et al.*, 2016), C-H (Rüger *et al.*, 2016) and C=C stretching modes (Premvardhan *et al.*, 2009; Rüger *et al.*, 2016) from carotenoids, respectively. The band 1198 cm^{-1} can be assigned to N-C stretching modes of the chlorophyll *a* (Rüger *et al.*, 2016). It is known that, pigment composition is similar between the diatom species, however, the ratio between the pigments is highly variable (Kuczynska *et al.*, 2015) as well as the concentrations of these molecules in different cell compartments (Heraud *et al.*, 2007). This may explain the differences among species in these pigment-related bands. Variations in the area of the pigment bands might be related with the amount of these compounds (Pinzaru *et al.*, 2016). The area of the band 1160 cm^{-1} was lower for *A. pediculus* and *A. minutissimum* two smaller and pioneer *taxa* capable of colonizing baring substrates and resist to environmental adversities (Rimet & Bouchez, 2012). The lower amounts of carotenoids reflect this resistance due to the fact that the production of carotenoids consists on a defence mechanism to prevent oxidative stress (Heraud *et al.*, 2007). The different values of the width of the bands might reflect pigment diversity (Premvardhan *et al.*, 2009). Contrarily to the area of 1160 cm^{-1}, the width of the bands 1013 and 1526 cm^{-1} showed higher values in *Achnanthidium minutissimum* and *A. pediculus* than in the remaning species. This may indicate the presence of a higher variety of carotenoids, especially when it

comes to the band 1526 cm^{-1}, a marker of the length of the polyene chain, which vary among different carotenoids (Merlin, 1985). The width of the band 1160 cm^{-1} was lower in *Cyclotella stelligera*, than in the other species, probably reflecting the presence of a lower variety of carotenoids. The frequency of some bands was also relevant to discriminate the species analyzed. In particular, the frequency of bands 1526 cm^{-1}, 1180 cm^{-1} and 1198 cm^{-1}. In diatom studies, variations of frequency are related to resonance phenomena caused by changes in the wavelength of the incident laser. This resonance phenomena occurs when the energy of the incident laser is similar to the energy of the transition of a determined compound causing the enhancement of the band corresponding to that compound. Frequency differences of the pigment bands in solution can also be derived from conformational changes due to the polarity of the solvent (Premvardhan *et al.*, 2009; Premvardhan *et al.*, 2010). In this study pigments were not extracted, so no solvent was used, and the incident laser was set to the same parameters in acquisitions. Hence, changes in frequency are most probably due to the presence of different molecular conformations in the measured cells. Variations in the frequency of the band 1270 cm^{-1} can be related to conformational changes of amide III, $=CH_2$ bending modes or nucleic acids (Notinger, 2007). Future studies should focus also on elucidating the differences in molecular conformations that could underlying the frequency shifts found in the diatom species. It is evident, however, that the species *C. stelligera* stood out in profile from the other species. This particular species is non-motile and planktonic contrarily to the other species described (Rimet & Bouchez, 2012). In consequence, metabolic and molecular adaptations can occur in these cells in response to challenging environmental conditions. Such responses would be reflected in the variations of the Raman bands considered. Variation in bands 1160, 1180 and 1198 cm^{-1} can also be assigned to C=S modes of the frustule (De Tommasi, 2016; De Tommasi *et al.*, 2018). Similarly to this study, some authors found differences among genus in bands related to the frustule components, which may reflect differences in frustule silification (Yuan *et al.*, 2004). Overall, interpretation of the PLS by calculation of vector modules and projection of vector species (y_i) on Raman variables (x_i) in the hyperspace defined by the six significant components obtained helped to obtain the Raman profile of each species, bringing information crucial to characterize the species and their physiological status.

Artifical Neural Networks have previously been used to predict *taxa* from diatoms (Pedraza *et al.*, 2018) and different organisms such as other algae and phytoplankton *(*Boddy *et al*., 2000; Medlin, 2012*)*, macroinvertebrates (Joutsijoki *et al.*, 2014), fungi (Naumann, 2009) and even plants (Wolski & Kruk, 2020). However, to our knowledge,

this is the first study concerning the prediction of diatom *taxa* from Raman spectral data. Through our results, the ANN model was effective in predicting diatom subclass with an accuracy of 76%, within category *good* of ECVAM classification (Winter *et al.*, 2008), followed by order prediction with an accuracy of 53.1%. The remaining taxonomical levels considered were predicted with lower accuracy. Nevertheless, within each subclass, order and species it is evident that some levels are predicted with higher performance than others. These results suggest the abundance of *taxa* could be interfering with the performance of the ANN model (Dedecker *et al.*, 2002; Fielding & Bell, 1997; Manel *et al.*, 1999). Indeed, other authors have found that when a *taxa* is rare models tend to learn that the *taxa* is always absent. Conversely, when a *taxa* is common models tend to learn that the *taxa* is always present (Dedecker *et al.*, 2004). In this work, each species differed slightly regarding their representation in the dataset. However, the differences in representation were higher when considering the number of species within higher *taxa* levels; some *taxa* contain many species and others only a few. This may be a source of bias in the analysis. Further studies using a wider sample with a more even distribution of species can help clarify this effect and minimize such interferences. The laser wavelength used for Raman measurements in this study is commonly used for the study of pigments in other diatom studies (Pinto *et al*., submitted and references herein). Raman bands associated with pigments have high sensitivity to light conditions (Alexandre *et al.*, 2014) and nutrients (Rüger *et al*., 2016). Therefore, Raman data of pigments might not be conservative enough for *taxa* identification using ANN models. Further studies based on other cell components (p.e. frustule) can be done in order to try to improve the disparity registered in ANN accuracy results *per taxa*.

In another diatom study, an ANN based on images belonging to ten different diatom species was used for *taxa* prediction (Pedraza *et al.*, 2018). The study adopted two approaches: pixelwise in which each pixel was considered as a unit of detection and objectwise in which each individual was considered as a unit of detection. The pixelwise method showed 68% sensitivity, while the objecwise method showed a sensitivity of 13%. Similarly to the present work, some species were better identified than others (Pedraza *et al.*, 2018). Comparatively, some species were identified by the present ANN Raman models with a sensitivity in between the pixelwise and the objectwise methods, as for example *Nitzschia inconspicua* (50% sensitivity *vs* 70% for the pixelwise method and 21% for the objectwise method) and *Nitzschia palea* (29% sensitivity vs 75% for the pixelwise method and 1% for the objectwise method) (Pedraza *et al.*, 2018). Overall, the species identified herein with higher sensitivity were: *Achananthidium exiguum* (67%),

Fragilaria crotonensis (67%), *Amphora pediculus* (71%), *Achananthidium minutissimum* (80%) and *Melosira varians* (82%). The higher sensitivity values obtained in our study are comparable to those obtained for the pixelwise method by Pedraza and colleagues. In this work, only two species – *Nitzschia capitellata* (87%) and *Staurosira venter* (83%) – had better sensitivity than *Melosira varians* (Pedraza *et al.* 2018). Our results thus suggest that Raman spectroscopy can be applied to identification of some species and taxonomic levels higher than the species, with consistent results. A most interesting characteristic of these Raman identification models is the high accuracy and sensitivity obtained when considering that the Raman spectra acquired reflect the physiological status of the diatoms rather than their morphological characteristics. Given the potential shown by the present results, a larger study including species from different geographical locations living under a diverse range environmental conditions will provide a sound dataset for ANN spot the common intrinsic profile of each species, improving species identification.

5. Conclusion

In conclusion, most Raman bands were found to differ among species. The Partial Least Squares regression allowed to depict a Raman profile for each species than can be used to better understand the physiological status of the different species. The Artificial Neural Networks models could predict better the diatom subclasses and order than the species, with accuracy varying from sufficient to excellent (67-89%). The model used to predict species showed a lower overall accuracy, however, some species were predicted with good results. This work also showed the constraints associated with morphological taxonomical identification can be easily eliminated for some species and higher taxonomical levels.

Acknowledgements

The authors would like to thank the EU and FCT/UEFISCDI/FORMAS for funding, in the frame of the collaborative international consortium REWATER, financed under the ERA-NET Cofund WaterWorks2015 (Water JPI). This research was also supported by national funds through FCT (Portuguese Foundation for the Science and Technology) within the scope of UIDB/04423/2020 and UIDP/04423/2020. The authors would also

like to thank to NECL (Network of Extreme Conditions Laboratories) for providing the Raman equipment necessary.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abbas, A., Josefson, M., & Abrahamsson, K. (2011). Characterization and mapping of carotenoids in the algae Dunaliella and Phaeodactylum using Raman and target orthogonal partial least squares. *Chemometrics and Intelligent Laboratory Systems, 107*(1), 174-177.

Alexandre, M. T., Gundermann, K., Pascal, A. A., van Grondelle, R., Buchel, C., & Robert, B. (2014). Probing the carotenoid content of intact Cyclotella cells by resonance Raman spectroscopy. *Photosynthesis Research, 119*(3), 273-281.

Almeida, S. F. P., & Gil, M. C. P. (2001). dÉcologie des diatomées d'eau douce de la région centrale du Portugal. *Cryptogamie Algologie, 22*(1), 109-126.

Bishop, C. M. (1995). *Neural networks for pattern recognition*: Oxford university press.

Boddy, L., Morris, C., Wilkins, M., Al-Haddad, L., Tarran, G., Jonker, R., & Burkill, P. (2000). Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. *Marine Ecology Progress Series, 195*, 47-59.

CEMAGREF, M. (1982). Etude des méthodes biologiques d'appréciation quantitative de la qualité des eaux. *Rapport Cemagref QE Lyon-AF Bassin Rhône Méditerranée Corse*.

Chen, L., Weng, D., Du, C., Wang, J., & Cao, S. (2019). Contribution of frustules and mucilage trails to the mobility of diatom Navicula sp. *Scientific reports, 9*(1), 1-12.

Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy, (2000).

De Tommasi, E. (2016). Light manipulation by single cells: the case of diatoms. *Journal of Spectroscopy, 2016*.

De Tommasi, E., Congestri, R., Dardano, P., De Luca, A. C., Managò, S., Rea, I., & De Stefano, M. (2018). UV-shielding and wavelength conversion by centric diatom nanopatterned frustules. *Scientific reports, 8*(1), 1-14.

Dedecker, A. P., Goethals, P. L., & De Pauw, N. (2002). Comparison of artificial neural network (ANN) model development methods for prediction of macroinvertebrate communities in the Zwalm river basin in Flanders, Belgium. *TheScientificWorldJOURNAL, 2*.

Dedecker, A. P., Goethals, P. L., Gabriels, W., & De Pauw, N. (2004). Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). *Ecological Modelling, 174*(1-2), 161-173.

Drobatz, K. J. (2009). Measures of accuracy and performance of diagnostic tests. *Journal of Veterinary Cardiology, 11*, S33-S40.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, 38-49.

Germain, H. (1981). *Flore Des Diatomees: Diatomophycees: Eaux Douces Et Saumatres Du Massif Armoricain Et Des Contrees Voisines D'Europe Occidentale*: Societe Nouvelle des Editions Boubee.

Heraud, P., Wood, B. R., Beardall, J., & McNaughton, D. (2007). Probing the Influence of the Environment on Microalgae Using Infrared and Raman Spectroscopy. In *New Approaches in Biomedical Spectroscopy* (Vol. 963, pp. 85-106): American Chemical Society.

INAG. (2008). Protocolo de amostragem e análise para o fitobentos-diatomáceas In *Manual para a Avaliação Biológica da Qualidade da Água em Sistemas Fluviais Segundo a Diretiva do Quadro da Água*.

Joutsijoki, H., Meissner, K., Gabbouj, M., Kiranyaz, S., Raitoharju, J., Ärje, J., . . . Juhola, M. (2014). Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecological Informatics, 20*, 1-12.

Kuczynska, P., Jemiola-Rzeminska, M., & Strzalka, K. (2015). Photosynthetic pigments in diatoms. *Marine drugs, 13*(9), 5847-5881.

Langlotz, C. P. (2003). Fundamental measures of diagnostic examination performance: usefulness for clinical decision making and research. *Radiology, 228*(1), 3-9.

Laviale, M., Beaussart, A., Allen, J., Quilès, F., & El-Kirat-Chatel, S. (2019). Probing the Adhesion of the Common Freshwater Diatom Nitzschia palea at Nanoscale. *ACS applied materials & interfaces, 11*(51), 48574-48582.

Lear, G., Dopheide, A., Ancion, P. Y., Roberts, K., Washington, V., Smith, J., & Lewis, G. (2012). Biofilms in freshwater: their importance for the maintenance and monitoring of freshwater health. *Microbial Biofilms: Current Research and Applications*, 129-151.

Manel, S., Dias, J.-M., Buckton, S., & Ormerod, S. (1999). Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology, 36*(5), 734-747.

Matos, A. I. (2014). *Development of molecular tools for the early warning of potentially toxic cyanobacteria.* (Master), University of Porto, Porto, Portugal.

Medlin, L. K. J., M. (2012). Artificial neural networks contribute to the identification of cryptomonad taxa. *Vie et milieu-life and enVironment, 62*(3), 121-127.

Meksiarun, P., Spegazzini, N., Matsui, H., Nakajima, K., Matsuda, Y., & Sato, H. (2015). In vivo study of lipid accumulation in the microalgae marine diatom Thalassiosira pseudonana using Raman spectroscopy. *Applied spectroscopy, 69*(1), 45-51.

Mendes, T., Almeida, S. F., & Feio, M. J. (2012). Assessment of rivers using diatoms: effect of substrate and evaluation method. *Fundamental and Applied Limnology/Archiv für Hydrobiologie, 179*(4), 267-279.

Merlin, J. C. (1985). Resonance Raman spectroscopy of carotenoids and carotenoid-containing systems. *Pure and Applied Chemistry, 57*(5), 785-792.

Morais, J. (2009). *Avaliação do Risco de Ocorrência de Cianobactérias Tóxicas nos Lagos do Parque da Cidade do Porto.* (Master), University of Porto, Porto, Portugal.

Morin, S., Gómez, N., Tornés, E., Licursi, M., & Rosebery, J. (2016). Benthic diatom monitoring and assessment of freshwater environments: standard methods and future challenges. *Aquatic Biofilms*, 111.

Naumann, A. (2009). A novel procedure for strain classification of fungal mycelium by cluster and artificial neural network analysis of Fourier transform infrared (FTIR) spectra. *Analyst, 134*(6), 1215-1223.

Notingher, I. (2007). Characterisation using Raman micro-spectroscopy. In *Tissue Engineering Using Ceramics and Polymers* (pp. 248-266): Elsevier.

Pandey, L. K., Bergey, E. A., Lyu, J., Park, J., Choi, S., Lee, H., . . . Han, T. (2017). The use of diatoms in ecotoxicology and bioassessment: insights, advances and challenges. *Water Research, 118*, 39-58.

Parker, F. S. (1983). *Applications of infrared, Raman, and resonance Raman spectroscopy in biochemistry*: Springer Science & Business Media.

Pedraza, A., Bueno, G., Deniz, O., Ruiz-Santaquiteria, J., Sanchez, C., Blanco, S., . . . Cristobal, G. (2018). *Lights and pitfalls of convolutional neural networks for diatom identification.* Paper presented at the Optics, Photonics, and Digital Technologies for Imaging Applications V.

Pinto, R., Mortágua, A., Almeida, S. F., Serra, S., & Feio, M. J. (2020). Diatom size plasticity at regional and global scales. *Limnetica, 39*(1), 387-403.

Pinzaru, S. C., Müller, C., Tomšić, S., Venter, M. M., Brezestean, I., Ljubimir, S., & Glamuzina, B. (2016). Live diatoms facing Ag nanoparticles: surface enhanced Raman scattering of bulk cylindrotheca closterium pennate diatoms and of the single cells. *RSC advances, 6*(49), 42899-42910.

Potapova, M., & Hamilton, P. B. (2007). Morphological and ecological variation within the Achnanthidium minutissimum (Bacillariophyceae) species complex 1. *Journal of phycology, 43*(3), 561-575.

Premvardhan, L., Bordes, L., Beer, A., Buchel, C., & Robert, B. (2009). Carotenoid structures and environments in trimeric and oligomeric fucoxanthin chlorophyll a/c2 proteins from resonance Raman spectroscopy. *J Phys Chem B, 113*(37), 12565-12574.

Premvardhan, L., Robert, B., Beer, A., & Büchel, C. (2010). Pigment organization in fucoxanthin chlorophyll a/c2 proteins (FCP) based on resonance Raman spectroscopy and sequence analysis. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1797*(9), 1647-1656.

Pytlik, N., Klemmed, B., Machill, S., Eychmüller, A., & Brunner, E. (2019). In vivo uptake of gold nanoparticles by the diatom Stephanopyxis turris. *Algal research, 39*, 101447.

Rimet, F., & Bouchez, A. (2012). Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. *Knowledge and management of Aquatic Ecosystems*(406), 01.

Round, F. E., Crawford, R. M., & Mann, D. G. (2007). *Diatoms: biology and morphology of the genera*: Cambridge university press.

Rüger, J., Mondol, A. S., Schie, I. W., Popp, J., & Krafft, C. (2019). High-throughput screening Raman microspectroscopy for assessment of drug-induced changes in diatom cells. *Analyst, 144*(15), 4488-4492.

Rüger, J., Unger, N., Schie, I. W., Brunner, E., Popp, J., & Krafft, C. (2016). Assessment of growth phases of the diatom Ditylum brightwellii by FT-IR and Raman spectroscopy. *Algal research, 19*, 246-252.

Squires, L. E., Rushforth, S. R., & Brotherson, J. D. (1979). Algal response to a thermal effluent: study of a power station on the provo river, Utah, USA. *Hydrobiologia, 63*(1), 17-32.

UNESCO-WHO-UNEP. (1996). *Water Quality Assessments*. Cambridge, UK: Chapman & Hall.

Vilbaste, S., & Truu, J. (2003). Distribution of benthic diatoms in relation to environmental variables in lowland streams. *Hydrobiologia, 493*(1-3), 81-93.

Winter, M. J., Redfern, W. S., Hayfield, A. J., Owen, S. F., Valentin, J.-P., & Hutchinson, T. H. (2008). Validation of a larval zebrafish locomotor assay for assessing the seizure liability of early-stage development drugs. *Journal of pharmacological and toxicological methods, 57*(3), 176-187.

Wolski, G. J., & Kruk, A. (2020). Determination of plant communities based on bryophytes: The combined use of Kohonen artificial neural network and indicator species analysis. *Ecological indicators, 113*, 106160.

Wood, B. R., Heraud, P., Stojkovic, S., Morrison, D., Beardall, J., & McNaughton, D. (2005). A portable Raman acoustic levitation spectroscopic system for the identification and environmental monitoring of algal cells. *Analytical chemistry, 77*(15), 4955-4961.

Wu, Q., Nelson, W., Treubig, J., Brown, P., Hargraves, P., Kirs, M., . . . Hanlon, E. (2000). UV resonance Raman detection and quantitation of domoic acid in phytoplankton. *Analytical chemistry, 72*(7), 1666-1671.

Yuan, P., He, H. P., Wu, D. Q., Wang, D. Q., & Chen, L. J. (2004). Characterization of diatomaceous silica by Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 60*(12), 2941-2945.

Chapter 4 - Environmental diagnosis with Raman Spectroscopy applied to diatoms

Luís Oliva-Teles[2], Raquel Pinto[2], Rui Vilarinho[3], António Paulo Carvalho[2], J. Agostinho Moreira[3], Laura Guimarães[1],

[1] CIIMAR/CIMAR - Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Avenida General Norton de Matos, s/n 4450-208 Matosinhos, Portugal
[2] Department of Biology, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n, 4169-007, Porto, Portugal
[3] Department of Physics, Faculty of Sciences of the University of Porto, Rua do Campo Alegre, s/n. 4169-007, Porto, Portugal

Abstract

Water quality in freshwater ecosystems has been changing due to ever greater use of water resources and the contamination load resulting from human activities. Therefore, it is important to develop fast and efficient methodologies to monitoring water quality. Diatoms are good bioindicators recommended for water quality assessment around the world. However, the traditional indexes used for this purpose are based on a taxonomical identification that requires significant expertise and is very time-consuming.
Raman spectroscopy is a simple, fast and label-free technique that can be applied to environmental diagnosis using diatoms. To test this hypothesis, Raman spectra were obtained from the most abundant species in three similar lakes of a city park located in the North of Portugal (Porto). Fourteen bands were found in the Raman spectra acquired. In general, there were significant differences in the area, width and frequency of diatom Raman bands among lakes or/and species common to the three lakes. Lake diagnosis was performed using an Artificial Neural Network (ANN). Diagnose accuracy was 89% when using only Raman data as input for the ANN. Accuracy increased to 96% when using Raman and *taxa* data as input for the ANN. The sensitivity obtained for both models is comparable to most clinical tests, including the recent ones developed for COVID-19, which varies between 71% and 98%. The method lays an important

foundation for future environmental studies aiming at assessing freshwater systems, to be replicated at larger scales and to varied geographic settings.

Keywords: Spectra, Environmental health, Artificial Neural Networks (ANN), Pigments, Frustule, Lipids,

Introduction

The increase in the human population enhances the necessity for producing goods and services, resulting in greater water use and pollution. The consequent decrease in water quality affects domestic, industrial, agricultural and recreative activities, fisheries and aquaculture production and the health of ecosystems (Boyd, 2019). Thus, it is necessary to improve methodologies for a faster and more efficient diagnose and monitoring of water quality.

Diatoms are well-recognized bioindicators and have been used to assess water quality for decades (Patrick, 1973; Dixit *et al.*, 1999; Blanco *et al.*, 2004; Desrosiers *et al.*, 2013). These microalgae are distributed all over the world in practically all kinds of aquatic systems (Round *et al.*, 2007), have fast responses to environmental changes and different species have different tolerance ranges (Almeida & Gil, 2001; Squires *et al.*, 1979; Vilbaste & Truu, 2003). According to the Water Framework Directive (WFD), diatoms are a biological element of the aquatic flora and, therefore, obligatory for the ecological assessment of water courses in European countries (European Comission Council, 2000). Diatoms are also used in routine water monitoring surveys in other parts of the world such as Canada, USA, Japan, South America and Australia (UNESCO-WHO-UNEP, 1996).

The commonly used methods for assessing water quality using diatoms are based on taxonomic identification and valve counting (Pandey *et al.*, 2017; Pinto *et al.*, 2020). The species abundance is then used in the calculation of autoecological indexes such as, for example, the IPS - Indice de Polluosensibilité Spécifique (CEMAGREF, 1982) or the IBD - Indice Biologique Diatomées (Coste *et al.*, 2009). This methodology is very practical in the case of diatoms due to the fact that they are easy to sample, process (Lear *et al.*, 2012) and store (Mendes *et al.*, 2012). Additionally, they possess a detailed siliceous wall – the frustule – which shape and patterns differ among species enabling the taxonomic identification at the species level (Germain, 1981; Pandey *et al.*, 2017). However, this approach has several drawbacks; it is time-consuming, *taxa* are

recurrently changing, and taxonomical identification requires years of expertise (Morin *et al.*, 2016). Furthermore, the ecological indexes derived are often region-specific, though they are frequently employed across different regions and, consequently, applied inaccurately (Morin *et al.*, 2016; Pinto *et al.*, 2020).

Raman spectroscopy (RS) is a promising alternative technique with high potential for application to diatoms in order to diagnose and monitor water quality (Pinto *et al.*, *submitted*). In biological analysis, this technique has many advantages in relation to other methods because it is label-free, water interference can be minimized (Parker, 1983) and it requires minimal or no sample preparation and processing (Heraud, *et al.*, 2007; Dantas *et al.*, 2020). In diatoms, RS has been used to understand the structure, conformation and location of a variety of substances such as pigments (Alexandre *et al.*, 2014; Premvardhan *et al.*, 2009; Premvardhan *et al.*, 2010), lipids (Meksiarun *et al.*, 2015), siliceous frustule (Yuan *et al.*, 2004), extracellular polymeric substances (EPS) (Laviale *et al.*, 2019), mucilage (Chen *et al.*, 2019) and toxins (Wu *et al.*, 2000) and their variation with some abiotic factors. For example, bands characteristic of carotenoids in diatom *Cyclotella meneghiniana* vary with light conditions due to different carotenoids produced and their conformation (Alexandre *et al.*, 2014). Also, in the diatom *Thalassiosira pseudonana* bands associated with fatty acids vary with high carbon dioxide levels due to the enhancement of their production (Meksiarun *et al., 2015)*. Raman Spectroscopy was also used in toxicological assays with *Phaeodactylum tricornutum,* to discriminate diothreitol effects elicited by high light and low light conditions (Rüger *et al.*, 2019), and with *Stephanopyxis turris* to study the mechanisms of incorporation of gold nanoparticles in the cell (Pytlik *et al.*, 2019).

Despite the relatively high number of RS studies applied to diatoms, their scope is directed to biotechnological applications or to obtain fundamental knowledge, with no sound demonstration of its utility in environmental diagnose. Nonetheless, RS readings of the structure, conformation and location of diatom constituents, conjugated with chemometric techniques, can provide a solid foundation to carry out diagnose tests with high sensitivity. Additionally, it is known that RS bands can vary with *taxa* eliminating the constraint associated with traditional taxonomical identification (Abbas *et al.*, 2011; Wood *et al.*, 2005; Yuan *et al.*, 2004). The main objective of this work was to develop a diagnose test of environmental quality using diatom RS. For that purpose, RS bands were obtained from diatoms sampled in three lakes located in a natural city park in Northern Portugal. The data were then analysed using partial least squares (PLS) and

with artificial neural network analysis (ANN) to obtain classification models discriminating the three lakes. Finally, model accuracy and sensitivity were evaluated.

2. Materials and Methods

2.1. Sampling Sites

Sampling was carried out in three lakes of Oporto City Park. This is a recreative park located in the North of Portugal, near Matosinhos beach. This park has a green area of 83 ha with four artificial lakes, numerated according to the proximity to the north entrance of the park. Water circulates successively from Lake 1 to Lake 4 and these lakes supply the irrigation system of the park (Morais, 2009). The fauna of this park encompasses numerous species of birds and fishes, and the flora is composed of various tree and shrub species. In terms of phytoplankton, some species of cyanobacteria were detected in these lakes (*Microcystis* sp*., M.aeruginosa* and *Planktothrix* sp*. Cylindrospermopsis racirborskii, Planktothrix agardhii*) (Matos, 2014; Morais, 2009). Sampling was done in three different points belonging to Lake 1, Lake 2 and Lake 3 as represented in Figure 8. These lakes were selected for their very similar hydrodynamic and physico-chemical characteristics. The rational for the selection was based on the reasoning that this is the worst case scenario to test the discriminatory power of this technological approach (RS applied to diatoms), and thus the best suitable to demonstrate its diagnose abitity.

2.2. Field Sampling

Biofilm samples were collected according to the protocol developed by the Portuguese Water Institute (INAG, 2008) with a few adaptions depending on the availability of biofilm substrates on sampling sites. Natural and artificial substrates were scrapped with a toothbrush preferring an area of about 100 cm^2. Alternatively, biofilm was pipetted from the same area of the substrate surface with a syringe. The sampled biofilm was suspended in lake water. In all the lakes sampled, biofilm was collected from rocks, wood, sediment, macrophytes or artificial substrates such as bricks or plastic. The use of different substrates has no influence on the quality of the sample as proven on previous studies (Mendes et al., 2012).

A total of ten mixed biofilm samples were collected in each lake: nine plastic flasks (120 mL) and one dark glass bottle (60 mL). Lake water for chemical parameters was also sampled into 500 mL bottles for each lake. Samples were temporarily stored in a thermal box. Biofilm samples on the plastic flasks, and water samples, were stored at -80ºC until further use. Biofilm samples in dark glass bottles were preserved on formaldehyde (33%). For all the analyses, the biofilm and water samples were defrosted at 4ºC.
Conductivity, pH, temperature, as well as dissolved oxygen saturation and concentration were also measured in three sites from each lake using a multiparametric meter (HQ40D, Hach USA).

2.3. Chemical Water Quality

Concentration of several ions was quantified on three water aliquotes from each lake: nitrate (NO_3^-), nitrite (NO_2^-), ammonia (NH_3), phosphate (PO_4^{3-}) and silica (SiO_2). Concentrations of NO_3^- and NO_2^- in each sample were determined using a NANOCOLOR® 500D spectrophotometer (Macherey Nagel, Germany) with Tests 0-64 (NANOCOLOR® Nitrate 50) and 0-69 (NANOCOLOR® Nitrite 4), respectively. Concentrations of NH_3, PO_4^{3-} and SiO_2 in each sample were determined using a Hanna C214 spectrometer (Hanna Instruments, United States of America) using the reagent sets HI93700-01, HI93713-01 and HI93705-01, respectively. A Water Quality Index (WQI; Brown *et al.*, 1970) was calculated using the Dissolved Oxygen (%), pH, Temperature (ºC), Total Phosphate (mg/L) and Nitrate (mg/L). The concentration of NO_2^- was always below the limit of detection of the equipment used.

2.4. Taxonomic identification

Diatom identification was carried out in the samples preserved in formaldehyde. For this, the organic matter in the samples was oxidized with nitric acid following the protocol of the Water Portuguese Institute (INAG, 2008). This process consisted on adding 10 mL of nitric acid and some crystals of potassium dichromate to each sample. Samples were oxidized for 24 hours at room temperature and then centrifuged at 1500 rpm for five minutes at room temperature using a Kubota 2420 centrifuge (Kubota Corporation, Japan). The supernatant was discarded, and the pellet was resuspended in distilled water. This process was repeated several times for each sample until the oxidants were totally removed. Samples were diluted in water to decrease turbidity and reach an ideal

number of cell frustules in the optical field enabling taxonomic identification. Each diluted sample was dropped on microscope lamella and then the frustule density was confirmed using a light microscope (Zeiss Primo Star, 40x, N.A.= 0.65). Microscope lamella were dried at room temperature until the water was totally evaporated. Permanent slides were mounted using *Naphrax* ® (Brunel Microscopes, Ltd.). Diatom taxonomic identification was carried out under a light microscope (Zeiss Primo Star, 100x, N.A.=1.25) consulting diatom floras (Germain, 1981). A total of 400 valves were identified in each sample. The valve counts of each lake were used for the calculation of IPS (CEMAGREF, 1982) using the software OMNIDIA 5.5. (Lecointe *et al.*, 1993). IPS values were interpreted according to Rimet (2012).

2.5. Raman Spectroscopy

Raman readings were done in the samples kept at -80ºC. For this, biofilm samples were dropped on microscope slides. Then, the slides were dried at room temperature until water was fully evaporated. This prevented the movement of the valves due to water evaporation during RS readings. All readings were performed using an InVia™ Qontor® confocal Raman microscope (Renishaw, United Kingdom) assembled with a Leica DM2700 microscope (Ernst Leitz GmbH, Germany) with a 50x objective used to focus the laser beam onto each diatom in the sample. The incident laser was a Cobolt 04-01 Series Samba™ (Hübner Photonics, Germany) with 532 nm regulated to 0.1 mW. The spectra were obtained with an acquisition time of 10s and 3 accumulations.
According to the results of the taxonomic identification, for each lake a total of 18 spectra were obtained per diatom species showing an abundance >1%. Only species common to the three lakes under analysis were used for spectra reading. Rare *taxa* were excluded from the analysis due to the difficulty in identifying specimens in the non-oxidized samples. Cells were read in the region between the central area and the apex excluding the raphe area and including the chloroplast. Raman Spectra were acquired with the software WiRE™ 5.2. (Renishaw Inc., UK) and the area, width and frequency of the bands were estimated with the oscillatory function in the software Igor Pro™ (Wavemetrics Inc., 1998).

2.6. Data Analysis

All the statistical and data treatment described below were done with the software Statsoft Statistica™ 64 (Statsoft Software Inc., 2014). A data normalization of the areas of the bands was first carried out to correct for intensity fluctuations in the spectra obtained, a common procedure in Raman data analysis. To achieve the most appropriate normalization of data, correlations between the three Raman variables (area, width and frequency) obtained were first investigated in the whole raw dataset. The area of Raman band 1526 cm^{-1} showed the highest number of significant correlations with other variables; it was correlated with most variables and was thus selected for the normalization. After normalization, those correlations strongly decreased. Normalization by this area was previously employed in other studies (Alexandre *et al.*, 2014; Premvardhan *et al.*, 2009).

A PLS was carried out on the normalized data to describe the studied lakes and identify the band components explaining the highest variability in the dataset. Patterns in Raman variables were also investigated through a Cluster analysis performed with the *x loadings* of the PLS components obtained to infer about relationships in the dataset. Differences between the lakes, the common species in the three lakes and the interactions between these two factors were subsequently investigated with a two-way ANOVA followed by a Tukey HSD test.

For lake diagnosis, a classification analysis was done with an ANN with supervised learning. The network architecture used was Multilayer Perceptron (MLP) with operations organized into a network with three dimensions: input layer, hidden layer and output layer (Bishop, 1995). Through this procedure, each neuron performs a weighted sum of its inputs and passes it through a transfer function to produce an output (Bishop, 1995). For this, the normalized dataset was randomly subdivided into three subsets: the training set for preliminary adjustment of the model to the data, the testing set to calibrate the model and the validation set to validate the model (Drobatz, 2009; Langlotz, 2003). The ANN was performed using the lakes as target, the Raman variables as a continuous input or the Raman variables combined with the taxa as a categorical input. Different model combinations were produced to investigated which ones could provide the best classification with the least number of variables, *e.g.* with and without the non-normalized area of band 1526 cm^{-1}, with all the abundant species or just the common species in the three lakes. The performance of the ANN models obtained was evaluated by calculating the accuracy and sensitivity rates. As a strategy for creating the predictive models, the automated network search performed by Statistica™ was used.

3. Results

3.1. Water Chemistry

The graphs of Figure 16 show the mean temperature (ºC), pH, dissolved oxygen (mgL^{-1}), conductivity (μScm^{-1}), salinity (‰) and the mean concentrations in mgL^{-1} of NH_3, NO_3^-, PO_4^{3-}, SiO_2 measured in the replicates of the three lakes studied. Lake 1 and Lake 3 were similar in physico-chemical characteristics, whereas Lake 2 showed a slightly different profile (Figure 16). Lake 1 had the lowest conductivity, mean water temperature and concentration of NH_3, as well as the highest concentration of NO_3^-, PO_4^{3-} and SiO_2. Lake 3 had the highest dissolved oxygen levels and pH, as well as the lowest SiO_2 concentration. In contrast, Lake 2 exhibited the lowest dissolved oxygen and pH, the lowest NO_3^- and PO_4^{3-} concentrations, and the highest water temperature and NH_3 concentration. According to the WQI, Lake 2 exhibited the lowest water quality, rated as medium water quality. For lake 1 and lake 3, the WQI indicated a good water quality (Table 7).

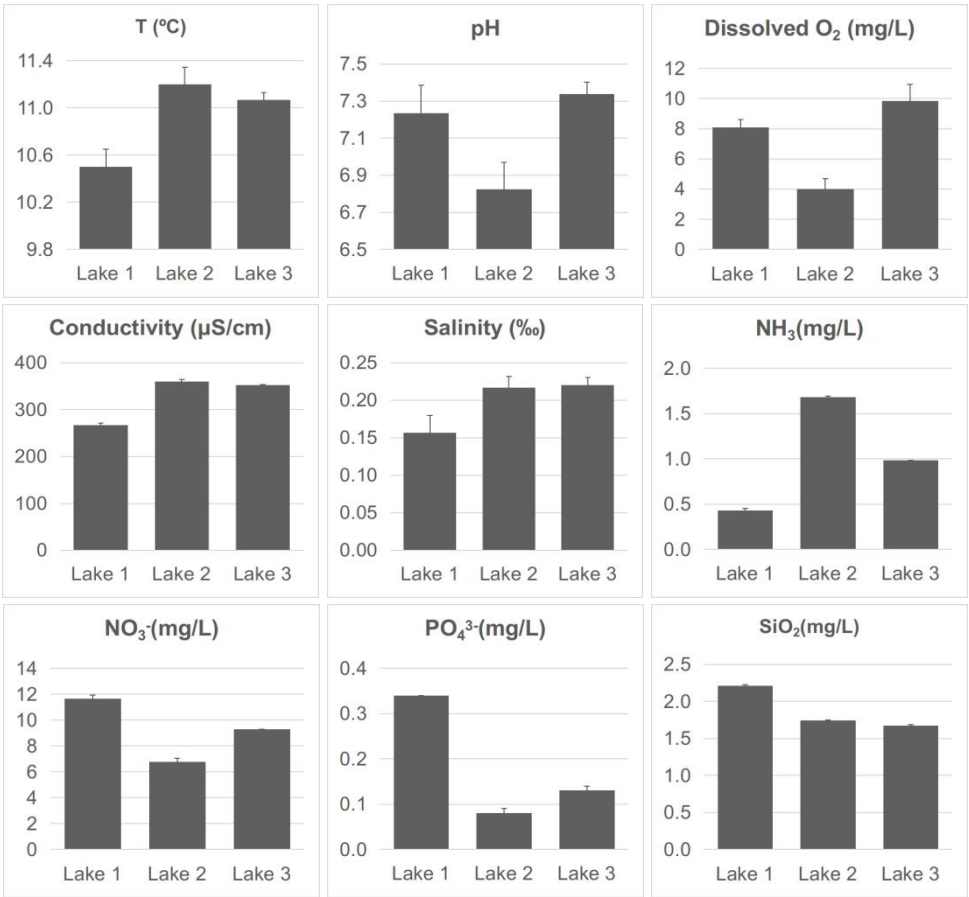


Figure 16 - Physical-chemical parameters obtained for the three studied lakes. Values represent the means and standard deviations of water temperature (T,), pH, conductivity, dissolved oxygen, salinity and concentrations of ammonia (NH_3), nitrates (NO_3^-), phosphates (PO_4^{3-}) and silica (SiO_2).

Table 7 - Water quality of the studied lakes, as indicated by the Water Quality Index (WQI; Brown *et al.*, 1970) and the "*Indice de Poluossensibilité Specifique"* (IPS; CEMAGREF, 1982).

	Water Quality Index		*Indice of Poluossensibilité Specifique*	
Lake	Value	Water quality category (classification interval)	IPS	Water quality category (classification category)
1	70	Good (70-89)	9.5	Intermediate (9-13)
2	57	Medium (50-69)	8.6	Bad (5-9)
3	77	Good (70-89)	12.5	Intermediate (9-13)

3.2. Diatom taxonomic identification

Diatom valve counts and species abundance percentages for each lake are presented on Table II of Appendix II. In Lake 1, ten species had an abundance over 1% and the most abundant species was *Gomphonema parvulum* (Figure 17C). In Lake 2, fourteen species showed abundance >1% and the most abundant species was *Tabularia Tabulata* (Figure 17A). In lake three, 21 species had an abundance >1% and the most abundant species was *Amphora pediculus* (Figure 17F). The species with an >1% abundance common to the three were *Gomphonema parvulum* (Figure 17C), *Melosira varians* (Figure 17D), *Navicula gregaria* (Figure 17E) and *Nitzschia palea* (Figure 17B). The IPS revealed that Lake 2 had a bad water quality and lake 1 and 3 had intermediate water quality (Table 7).

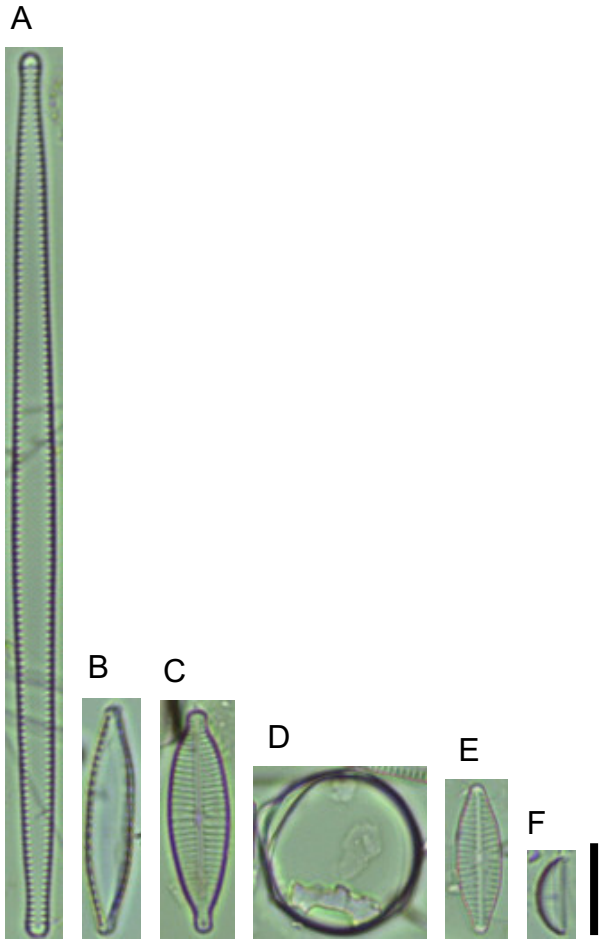


Figure 17 - Most abundant and common taxa in the three lakes of Oporto natural City Park: A - *Tabularia tabulata,* the most abundant in Lake 1; B – *Nitzschia palea,* abundant in the three lakes; C - *Gomphonema parvulum*, the most abundant in Lake 2 and abundant in the three lakes.; D – *Melosira varians*, abundant in the three lakes.; E – *Navicula gregaria*, abundant in the three lakes; F - *Amphora pediculus*, the most abundant in Lake 3; Scale bar = 10 µm.

3.3. Raman Spectroscopy

Figure 18 shows an example of Raman spectra obtained for diatom species common to the three lakes. Overall, a total of 14 Raman bands were identified in the dataset: around 867 cm^{-1}, 920 cm^{-1}, 963 cm^{-1}, 1013 cm^{-1}, 1160 cm^{-1}, 1180 cm^{-1}, 1198 cm^{-1}, 1270 cm^{-1}, 1315 cm^{-1}, 1390 cm^{-1}, 1445 cm^{-1}, 1526 cm^{-1}, 1606 cm^{-1} and 1656 cm^{-1}.

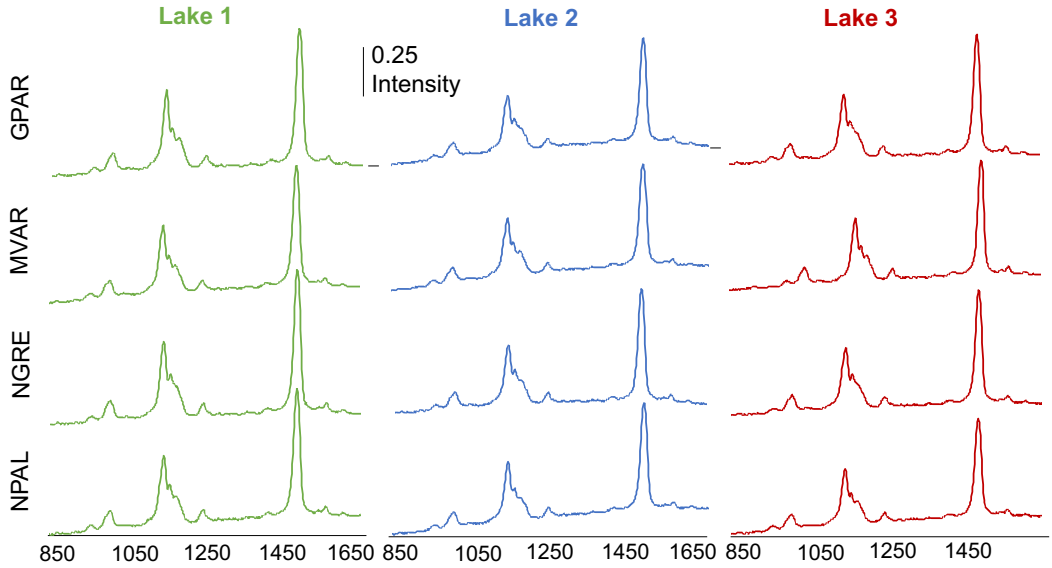


Figure 18 - Example of Raman Spectra obtained for *Gomphonema parvulum* (GPAR), *Melosira varians* (MVAR), *Navicula gregaria* (NGRE) and *Nitzschia palea* (NPAL) sampled in the three lakes studied.

The PLS results showed two significant components explaining a total of 24,8% of the variance (R^2Y=0.248). In Figure 19, Raman variables are ordered according to their importance to the components found. The most important variables in the model are highlighted in red. The significant PLS components found are presented in Figure 20. The separation of lakes and the four species common to all lakes provided by these components is also illustrated in the lower corner of the figure. The results show a very good separation of the three lakes. An horizontal gradient opposing Lake 2 to the remaining lakes is clearly observed. A vertical gradient further opposes Lake 1 to Lake 3. The Cluster Analysis based on the *x loadings* of the PLS identified 6 significant clusters of Raman variables contributing to the discrimination observed in the PLS (Figure 20). Table 8 summarizes the bands related to the nine important Raman variables identified and their respective molecular assignment, as retrieved from the available literature. To the best of our knowledge no molecular assignment has yet been done to band 1390 cm^{-1}.

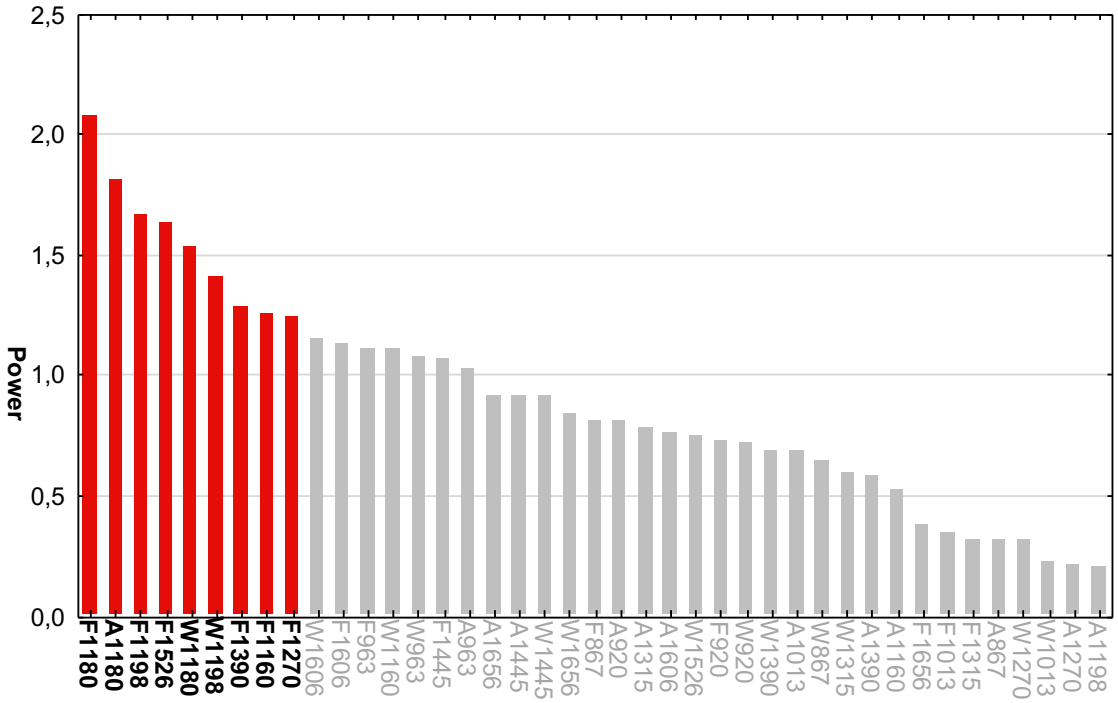


Figure 19 - Importance of Raman variables to the significant components identified in the Partial Least Squares (PLS) analysis. The most important variables are highlighted in red: Frequency (F), area (A) and Width (W) of the band 1180 cm^{-1}, Frequency and Width of the band 1198 cm^{-1} and the frequency of the bands 1526, 1390, 1160 and 1270 cm^{-1}.

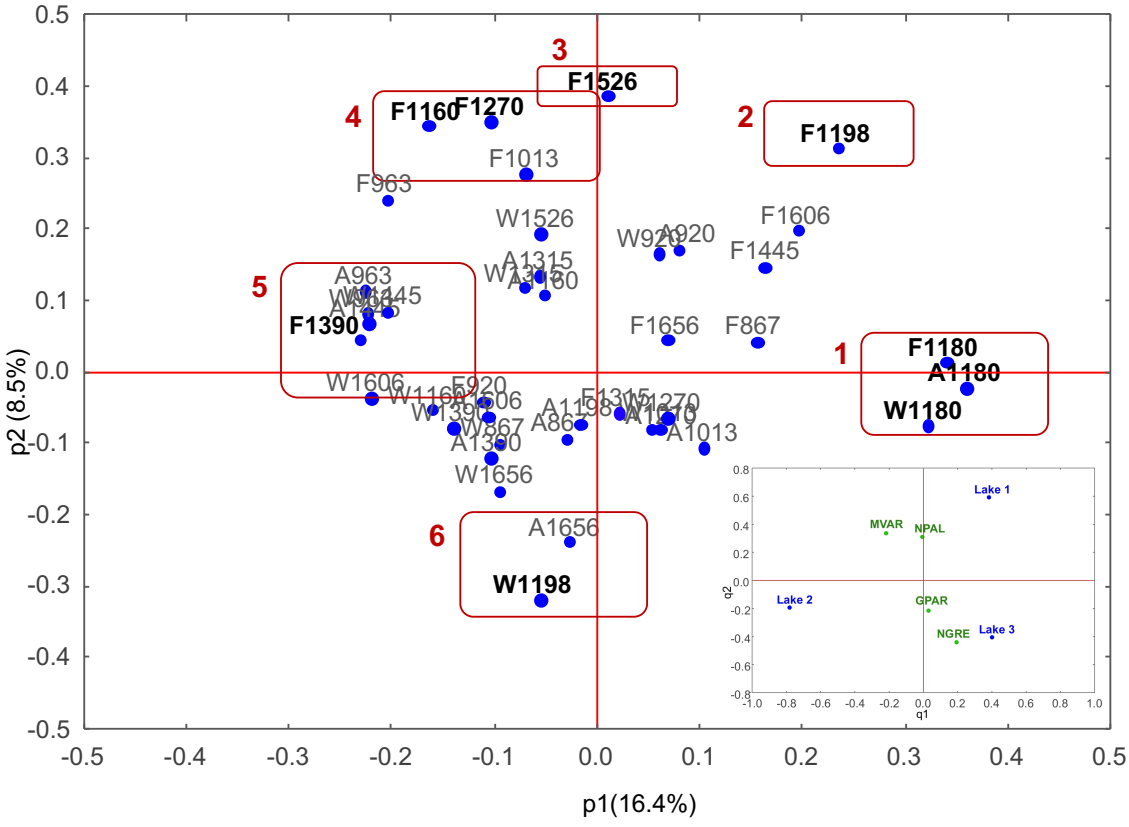


Figure 20 - Results of the Partial Least Squares (PLS) analysis performed on the Raman variables obtained. Two significant components were extracted, explaining 16.4% and 8.5% of the total variance. The most important variables

are highlighted in bold. The significant groups identified by the Cluster Analysis carried out on the PLS *x loadings* are numbered from 1 to 6. The inset shows the PLS separation of lakes and the species *Gomphonema parvulum* (GPAR), *Melosira varians* (MVAR), *Navicula gregaria* (NGRE) and *Nitzschia palea* (NPAL), common to all lakes.

Table 8 - Molecular assignments found in the available literature for the bands related to the significant variables identified by the Partial Least Squares analysis.

Band (cm^{-1})	**Assignments**	**Reference**
1160	C=S (Frustule)	De Tommasi, 2016; De Tommasi *et al.*, 2018
	C-C stretching modes from Carotenoids	Premvardhan *et al.*, 2009; Ruger *et al.*, 2016
1180	C=S (Frustule)	De Tommasi, 2016; De Tommasi *et al.*, 2018
	C-H deformational modes from oids	Premvardhan *et al.*, 2009
1198	C=S (frustule)	De Tommasi, 2016; De Tommasi *et al.*, 2018
	N-C stretching modes from Chl a	Ruger *et al.*, 2016
1270	Amide III; =CH bend (lipids); T,A	Notingher, 2007
1390	--	--
1526	C=C stretching modes from Carotenoids	Premvardhan *et al.*, 2009; Alexandre *et al.*, 2014; Ruger *et al*., 2016

The two-way ANOVAs carried out for each of the nine Raman variables significantly separating Lakes and Common species are shown in Table 9. A two-way ANOVA was also done for the area of band 1526 cm^{-1}, due to its great discriminant capability *per se* (Table 9). Significant differences were found for the main factors in most of the variables. For Width1180, Area1180 and Freq1390, significant differences among Common species and/or among Lakes were identified (Table 9). Variables A1180 and F1390 were significantly affected by the lake, while variable W1180 was affected by the two factors separately. For the remaining variables significance of the interaction term (L*CS) was found, indicating that the effect of the lake of origin was dependent on the species analyzed. The homogeneous subsets found through the Tukey HSD test are presented in Figure 21. The three Raman variables (area, width and frequency) of band 1180 cm^{-1} (Cluster 1) generally exhibited the lowest values in Lake 2 (all 3 variables), and highest values in *Navicula gregaria* species (F1180, W1180). For variable F1180, as a result of

the interaction between the two main factors, some conditions deviate from this general trend, such as the values of *Gomphonema parvulum* (GPAR) from Lake 3, which were similar to those of *Navicula gregaria* from the same lake. Overall these three variables clearly distinguished all three lakes from each other. The frequency of band 1390 cm^{-1} (F1390, Cluster 5) contributed to this distinction. Though, by showing a completely opposite trend, where Lake 2 exhibited the highest values and Lake 3 the lowest. For variable F1198 (Cluster 2) the lowest values were again found in lake 2, compared to the other two lakes. In the remaining lakes, this general trend is significantly influenced by the species factor, in particular in Lake 1, where the highest values were obtained for *Navicula gregaria* collected from Lake 1. Variable W1198 (Cluster 6) tended to show the lowest values in Lake 1, than in the remaining lakes (particularly in *Gomphonema parvulum*) and the highest values in *Gomphonema parvulum* and *Navicula gregaria* collected from Lake 3. For the remaining variables, the strong interaction between the two main factors (Lakes and Common species) reflected in even a higher variability in the values displayed by each species x lake combination. Here, general trends were not evident. Some conditions, however, stand out for exhibiting significantly different opposite values. Namely, F1526 (Cluster 3) separates *Nitzschia palea* of Lake 1 (higher values) and *Melosira varians* of Lake 3 from *Navicula gregaria* of Lake 3 (lower values). Conversely, F1270 and F1160 (Cluster 4) mainly separated *Nitzschia palea* of Lake 2 (higher than the remaining) and *Navicula gregaria* of Lake 3 (lower than the remaining) from each other and from the remaining combinations of lakes and species. Finally, A1526 (outside the clusters) mainly separates *Melosira varians* of Lake 2 (higher values) from *Navicula gregaria* and *Nitzschia palea* of Lake 1 (lower values).

Table 9 - Results of the two-way ANOVAs performed for the significant Raman variables identified through the PLS. The sources of variation were the Lakes (L) with two degrees of freedom (*df*), the Common Species (CS, *df=3*) and the interaction between these two factors (L*CS, *df=6*). Total df=11 and *df* of the error term were 204.

Cluster number	Raman variable	Multiple R^2	P-value			
			Total	L	CS	L*CS
1	Width1180	0.39	<0.0001	<0.0001	0.0023	0.0718
	Area1180	0.52	<0.0001	<0.0001	0.0107	0.1296
	Freq1180	0.69	<0.0001	<0.0001	<0.0001	0.0009
2	Freq1198	0.50	<0.0001	<0.0001	0.2797	<0.0001
3	Freq1526	0.41	<0.0001	<0.0001	<0.0001	0.0009
4	Freq1270	0.42	<0.0001	<0.0001	<0.0001	<0.0001
	Freq1160	0.32	<0.0001	<0.0001	<0.0001	<0.0001

5	Freq1390	0.29	<0.0001	<0.0001	0.0133	0.0660
6	Width1198	0.33	<0.0001	<0.0001	0.1176	0.0001
	Area1526	*0.30*	*<0.0001*	*<0.0001*	*<0.0001*	*0.0001*

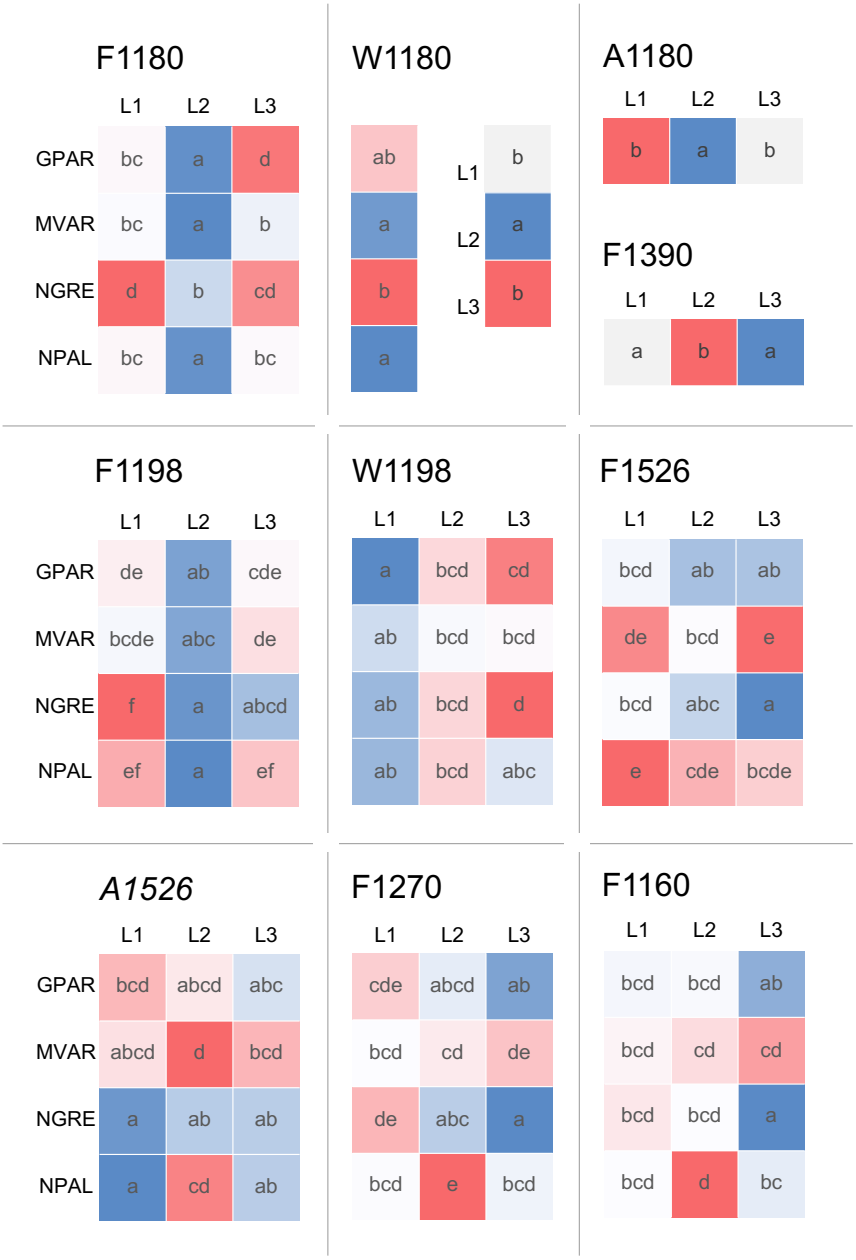


Figure 21 - Homogeneous subsets (Tukey HSD test, p<0.01) identified for the relevant PLS Raman variables. The capital letter in the name of the variables represent the area (A), width (W) and frequency (F) of the bands. The species common to Lake 1 (L1), Lake 2 (L2) and Lake 3 (L3) were *Gomphonema parvulum* (GPAR), *Melosira varians* (MVAR), *Navicula gregaria* (NGRE) and Nitzschia palea (NPAL). For each variable, shades of blue represent values lower than average and red shades represent values above average.

3.4. Lake diagnosis

To identify the best models classifying the three lakes with the least information possible, artificial neural network (ANN) analysis was used. For this, all combinations of input data (continuous and categorical) were first investigated. In particular, the continuous input to the ANN was all the Raman variables or combinations of width or frequency with the non-normalized area of band 1526 cm^{-1}. Overall, eight Raman classification models were selected for interpretation; their data characteristics are shown in Table 10. These were the models showing the highest accuracy rates in the test series for each combination of categorical and continuous input data and case (species) selection. These models with Raman variables were found to classify the lakes with excellent accuracy (87.4% to 95.5%), as determined in the validation series (models A to H in Table 10). Also included in Table 10 is model I, which only takes taxa as categorical input, with no Raman variables were included. Comparison of model I with model A, clearly shows the contribution of Raman data for the classification performance. For model I, the accuracy obtained was insufficient (61.4%) for lake classification. Model A, in which the opposite was done, i.e. only the Raman variables were entered in, discriminated very well the three lakes, with much higher validation accuracy (88.6%, Table 5), i.e. 27.2% over model I and using no taxonomic information. The classification done by model D further stresses the relevance of Raman data for accurate lake discrimination. Model D was the best classification model without taxa as categorical input, exhibiting 93.2% accuracy in the validation series. Furthermore, this model was based on data (the frequency of all bands and the non-normalized area of band 1526 cm^{-1}) obtained for the four species that were invariably common to the three lakes (*Gomphonema parvulum*, *Melosira varians*, *Navicula gregaria* and *Nitzschia palea*). Models G and H were very similar and showed the highest classification performance (95.5%). However, they required the taxa data as categorical input. Their difference relies on using Raman data acquired for all species (model G) or just the four common species (model H). Model H is thus preferable to model G, given that it uses comparatively less amount of data. To proceed to a more refined analysis, measures of diagnostic performance (i.e. sensitivity, accuracy) were calculated for all models (Table 11).

Table 10 - Information entered in the best Artificial Neuronal Network (ANN) classification models obtained. The network architecture used was Multilayer Perceptron (MLP). Continous input data were all the Raman variables, or different combinations of these variables with the non-normalized area of band 1526 cm^{-1} (NN A1526).

Data entered	Model								
	A	B	C	D	E	F	G	H	I
Categorical input	None	None	None	None	Taxa	*Taxa*	*Taxa*	*Taxa*	*Taxa*
Continuous input	Raman	Raman	Raman	Raman	Raman	Raman	Raman	Raman variables	None
Species selected	All	Common species	All	Common species	All	Common species	All	Common species	All
Raman variables selected	All	All	All; NN A1526	F; NN A1526	W; F	W; F	F; NN A1526	F; NN A1526	None

Table 11 - Measures of diagnostic performance for the best artificial neural network models obtained for the validation series. Model accuracy and sensitivity by lake are presented.

Model	Accuracy (%)	Sensitivity (%)		
		Lake 1	**Lake 2**	**Lake 3**
A	88.6	75.0	92.7	92.5
B	88.6	94.1	100	73.3
C	87.4	70.0	92.7	92.5
D	93.2	88.2	100	93.3
E	94.9	87.5	98.2	96.3
F	90.9	94.1	100	80.0
G	95.5	97.5	92.7	96.3
H	95.5	94.1	100	93.3
I	61.4	52.3	50.0	73.8

Globally, all models using Raman data showed excellent results in the validation series. In particular, models D and H were especially sensitive (100%) in predicting Lake 2, correctly classifying all cases for this lake (Table 11). Overall, model A and model D were the most parsimonious, while still showing very high performance.

4. Discussion

Evidence produced over the last decade has suggested the importance of Raman vibrational spectroscopy to diagnose aquatic ecosystems under different levels of physico-chemical stress (Pinto *et al*., submitted and references herein). In the present work Raman spectroscopy was applied to freshwater diatoms from three lakes with similar hydromorphological conditions. The empirical Raman data acquired was analysed by ANN to produce a model diagnosing the three similar lakes under investigation. A characterization of the lakes using physico-chemical parameters and a diatom index revealed that in terms of water quality, Lake 2 stood out from the other lakes studied. The poorer chemical water quality of this lake, compared to the other two, was reflected by its lower WQI. The higher concentration of ammonia, lower pH and lower oxygen level found in this lake, compared to the other two lakes, might indicate an organic contamination from an external source, possible due to the high number of birds in the lake that were observed locally. For diatom analysis, globally a total of 45 different species were identified in the three lakes. Species richness was lower in Lake 1 (20 different species identified) than in the other two, and fairly similar between Lake 2 (27 species) and Lake 3 (32 species). According to the species identified and their abundance, Lake 2 also exhibited a poorer biological water quality as indicated by the IPS, which is in agreement with the chemical data. The common diatom species abundant in the three lakes were *Gomphonema parvulum, Melosira varians*, *Navicula gregaria* and *Nitzschia palea*. *Gomphonema parvulum* and *M. varians* are high-profile species, meaning that in a biofilm these individuals are found in the upper layer with easier access to light and nutrients but also more susceptible to pollutants and other chemicals (Rimet & Bouchez, 2012). *Navicula gregaria* and *N. palea* are motile species (Rimet & Bouchez, 2012). Except for *M. varians*, which is abundant in high-nutrient waters (Passy & Larson, 2011), these species are characteristic of practically all kinds of freshwater environments (Abarca *et al.,* 2014; Cox, 1987; Trobajo *et al.*, 2009). *Melosira varians*, *Navicula gregaria* and *Nitzschia palea* are also known by their tolerance to physical and chemical disturbing (Cox, 1987; Passy & Larson, 2011; Trobajo *et al*., 2009). In contrast, *Gomphonema parvulum* seems to be sensitive to water quality changes (Abarca *et al*., 2014).

From the Raman data acquired in this study, 14 RS bands were detected using a 532 nm wavelength with 0.1 mW potency and an acquisition time of 10s with three accumulaitons: 867 cm^{-1}, 920 cm^{-1}, 963 cm^{-1}, 1013 cm^{-1}, 1160 cm^{-1}, 1180 cm^{-1}, 1198 cm^{-}

^{1}, 1270 cm^{-1}, 1315 cm^{-1}, 1390 cm^{-1}, 1445 cm^{-1}, 1526 cm^{-1}, 1606 cm^{-1} and 1656 cm^{-1}. Similar Raman spectra were previously obtained for *Thalassiosira pseudonana* (Meksiarun *et al*., 2015), *Ditylum brightwellii* (Rüger *et al*., 2016) and *Cylindrotheca closterium* (Pinzaru *et al*., 2016). For *T. pseudonana* the conditions for data acquisition were an incident laser with 785 nm wavelength, a 30 nmW potency and 30s of acquisition time (Meksiarun *et al*., 2015). For *D. brightwellii*, 785 nm wavelength was used with 30 mW potency and 2s acquisition time (Rüger *et al*., 2016). For *C. closterium*, the authors employed acquisition conditions similar to the ones in the present study; 532 nm with a 0.1-1 mW potency and 1s acquisition time (Pinzaru *et al*., 2016). The results of the PLS analysis indicated 6 bands in the Raman spectra as significantly relevant for lake characterization. These were found to be within the fingerprint region 1160 to 1526 cm^{-1} (Table 8). Most of these bands have been previously assigned to vibrations of frustule components, carotenoids and chlorophyll *a* (Premvardhan *et al*., 2009; Rüger *et al*., 2016). Specifically, bands 1160, 1180 and 1526 cm^{-1} were assigned to C-C , C-H and C=C stretching modes from carotenoids, respectively. Bands 1198 and 1270 cm^{-1} were assigned to N-C and C-N stretching modes of chlorophyll *a*, respectively. For band 1390 cm^{-1} no assignments could be found in the available literature. Variations in the area of the pigment bands might be related with the amount of these compounds in the cells (Pinzaru *et al*., 2016). The results of this study therefore suggest the presence of distinct molecular conformations of carotenoids and chlorophyll *a* in the species collected from the different lakes, as indicated by the significant differences found in the ANOVA analysis. Among them, for example, *Navicula gregaria* from Lake 1 and *Gomphonema parvulum* from Lake 3, would have a higher number of molecular confirmations than *Gomphonema parvulum*, *Melosira varians* and *Nitzschia palea* from Lake 2, as indicated by F1180. Species from Lake 2 should also present slightly different molecular conformations of Chl *a*, as indicated by F1198. Despite the fact that diatom Raman spectra is usually dominated by pigment contribution (Pinzaru *et al*., 2016; Rüger *et al*., 2016), differences in $=CH_2$ lipids composition appear also to occur among the studied lakes and species, as indicated by the frequencies acquired to band 1270 cm^{-1}. According to other studies in diatoms, changes in the wavelength of the incident laser cause frequency shifts of pigment bands due to resonance phenomena caused by the similarity with the energy of the different pigments (Premvardhan *et al*., 2009; Premvardhan *et al*., 2010). Frequency differences of the pigment bands in solution can also derive from conformational changes due to the polarity of the solvent (Premvardhan *et al*., 2009; Premvardhan *et al*., 2010). However, as in this study the same excitation

wavelength was always used, and the pigments were not extracted. Hence, differences in Raman band frequency should be due to the presence of different pigments or pigments with different conformations. In contrast, differences in band width reflect the variety of those molecules and molecular conformations within an individual. Previous studies have also shown that environmental stress (e.g. nutrient starvation, high light exposure, high temperature, lower dissolved oxygen concentration) can alter the levels of carotenoids in microalgae, and their photosynthetic activity, to cope with the living/exposure conditions (Faraloni & Torzillo, 2017 and references therein). For example, one study employed Raman spectroscopy together with chemometric methods to assess the *in vivo* nutrient status of single microalgal cells (Heraud, Beardall, *et al*., 2007). Photosynthetic activity may also be reduced, by decreasing the transport of photosynthetic electrons, but this reduction may also be associated with the formation of free radicals and oxidant species. Conversely, carotenoids can be increased to protect against oxidative damage caused by those stressful conditions. Namely, astaxanthin (a secondary carotenoid), the pigment lutein, as well as antheraxanhitn, violaxanthin and zeaxanthin (pigments of the xanthophyll cycle) (Demmig-Adams *et al*., 1989; Eskling *et al*. 1997; Faraloni & Torzillo, 2017 and references therein). Moreover, Raman studies have shown that in *C. meneghiniana* carotenoid bands can vary with light conditions, due to the production of different carotenoids and their conformation (Alexandre *et al*., 2014). Another Raman study found that carotenoid bands seemed to decrease in diatoms that were in the exponential growth phase (Rüger *et al*., 2016). As to width, the differences and significance found for bands 1180 cm^{-1} (carotenoids with C-H deformational modes) and 1198 cm^{-1} (Chl a with N-C stretching modes) indicates the occurrence of pigment diversity occurring among the lakes x species investigated. Another interesting result is related to band 1526 cm^{-1}. The area of this band was correlated with all other Raman variables and showed significant lake by species differences. The results suggest in particular that *Melosira varians* and *Nitzschia palea* from Lake 2 had significantly higher concentration of carotenoids with C=C stretching modes than the other microalgae, while *Navicula gregaria* and *Nitzschia palea* from Lake 1 had significantly lower concentration of these carotenoids. Variability in Raman spectra intensity in algal species, in relation to pigment concentrations in different intracellular locations was previously shown (Heraud, Beardall, *et al*., 2007). Further controlled laboratory studies conjugated employed chemical and biological methods (e.g. qPCR) are necessary to complement the information acquired and better understand the

responses of these diatoms, and their frustule variation, to the environmental conditions of the three lakes.

From the diagnostic tests carried out with the Artificial Neuronal Network, it was found that using only Raman data provided a notable improvement in the classification performance of the models, relative to using only the *taxa* data. Despite the fact that the three lakes are closely located, with and bear an interconnected water system (Morais, 2009), their different chemical conditions were remarkably predicted by Raman models, and much better than the one based on *taxa* identification. The models empolying data on Raman band frequency and the non-normalized area of the band 1526 cm^{-1} were those providing the highest classification accuracy (95.5%). However, besides the Raman data, these models also required the *taxa* data to attain such a performance. The most expedite model obtained was in fact Model A, which showed an excellent classification accuracy of 88.6% using all Raman data as input and requiring no species taxonomic identification. In fact, according to the classification criteria proposed by the European Centre for the Validation of Alternative Methods (ECVAM), the overall performance of a diagnostic test (given by its classification accuracy) is considered sufficient for accuracy values in the range 0.65-0.74, good for accuracy values in the range 0.75-0.84 and excellent for accuracy values >0.85 (Winter *et al*., 2008). On this, it is of note that all the Raman models obtained in this work fall in the excellent accuracy category. In other studies, ANN was used to predict the water quality in Odivelas reservoir for a period of 10 years using abiotic parameters such as chemical oxygen demand (COD), dissolved oxygen (DO) and total suspended solids (TSS) (Couto *et al*., 2012). The results obtained were comparable to those of the present work, showing that the model was 100% accurate in predicting the most polluted areas and 95.5% accurate in predicting the least polluted areas (Couto *et al*., 2012). The models presented in this study were also more accurate in predicting the most polluted lake (Lake 2). The use of ANN to predict chemical conditions using bioindicators was also done in ecotoxicological studies with equally good results (Teles *et al*., 2015; Amorim *et al*., 2018). For example, Amorim *et al*. (2018) used ANN to predict water contamination by *Escherichia coli* as measured by the swimming behavioural response of zebrafish (*Danio rerio*). Once more, the sensitivity in detecting system disturbances, i.e. higher concentrations of *E. coli* was 100%. The low pollution level found in Lake 2, and the 100% sensitivity obtained for this lake, further reinforce the usefulness of ANN for application in diagnotic tests. A very interesting finding of this study is the low number of incorrectly classifed cases obtained with the Raman classification models. The overall sensitivity rates (≥89%) obtained for

the two best Raman models (A and H) are included in the interval returned by RT-PCR tests set for the detection of COVID-19, which varies between 71% and 98% sensitivity, according to a recent systematic review conducted by Arevalo-Rodriguez and colleagues (2020). As indicated above, 71% falls in the sufficient accuracy category (65-74%) and 98% in the excellent accuracy category (≥85%) (Winter *et al*., 2008). The sensitivity of model A, the most parsimonious model, was 75% for Lake 1 and 93% for Lake 2 and Lake 3, while the sensitivity of model H, the highest accuracy model, was 100% for Lake 2, 94% for Lake 1 and 93% for Lake 3. For serological COVID-19 tests, a systematic review showed the detection of Immunoglobulin M (IgM) or Immunoglobulin G (IgG) with Enzymatic-Linked Immunosorbent Assays (ELISA) had 84.3% sensitivity (Bastos et al., 2020), lower than the best Raman models. Lateral Flow Immunoassays (LFIAs) showed even lower sensitivity (66,0%) in the detection of IgM or IgG (Bastos *et al*., 2020). In contrast, Chemiluminescent Immunoassays (CLIAs) showed excellent sensitivity (97,8%) in IgM or IgG detection (Bastos *et al*., 2020), similar to models G and H presented herein. This ANN sensitivity, and ability to discriminate near normal conditions, represents a clear advantage for its application as follow up of ecosystem restoration measures. Ability to detect minimal health deviations can allow determining system thresholds and protecting the least tolerant species, usually the first ones to be lost due to environmental degradation.

5. Conclusion

Application of Raman spectroscopy to diatoms, combined with artificial neural network analysis, provided an excellent diagnose of three interconnected lakes from a City Park, bearing high similarity among them. The diagnose accuracy was 89% for the model using only Raman data and 96% for the model using both Raman and *taxa* data. The former represents a very important foundation for future environmental studies, to be replicated at larger scales and geographic settings. To the best of our knowledge this is the first study providing empirical proof on the advantages and accuracy of diatom Raman spectroscopy for environmental diagnosis of freshwater ecosystems. As presented herein, this can used as a simple and fast method, not requiring taxonomic identification of sampled diatoms. The results further indicate that diatoms cell components and metabolites contributing to Raman spectra, and related lake classification, vary in quantity and/or at the molecular level, depending on the environmental conditions and the species analyzed.

Acknowledgements

The authors would like to thank the EU and FCT/UEFISCDI/FORMAS for funding, in the frame of the collaborative international consortium REWATER, financed under the ERA-NET Cofund WaterWorks2015 (Water JPI). This research was also supported by national funds through FCT (Portuguese Foundation for the Science and Technology) within the scope of UIDB/04423/2020 and UIDP/04423/2020. The authors would also like to thank to NECL (Network of Extreme Conditions Laboratories) for providing the Raman equipment necessary.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abarca, N., Jahn, R., Zimmermann, J., & Enke, N. (2014). Does the cosmopolitan diatom Gomphonema parvulum (Kützing) Kützing have a biogeography? PLoS One, 9(1), e86885.

Abbas, A., Josefson, M., & Abrahamsson, K. (2011). Characterization and mapping of carotenoids in the algae Dunaliella and Phaeodactylum using Raman and target orthogonal partial least squares. Chemometrics and Intelligent Laboratory Systems, 107(1), 174-177.

Akkas, S. B., & Severcan, F. (2012). Diagnosis and Screening of Aquatic Environments by Vibrational Spectroscopy. Vibrational Spectroscopy in Diagnosis and Screening, 6, 321.

Alexandre, M. T., Gundermann, K., Pascal, A. A., van Grondelle, R., Buchel, C., & Robert, B. (2014). Probing the carotenoid content of intact Cyclotella cells by resonance Raman spectroscopy. Photosynthesis Research, 119(3), 273-281.

Almeida, S. F. P., & Gil, M. C. P. (2001). dÉcologie des diatomées d'eau douce de la région centrale du Portugal. Cryptogamie Algologie, 22(1), 109-126.

Amorim, J., Fernandes, M., Abreu, I., Tavares, F., & Oliva-Teles, L. (2018). Escherichia coli's water load affects zebrafish (Danio rerio) behavior. Science of The Total Environment, 636, 767-774.

Arevalo-Rodriguez, I., Buitrago-Garcia, D., Simancas-Racines, D., Zambrano-Achig, P., del Campo, R., Ciapponi, A., . . . Low, N. (2020). False-negative results of initial RT-PCR assays for COVID-19: a systematic review. medRxiv.

Bastos, M. L., Tavaziva, G., Abidi, S. K., Campbell, J. R., Haraoui, L.-P., Johnston, J. C., . . . Trajman, A. (2020). Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. bmj, 370.

Bishop, C. M. (1995). Neural networks for pattern recognition: Oxford university press.

Blanco, S., Ector, L., & Bécares, E. (2004). Epiphytic diatoms as water quality indicators in Spanish shallow lakes. Vie et Milieu, 54(2-3), 71-80.

Boyd, C. E. (2019). Water quality: an introduction: Springer Nature.

Brown, R. M., McClelland, N. I., Deininger, R. A., & Tozer, R. G. (1970). A WATER QUALITY INDEX- DO WE DARE.

CEMAGREF, M. (1982). Etude des méthodes biologiques d'appréciation quantitative de la qualité des eaux. Rapport Cemagref QE Lyon-AF Bassin Rhône Méditerranée Corse.

Chen, L., Weng, D., Du, C., Wang, J., & Cao, S. (2019). Contribution of frustules and mucilage trails to the mobility of diatom Navicula sp. Scientific reports, 9(1), 1-12.

Coste, M., Boutry, S., Tison-Rosebery, J., & Delmas, F. (2009). Improvements of the Biological Diatom Index (BDI): Description and efficiency of the new version (BDI-2006). Ecological indicators, 9(4), 621-650.

Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy, (2000).

Couto, C., Vicente, H., Machado, J., Abelha, A., & Neves, J. (2012). Water quality modeling using artificial intelligence-based tools. International Journal of Design & Nature and Ecodynamics, 7(3), 300-309.

Cox, E. J. (1987). Studies on the diatom genus Navicula Bory. VI. The identity, structure and ecology of some freshwater species. Diatom research, 2(2), 159-174.

De Tommasi, E. (2016). Light manipulation by single cells: the case of diatoms. Journal of Spectroscopy, 2016.

De Tommasi, E., Congestri, R., Dardano, P., De Luca, A. C., Managò, S., Rea, I., & De Stefano, M. (2018). UV-shielding and wavelength conversion by centric diatom nanopatterned frustules. Scientific reports, 8(1), 1-14.

Desrosiers, C., Leflaive, J., Eulin, A., & Ten-Hage, L. (2013). Bioindicators in marine waters: benthic diatoms as a tool to assess water quality from eutrophic to oligotrophic coastal ecosystems. Ecological indicators, 32, 25-34.

Dixit, S. S., Smol, J. P., Charles, D. F., Hughes, R. M., Paulsen, S. G., & Collins, G. B. (1999). Assessing water quality changes in the lakes of the northeastern United States using sediment diatoms. Canadian Journal of Fisheries and Aquatic Sciences, 56(1), 131-152.

Drobatz, K. J. (2009). Measures of accuracy and performance of diagnostic tests. Journal of Veterinary Cardiology, 11, S33-S40.

Faraloni, C., & Torzillo, G. (2017). Synthesis of antioxidant carotenoids in microalgae in response to physiological stress. Carotenoids. IntechOpen, 143-157.

Germain, H. (1981). Flore Des Diatomees: Diatomophycees: Eaux Douces Et Saumatres Du Massif Armoricain Et Des Contrees Voisines D'Europe Occidentale: Societe Nouvelle des Editions Boubee.

Heraud, P., Beardall, J., McNaughton, D., & Wood, B. R. (2007). In vivo prediction of the nutrient status of individual microalgal cells using Raman microspectroscopy. FEMS microbiology letters, 275(1), 24-30.

Heraud, P., Wood, B. R., Beardall, J., & McNaughton, D. (2007). Probing the Influence of the Environment on Microalgae Using Infrared and Raman Spectroscopy. In New Approaches in Biomedical Spectroscopy (Vol. 963, pp. 85-106): American Chemical Society.

INAG. (2008). Protocolo de amostragem e análise para o fitobentos-diatomáceas In Manual para a Avaliação Biológica da Qualidade da Água em Sistemas Fluviais Segundo a Diretiva do Quadro da Água.

Langlotz, C. P. (2003). Fundamental measures of diagnostic examination performance: usefulness for clinical decision making and research. Radiology, 228(1), 3-9.

Laviale, M., Beaussart, A., Allen, J., Quilès, F., & El-Kirat-Chatel, S. (2019). Probing the Adhesion of the Common Freshwater Diatom Nitzschia palea at Nanoscale. ACS applied materials & interfaces, 11(51), 48574-48582.

Lear, G., Dopheide, A., Ancion, P. Y., Roberts, K., Washington, V., Smith, J., & Lewis, G. (2012). Biofilms in freshwater: their importance for the maintenance and monitoring of freshwater health. Microbial Biofilms: Current Research and Applications, 129-151.

Lecointe, C., Coste, M., & Prygiel, J. (1993). “Omnidia”: software for taxonomy, calculation of diatom indices and inventories management. Hydrobiologia, 269(1), 509-513.

Matos, A. I. (2014). Development of molecular tools for the early warning of potentially toxic cyanobacteria. (Master), University of Porto, Porto, Portugal.

Meksiarun, P., Spegazzini, N., Matsui, H., Nakajima, K., Matsuda, Y., & Sato, H. (2015). In vivo study of lipid accumulation in the microalgae marine diatom Thalassiosira pseudonana using Raman spectroscopy. Applied Spectroscopy, 69(1), 45-51.

Mendes, T., Almeida, S. F., & Feio, M. J. (2012). Assessment of rivers using diatoms: effect of substrate and evaluation method. Fundamental and Applied Limnology/Archiv für Hydrobiologie, 179(4), 267-279.

Morais, J. (2009). Avaliação do Risco de Ocorrência de Cianobactérias Tóxicas nos Lagos do Parque da Cidade do Porto. (Master), University of Porto, Porto, Portugal.

Morin, S., Gómez, N., Tornés, E., Licursi, M., & Rosebery, J. (2016). Benthic diatom monitoring and assessment of freshwater environments: standard methods and future challenges. Aquatic Biofilms, 111.

Pandey, L. K., Bergey, E. A., Lyu, J., Park, J., Choi, S., Lee, H., . . . Han, T. (2017). The use of diatoms in ecotoxicology and bioassessment: insights, advances and challenges. Water Research, 118, 39-58.

Parker, F. S. (1983). Applications of infrared, Raman, and resonance Raman spectroscopy in biochemistry: Springer Science & Business Media.

Passy, S. I., & Larson, C. A. (2011). Succession in stream biofilms is an environmentally driven gradient of stress tolerance. Microbial ecology, 62(2), 414.

Patrick, R. (1973). Use of algae, especially diatoms, in the assessment of water quality. In Biological methods for the assessment of water quality: ASTM International.

Pinto, R., Mortágua, A., Almeida, S. F., Serra, S., & Feio, M. J. (2020). Diatom size plasticity at regional and global scales. Limnetica, 39(1), 387-403.

Pinzaru, S. C., Müller, C., Tomšić, S., Venter, M. M., Brezestean, I., Ljubimir, S., & Glamuzina, B. (2016). Live diatoms facing Ag nanoparticles: surface enhanced Raman scattering of bulk cylindrotheca closterium pennate diatoms and of the single cells. RSC advances, 6(49), 42899-42910.

Premvardhan, L., Bordes, L., Beer, A., Buchel, C., & Robert, B. (2009). Carotenoid structures and environments in trimeric and oligomeric fucoxanthin chlorophyll a/c2 proteins from resonance Raman spectroscopy. J Phys Chem B, 113(37), 12565-12574.

Premvardhan, L., Robert, B., Beer, A., & Büchel, C. (2010). Pigment organization in fucoxanthin chlorophyll a/c2 proteins (FCP) based on resonance Raman spectroscopy and sequence analysis. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1797(9), 1647-1656.

Pytlik, N., Klemmed, B., Machill, S., Eychmüller, A., & Brunner, E. (2019). In vivo uptake of gold nanoparticles by the diatom Stephanopyxis turris. Algal research, 39, 101447.

Rimet, F. (2012). Diatoms: an ecoregional indicator of nutrients, organic matter and micropolluants pollution. Université de Grenoble,

Rimet, F., & Bouchez, A. (2012). Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. Knowledge and management of Aquatic Ecosystems(406), 01.

Round, F. E., Crawford, R. M., & Mann, D. G. (2007). Diatoms: biology and morphology of the genera: Cambridge university press.

Rüger, J., Mondol, A. S., Schie, I. W., Popp, J., & Krafft, C. (2019). High-throughput screening Raman microspectroscopy for assessment of drug-induced changes in diatom cells. Analyst, 144(15), 4488-4492.

Rüger, J., Unger, N., Schie, I. W., Brunner, E., Popp, J., & Krafft, C. (2016). Assessment of growth phases of the diatom Ditylum brightwellii by FT-IR and Raman spectroscopy. Algal research, 19, 246-252.

Squires, L. E., Rushforth, S. R., & Brotherson, J. D. (1979). Algal response to a thermal effluent: study of a power station on the provo river, Utah, USA. Hydrobiologia, 63(1), 17-32.

Tan, X., Zhang, Q., Burford, M. A., Sheldon, F., & Bunn, S. E. (2017). Benthic diatom based indices for water quality assessment in two subtropical streams. Frontiers in microbiology, 8, 601.

Teles, L. O., Fernandes, M., Amorim, J., & Vasconcelos, V. (2015). Video-tracking of zebrafish (Danio rerio) as a biological early warning system using two distinct artificial neural networks: Probabilistic neural network (PNN) and self-organizing map (SOM). Aquatic Toxicology, 165, 241-248.

Trobajo, R., Clavero, E., Chepurnov, V. A., Sabbe, K., Mann, D. G., Ishihara, S., & Cox, E. J. (2009). Morphological, genetic and mating diversity within the widespread bioindicator Nitzschia palea (Bacillariophyceae). Phycologia, 48(6), 443-459.

UNESCO-WHO-UNEP. (1996). Water Quality Assessments. Cambridge, UK: Chapman & Hall.

Vilbaste, S., & Truu, J. (2003). Distribution of benthic diatoms in relation to environmental variables in lowland streams. Hydrobiologia, 493(1-3), 81-93.

Winter, M. J., Redfern, W. S., Hayfield, A. J., Owen, S. F., Valentin, J.-P., & Hutchinson, T. H. (2008). Validation of a larval zebrafish locomotor assay for assessing the

seizure liability of early-stage development drugs. Journal of pharmacological and toxicological methods, 57(3), 176-187.

Wood, B. R., Heraud, P., Stojkovic, S., Morrison, D., Beardall, J., & McNaughton, D. (2005). A portable Raman acoustic levitation spectroscopic system for the identification and environmental monitoring of algal cells. Analytical chemistry, 77(15), 4955-4961.

Wu, Q., Nelson, W., Treubig, J., Brown, P., Hargraves, P., Kirs, M., . . . Hanlon, E. (2000). UV resonance Raman detection and quantitation of domoic acid in phytoplankton. Analytical chemistry, 72(7), 1666-1671.

Yuan, P., He, H. P., Wu, D. Q., Wang, D. Q., & Chen, L. J. (2004). Characterization of diatomaceous silica by Raman spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 60(12), 2941-2945.

Chapter 5 – General Discussion

Prior to the studies presented in this dissertation, not much was known about the applicability of Raman Spectroscopy for *taxa* identification and environmental diagnosis. From the works reviewed (Chapter 2), it was reported that some bands varied with *taxa* (Abbas et al., 2011; Wood *et al*., 2005; Yuan *et al*., 2004) and environmental conditions such as nutrients (Rüger *et al.*, 2016), light (Alexandre *et al.*, 2014) and CO_2 (Meksiarun *et al.*, 2015). Additionally, some laboratory studies disentangled the toxicological effects of DTT reflected in Raman spectra (Rüger *et al.*, 2019) or the mechanisms for the uptake of gold nanoparticles using Raman Imaging (Pytlik *et al.*, 2019). Despite the fact that vibrational spectroscopy, mostly Fourier-transform infrared spectroscopy, applied to other algae (including diatoms) had been previously suggested as an hypothesis for monitoring aquatic systems (Akkas & Severcan, 2012), field environmental approaches for diagnostic testing were not available. Nevertheless, from the studies explored in the literature review presented in this dissertation, it was possible to collect a variety of spectral signatures related to pigments (Chls *a* and *c1*, Fx, Ddx and Dtx), frustule, lipids and other substances such as DA and EPS. It was also possible to understand how Raman parameters (*i.e.* laser wavelength, time acquisition and objective) can be set to target certain cellular components. This information reinforced the potential of this technique applied to diatoms to fulfil the main objective of this dissertation, which was to test its applicability for *taxa* identification and environmental diagnosis. Importantly, it also provided a strong knowledge foundation and encouraged the realization of further studies supporting the application of Raman Spectroscopy in diatoms as a practical method to decrease the inherent constraints in diatom taxonomic identification using morphometric features and to investigate the effects of toxicants and environmental changes at individual, population and community levels.

In the studies related to the diatom taxa and the three lakes of Oporto City Park (Chapters 3 and 4) 14 bands were found to be characteristic of diatom Raman spectra. Significant differences among species, species common to the three lakes and the lakes were found for RS bands. According to the literature, differences regarding the normalized area of a Raman band might be due to the higher or lower quantity of a specific compound (Pinzaru *et al.*, 2016). Differences in Raman band frequency, however, must be cautiously interpreted. In diatom studies, frequency shifts are related to changes in the wavelength of the incident laser (Premvardhan *et al.*, 2009; Premvardhan *et al.*, 2010). If the wavelength of the incident laser matches the energy of

the transition of a specific compound, the band corresponding to that compound is enhanced as a result of resonance phenomenon (McCreery, 2005). Frequency differences of the pigment bands in solution can also be derived from conformational changes due to the polarity of the solvent (Premvardhan *et al.*, 2009; Premvardhan *et al.*, 2010). In the present study the incident laser used was invariable. Also, pigments were not extracted. Hence, the differences in frequency found for several bands might be related to differences in molecular conformation of the molecules assigned to these bands. In addition, differences in band width may be related to the variety of conformations within a single compound. (Premvardhan *et al.*, 2009; Premvardhan *et al.*, 2010). Indeed, a smaller or larger width results from the variety of different frequencies vibrating around a central peak, which are produced by the different molecular conformations present in the sample. Further studies are, thus, necessary to correctly interpret the biochemical implications of these conformational differences.

In the *taxa* approach of Chapter 3, six bands assigned to pigments (Premvardhan *et al.*, 2009; Rüger *et al.*, 2016) contributed the most for the principal components in the PLS analysis and were responsible for the majority of significant differences found among diatom species: 1013, 1160, 1180, 1198, 1270 and 1526 cm^{-1}. Despite the greater importance of these bands, most Raman variables contributed to species characterization and identification with high sensitivity, as for example *Achananthidium exiguum* (67%), *Fragilaria crotonensis* (67%), *Amphora pediculus* (71%), *Achananthidium minutissimum* (80%) and *Melosira varians* (82%). Among these are the pioneer *taxa Achnanthidium minutissimum* and *Amphora pediculus*, which are typical of baring substrates are tolerant to a wide range of environmental stressors (Rimet & Bouchez, 2012). The Raman profile obtained for the species analyzed, together with previous evidence of RS band variation among *taxa* (Abbas *et al*., 2011; Wood *et al*., 2005; Yuan *et al*., 2004), contribute to support the idea that RS has relevant potential to identify diatom species.

In Chapter 4, concerning lake diagnosis, several Raman bands contributed to detected differences among lakes and species as indicated by the PLS analysis. Except for band 1013 cm^{-1}, such bands also explained the differences found among the common species inhabiting the three lakes. In this study, differences were found among species and lakes that reflected a strong interaction between these two factors. The empirical data gathered was successfully used to derive classification models with the ANN. These models showed excellent accuracy levels.

The ANN models based on RS applied to diatoms proved to be a lot more effective for environmental diagnosis than for *taxa* identification. For *taxa* identification (Chapter 3) the subclass was the taxonomic level identified with highest accuracy (89%). The model identifying the order was the second more accurate (63%) and had better data distribution than the one identifying the subclass. Similar results were obtained for diatom identification using diatom images with sensitivity varying between 13% and 68% sensitivity (Pedraza *et al*., 2018). For environmental diagnosis (Chapter 4) the results showed much higher accuracy (>90%). Despite the fact that all the lakes were closely located, have a similar hydromorphological structure and are connected through water circulation (Morais, 2009), the accuracy of the most parcimonious models produced for lake diagnosis was 88.6% and 95.5%. It was interesting to verify that, results derived from field investigations, in which organism are influenced by a variety of uncontrollable conditions, can be comparable or even more accurate than studies done under controlled laboratory experiments (Amorim *et al*. 2018). ANN models seem to be very effective to discriminate sites under similar aquatic conditions. This ability to classify systems with so close chemical and ecological status favours the use in environmental diagnosis. Combined with Artificial Neural Network (ANN) methods, Raman spectroscopy can be used to discriminate different environmental conditions. This can decrease the constraints associated with the traditional morphometric methods. This was verified for the diagnosis of Lake 2 with 100% sensitivity. With these results it also important to highlight that one of the models is purely based on Raman variables since is based in common species that appear invariably in the three lakes. This demonstrates that, in the future, similar approaches could be adopted in environmental studies that are independent of the challenging morphometric diatom identification.

References

Abbas, A., Josefson, M., & Abrahamsson, K. (2011). Characterization and mapping of carotenoids in the algae Dunaliella and Phaeodactylum using Raman and target orthogonal partial least squares. *Chemometrics and Intelligent Laboratory Systems, 107*(1), 174-177.

Akkas, S. B., & Severcan, F. (2012). Diagnosis and Screening of Aquatic Environments by Vibrational Spectroscopy. *Vibrational Spectroscopy in Diagnosis and Screening, 6*, 321.

Alexandre, M. T., Gundermann, K., Pascal, A. A., van Grondelle, R., Buchel, C., & Robert, B. (2014). Probing the carotenoid content of intact Cyclotella cells by resonance Raman spectroscopy. *Photosynthesis Research, 119*(3), 273-281

Amorim, J., Fernandes, M., Abreu, I., Tavares, F., & Oliva-Teles, L. (2018). Escherichia coli's water load affects zebrafish (Danio rerio) behavior. *Science of The Total Environment, 636*, 767-774.

Couto, C., Vicente, H., Machado, J., Abelha, A., & Neves, J. (2012). Water quality modeling using artificial intelligence-based tools. *International Journal of Design & Nature and Ecodynamics, 7*(3), 300-309.

McCreery, R. L. (2005). *Raman spectroscopy for chemical analysis* (Vol. 225): John Wiley & Sons.

Meksiarun, P., Spegazzini, N., Matsui, H., Nakajima, K., Matsuda, Y., & Sato, H. (2015). In vivo study of lipid accumulation in the microalgae marine diatom Thalassiosira pseudonana using Raman spectroscopy. *Applied Spectroscopy, 69*(1), 45-51

Morais, J. (2009). *Avaliação do Risco de Ocorrência de Cianobactérias Tóxicas nos Lagos do Parque da Cidade do Porto.* (Master), University of Porto, Porto, Portugal.

Pedraza, A., Bueno, G., Deniz, O., Ruiz-Santaquiteria, J., Sanchez, C., Blanco, S., . . . Cristobal, G. (2018). *Lights and pitfalls of convolutional neural networks for diatom identification.* Paper presented at the Optics, Photonics, and Digital Technologies for Imaging Applications V.

Pinzaru, S. C., Müller, C., Tomšić, S., Venter, M. M., Brezestean, I., Ljubimir, S., & Glamuzina, B. (2016). Live diatoms facing Ag nanoparticles: surface enhanced Raman scattering of bulk cylindrotheca closterium pennate diatoms and of the single cells. *RSC advances, 6*(49), 42899-42910

Premvardhan, L., Bordes, L., Beer, A., Buchel, C., & Robert, B. (2009). Carotenoid structures and environments in trimeric and oligomeric fucoxanthin chlorophyll a/c2 proteins from resonance Raman spectroscopy. *J Phys Chem B, 113*(37), 12565-12574.

Premvardhan, L., Robert, B., Beer, A., & Büchel, C. (2010). Pigment organization in fucoxanthin chlorophyll a/c2 proteins (FCP) based on resonance Raman spectroscopy and sequence analysis. *Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1797*(9), 1647-1656

Pytlik, N., Klemmed, B., Machill, S., Eychmüller, A., & Brunner, E. (2019). In vivo uptake of gold nanoparticles by the diatom Stephanopyxis turris. *Algal research, 39*, 101447.

Rimet, F., & Bouchez, A. (2012). Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. *Knowledge and management of Aquatic Ecosystems*(406), 01.

Rüger, J., Mondol, A. S., Schie, I. W., Popp, J., & Krafft, C. (2019). High-throughput screening Raman microspectroscopy for assessment of drug-induced changes in diatom cells. *Analyst, 144*(15), 4488-4492.

Rüger, J., Unger, N., Schie, I. W., Brunner, E., Popp, J., & Krafft, C. (2016). Assessment of growth phases of the diatom Ditylum brightwellii by FT-IR and Raman spectroscopy. *Algal research, 19*, 246-252.

Winter, M. J., Redfern, W. S., Hayfield, A. J., Owen, S. F., Valentin, J.-P., & Hutchinson, T. H. (2008). Validation of a larval zebrafish locomotor assay for assessing the seizure liability of early-stage development drugs. *Journal of pharmacological and toxicological methods, 57*(3), 176-187.

Wood, B. R., Heraud, P., Stojkovic, S., Morrison, D., Beardall, J., & McNaughton, D. (2005). A portable Raman acoustic levitation spectroscopic system for the identification and environmental monitoring of algal cells. *Analytical chemistry, 77*(15), 4955-4961.

Yuan, P., He, H. P., Wu, D. Q., Wang, D. Q., & Chen, L. J. (2004). Characterization of diatomaceous silica by Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 60*(12), 2941-2945.

Chapter 6 – Conclusions and future perspectives

- Studies involving the structure, location and conformation of diatom cell components and their variation under different conditions provided a strong foundation for the development of environmental protocols using Raman spectroscopy in diatoms. A variety of studies were summarized to encourage further research on the application of Raman spectroscopy as a tool to assess physiological changes and water quality under a changing climate.
- Fourteen different bands were found in diatom Raman spectra. These bands were mainly assigned to pigments (e.g. carotenoids, *Chl a*) and frustule components. Raman profiles were depecited for various species, including *Cyclotella stelligera, Achnanthidium minutissimum* and *Amphora pediculus*.
- Bands also differed according to the lake of origin. The interaction between the factor lake and species greatly influenced the responses measure through Raman spectroscopy.
- Artificial Neural Network models based on Raman data were highly accurate to diagnose lakes and could also be used for identification of various *taxa*.
- Despite the proximity and connectivity of the three lakes in Parque da Cidade, all of them were successfully diagnosed. It was also evident that models only based in Raman variables can be created eliminating the inherent constraints associated to diatom morphometric identification.

The results obtained raised interesting research questions, opening perspectives for future research. In the future, this methodology applied to diatoms should be employed also in laboratory and controlled studies using monocultures and complementary methods such as biomaker and gene expression determinations. This simpler approach combined with chemical and chemometric methods will further help to understand individual physiological responses to contaminants and conditions related to climate change such as water acidification. Moreover, it will shed light on the specific pigments and other molecules indicated by the Raman spectra, allowing to relate changes in molecular conformations in the diverse diatom cell components with changes in Raman band width and frequency. The astonishing results of ANN methods based on Raman spectra of diatoms in detecting closely located and connected lakes point to the need to extend its application to other freshwater ecosystems and test its applicability to assess

restauration of polluted ecosystems where recovery actions have been applied. In short, a variety of investigation paths were identified after the present dissertation about the use of RS applied to diatoms for environmental diagnostic testing

Appendix I

Table I - Raman bands corresponding to different cell components, species tested, laser wavelengths used and the respective assignments according to the literature. YP - Sample from Yuanjiawan deposit in Shengxian county of Zhejiang province, China; BP - Sample from Buchang deposit in Haikang county of Guangdong province, China.

RS band position (cm-1)	**Species**	**Component**	**Laser Wavelength (nm)**	**Assignment**	**Reference**
800	*Coscionodiscus wailesii*	Frustule	488	Silica network vibrations and deformations	Kammer et al. 2010
1060	*Coscionodiscus wailesii*	Frustule	488	Unknown inorganic compound	Kammer et al. 2010
1450	*Coscionodiscus wailesii*	Frustule	488	CH deformation modes	Kammer et al. 2010
2930	*Coscionodiscus wailesii*	Frustule	488	CH valence vibrations	Kammer et al. 2010
1050-1100	*Coscionodiscus wailesii*	Frustule	532	SiO/SiO2 symmetric stretching	De Tommasi et al., 2016
1050-1100	*Coscionodiscus wailesii*	Frustule	532	SiO, SiO2, SiO3 symmetric stretching	De Tommasi et al. 2018

1150–1200	*Coscionodiscus wailesii*	Frustule	532	C=S	De Tommasi et al., 2016; 2018
1450-1550	*Coscionodiscus wailesii*	Frustule	532	CH, CH2, CH3, C=C	De Tommasi et al., 2016; 2018
2100-2250	*Coscionodiscus wailesii*	Frustule	532	CC	De Tommasi et al., 2016
2100-2250	*Coscionodiscus wailesii*	Frustule	532	C≡C, C≡N	De Tommasi et al. 2018
2550-2600	*Coscionodiscus wailesii*	Frustule	532	SH stretching	De Tommasi et al., 2016; 2018
2800-3000	*Coscionodiscus wailesii*	Frustule	532	C-H stretching vibration.	De Tommasi et al., 2016
2800-3000	*Coscionodiscus wailesii*	Frustule	532	CH, CH2, CH3 stretching	De Tommasi et al. 2018
450-550	*Coscionodiscus wailesii*	Frustule	532	Si-O-Si stretching vibrations	De Tommasi et al., 2016; 2018
480 and 970	*Coscionodiscus wailesii*	Frustule	488	Stretching vibrations of silica	Kammer et al. 2010
800-850	*Coscionodiscus wailesii*	Frustule	532	SiO4 symmetric stretching	De Tommasi et al., 2016

900-1050	*Coscionodiscus wailesii*	Frustule	532	SiO2, Si-OH stretching	De Tommasi et al. 2018
900-950	*Coscionodiscus wailesii*	Frustule	532	SiO3 /SiO2 symmetric stretching	De Tommasi et al., 2016; 2018
950-1050	*Coscionodiscus wailesii*	Frustule	532	CC aromatic ring chain vibrations	De Tommasi et al., 2016
446	N.A.	Frustule - Amorphous Silica	488	Si-O-Si bond rocking and bending in the SiO4 tetrahedra	Biswas et al., 2018
492	N.A.	Frustule - Amorphous Silica	488	Three and four membered ring of silicon (siloxane rings) or structural defects associated with the broken Si-O-Si bonds	Biswas et al., 2018
605	N.A.	Frustule - Amorphous Silica	488	Three and four membered ring of oxygen (siloxane rings) or structural defects associated with the broken Si-O-Si bonds	Biswas et al., 2018
1050 and 1200	N.A.	Frustule - Amorphous Silica	488	Symmetrical stretching of silicon and oxygen in the silicate tetrahedral with non-bridging oxygen atom	Biswas et al., 2018
373	Sample YP	Frustule - Diatomite	632.8	O–Si–O deformation (silicate impurities)	Yuan et al. 2004
440	Sample BP	Frustule - Diatomite	632.8	Unknown	Yuan et al. 2004

493	Sample BP	Frustule - Diatomite	632.8	O3SiOH tetrahedral vibration	Yuan et al. 2004
495	Sample YP	Frustule - Diatomite	632.8	O3SiOH tetrahedral vibration	Yuan et al. 2004
607	Sample BP e YP	Frustule - Diatomite	632.8	(SiO)3-ring breathing	Yuan et al. 2004
1076	Sample YP	Frustule - Diatomite	632.8	Si–O–Si asymmetric stretch	Yuan et al. 2004
1098	Sample BP	Frustule - Diatomite	632.8	Si–O–Si asymmetric stretch	Yuan et al. 2004
300; 350 and 380	Sample BP	Frustule - Diatomite	632.8	O–Si–O deformation (silicate impurities)	Yuan et al. 2004
698 and 800	Sample YP	Frustule - Diatomite	632.8	Si–O–Si symmetric stretch	Yuan et al. 2004
703 and 794	Sample BP	Frustule - Diatomite	632.8	Si–O–Si symmetric stretch	Yuan et al. 2004
463	N.A.	Frustule - Quartz	488	symmetric stretching- bending modes of Si-O-Si	Biswas et al., 2018
>600	N.A.	Frustule - Quartz	488	Si-O stretching modes	Biswas et al., 2018
455	*Chaetoceros calcitrans*	Frustule (heated up to 800 ºC)	514.5	Similar cristalization to partially ordered law-quartz or cristolabite percusor	Arasuna & Okuno, 2018
405	*Chaetoceros calcitrans*	Frustule (heated up tp 1200ºC)	514.5	Similar cristalization to low-cristolabite	Arasuna & Okuno, 2018

1080	*Thalassiosira pseudonana*	Lipids - Liquid-phase fatty acids	785	C-C gauche stretching	Meksiarun et al., 2015
1120	*Thalassiosira pseudonana*	Lipids - Liquid-phase fatty acids	785	C-C gauche stretching	Meksiarun et al., 2014;2015
865	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	C-C bending	Meksiarun et al., 2015
931	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	C-C stretching	Meksiarun et al., 2015
974	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	=C-H out-of-plane bending	Meksiarun et al., 2015
1265	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	=C-H in-plane bending	Meksiarun et al., 2014;2015
1304	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	CH2 twisting bending	Meksiarun et al., 2014;2015
1440	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	CH2 scissor bending	Meksiarun et al., 2014;2015
1660	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	C=C cis stretching	Meksiarun et al., 2015
1748	*Thalassiosira pseudonana*	Lipids - Polyunsaturated fatty acids	785	C=O stretching	Meksiarun et al., 2015
1068	*Thalassiosira pseudonana*	Lipids - Solid fatty acids	785	C-C trans stretching	Meksiarun et al., 2015
1652	*Pseudonitzschia* sp.	Other Substances - Domoic Acid	251	Coupling of the symmetric C=C mode	Wu et al. 2000

3008	*Nitzschia* sp.	Other substances - EPS (Aggregated cells)	532	C=C bonds indicating the insaturation of alkyl chains	Laviale et al. 2019
1302; 1440; 1656; 1730 and 1746	*Nitzschia* sp.	Other substances - EPS (Aggregated cells)	532	Triacylglycerol lipids	Laviale et al. 2019
800-900 and 1000-1200	*Nitzschia* sp.	Other substances - EPS (Aggregated cells)	532	Carbohydrate moieties possibly linked to the lipids forming hydrophobic EPS	Laviale et al. 2019
1437	*Nitzschia* sp.	Other Substances - EPS (Mucilage strands)	514	—CH2— deformation	Chen et al. 2019
1655	*Nitzschia* sp.	Other Substances - EPS (Mucilage strands)	514	C=O	Chen et al. 2019
2882	*Nitzschia* sp.	Other Substances - EPS (Mucilage strands)	514	—CH2— symetric and asymetric stretch vibrations of a carbohydrate	Chen et al. 2019
2936	*Nitzschia* sp.	Other Substances - EPS (Mucilage strands)	514	—CH3 symetric and asymetric stretch vibrations of a carbohydrate	Chen et al. 2019
1088 and 1090	*Nitzschia* sp.	Other substances - EPS (Mucilage trails and strands)	514	Polysacharide	Chen et al. 2019
1612 and 1619	*Nitzschia* sp.	Other substances - EPS (Mucilage trails and strands)	514	Tyrosine	Chen et al. 2019
594 and 599	*Nitzschia* sp.	Other substances - EPS (Mucilage trails and strands)	514	Phenylalanine	Chen et al. 2019

1000-1150	*Nitzschia* sp.	Other substances - EPS (Single cells)	532	Carbohydrates	Laviale et al. 2019
1003; 1250; 1600-1700	*Nitzschia* sp.	Other substances - EPS (Single cells)	532	Proteins	Laviale et al. 2019
1303; 1454 and 1735	*Nitzschia* sp.	Other substances - EPS (Single cells)	532	Fatty esters	Laviale et al. 2019
1157	*Nitzschia* sp.	Pigments - Beta-carotene	N.A.	-CC; CH	Supryia et al., 2014
1525	*Nitzschia* sp.	Pigments - Beta-carotene	N.A.	C=C	Supryia et al., 2018
918	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
1002	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
1014	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
1014	*Dytilium brightwellii*	Pigments - Carotenoids	785	CH3 streching modes (?)	Ruger et al. 2016
1048	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
1114	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
1162	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008

1162	*Dytilium brightwellii*	Pigments - Carotenoids	785	C-C streching modes	Ruger et al. 2016
1180	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
1180	*Dytilium brightwellii*	Pigments - Carotenoids	785	CH deformation modes	Ruger et al. 2016
1224	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
1528	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	C=C streching modes	Legesse et al. 2008
1528	*Dytilium brightwellii*	Pigments - Carotenoids	785	C=C streching modes	Ruger et al. 2016
1608	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Carotenoids	785	N.A.	Legesse et al. 2008
744	*Nitzschia* sp.	Pigments - Chl a	N.A.	-H-C-O-; -C-C-C-	Supryia et al., 2012
745	*Dytilium brightwellii*	Pigments - Chl a	785	H-C-O deformation modes, N-C-C deformation modes	Ruger et al. 2016
915	*Nitzschia* sp.	Pigments - Chl a	N.A.	-N-C-C-; -C-C-C-	Supryia et al., 2013
917	*Dytilium brightwellii*	Pigments - Chl a	785	N-C-C deformation modes, C-C-C deformation modes	Ruger et al. 2016
986	*Dytilium brightwellii*	Pigments - Chl a	785	CH3 deformation modes	Ruger et al. 2016

1224	*Dytilium brightwellii*	Pigments - Chl a	785	N-C stretching modes	Ruger et al. 2016
1327	*Nitzschia* sp.	Pigments - Chl a	N.A.	-CN; -CH	Supryia et al., 2016
1328	*Dytilium brightwellii*	Pigments - Chl a	785	C-N streching modes	Ruger et al. 2016
1495	*Nitzschia* sp.	Pigments - Chl a	N.A.	C-C; -CH3	Supryia et al., 2017
1605	*Nitzschia* sp.	Pigments - Chl a	N.A.	CC	Supryia et al., 2019
1187	*Nitzschia* sp.	Pigments - Chl a and Beta-carotene	N.A.	-CH; -N-C	Supryia et al., 2015
1438	*Dytilium brightwellii*	Pigments - Chl a and carotenoids	785	CH3 deformation modes	Ruger et al. 2016
1663	*Cyclotella meneghiniana*	Pigments - Chl a in oligomeric FCP	406.7	C=O keto-carbonyl groups with strong hydrogen bonding	Premvardhan et al. 2010
1685	*Cyclotella meneghiniana*	Pigments - Chl a in oligomeric FCP	406.7	C=O keto-carbonyl groups with no or weak hydrogen bonding	Premvardhan et al. 2010
1665	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric and oligomeric FCP	413,1 and 441,6	C=O keto-carbonyl groups with strong hydrogen bonding	Premvardhan et al. 2010
1675	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric and oligomeric FCP	413,1 and 441,6	C=O keto-carbonyl groups with moderate hydrogen bonding	Premvardhan et al. 2010

1677	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric and oligomeric FCP	406.7	C=O keto-carbonyl groups with moderate hydrogen bonding	Premvardhan et al. 2010
1690	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric and oligomeric FCP	413,1 and 441,6	C=O keto-carbonyl groups with no or weak hydrogen bonding	Premvardhan et al. 2010
1610 to 1615	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric and oligomeric FCP	441.6	CaCm methine bridge	Premvardhan et al. 2010
1654	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric FCP	406.7	C=O keto-carbonyl groups with strong hydrogen bonding	Premvardhan et al. 2010
1688	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric FCP	406.7	C=O keto-carbonyl groups with no or weak hydrogen bonding	Premvardhan et al. 2010
1610 and 1615	*Cyclotella meneghiniana*	Pigments - Chl a in trimeric FCP	406,7 and 413,1	CaCm methine bridge	Premvardhan et al. 2010
1670	*Nitzschia* sp.	Pigments - Chl a or Protein	N.A.	C=O or amide 1	Supryia et al., 2020
1615	*Cyclotella meneghiniana*	Pigments - Chl c2 in oligomeric FCP	457.9	CaCm methine bridge	Premvardhan et al. 2010
1620	*Cyclotella meneghiniana*	Pigments - Chl c2 in oligomeric FCP	457.9	C3^1=C3^2; C8^1=C8^2 vinyl groups	Premvardhan et al. 2010
1355 and 1359	*Cyclotella meneghiniana*	Pigments - Chl c2 in oligomeric FCP	476.5	C-N ring breathing mode	Premvardhan et al. 2010
1355 and 1360	*Cyclotella meneghiniana*	Pigments - Chl c2 in oligomeric FCP	457.9	C-N ring breathing mode	Premvardhan et al. 2010

1676 and 1690	*Cyclotella meneghiniana*	Pigments - Chl c2 in oligomeric FCP	457,9 and 476,5	C=O keto carbonyl	Premvardhan et al. 2010
1615	*Cyclotella meneghiniana*	Pigments - Chl c2 in trimeric FCP	457.9	CaCm methine bridge	Premvardhan et al. 2010
1620	*Cyclotella meneghiniana*	Pigments - Chl c2 in trimeric FCP	457.9	C3^1=C3^2; C8^1=C8^2 vinyl groups	Premvardhan et al. 2010
1355 and 1360	*Cyclotella meneghiniana*	Pigments - Chl c2 in trimeric FCP	476.5	C-N ring breathing mode	Premvardhan et al. 2010
1355 and 1362	*Cyclotella meneghiniana*	Pigments - Chl c2 in trimeric FCP	457.9	C-N ring breathing mode	Premvardhan et al. 2010
1677 and 1695	*Cyclotella meneghiniana*	Pigments - Chl c2 in trimeric FCP	457,9 and 476,5	C=O keto carbonyl	Premvardhan et al. 2010
746	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Chla	785	N.A.	Legesse et al. 2008
918	*Gomphonema parvulum* e *Phaeodactylum tricornutum*	Pigments - Chla	441.6	N.A.	Wagner & Weidlich, 1986
988	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Chla	785	N.A.	Legesse et al. 2008
1328	*Ditylum brightwellii* and *Stephanopyxis turris*	Pigments - Chla	785	N.A.	Legesse et al. 2008
340	*Gomphonema parvulum* e *Phaeodactylum tricornutum*	Pigments - Chlc1+Chlc2	457.9	In-plane bending modes of Cb peripheral groups	Wagner & Weidlich, 1986

714	*Gomphonema parvulum* e *Phaeodactylum tricornutum*	Pigments - Chlc1+Chlc2	457.9	In-plane deformation of pyrrole rings	Wagner & Weidlich, 1986
800	*Gomphonema parvulum* e *Phaeodactylum tricornutum*	Pigments - Chlc1+Chlc2	457.9	Deformation of peripheral groups	Wagner & Weidlich, 1986
1116	Standard solution	Pigments - Chlc1+Chlc2	457.9	Cb-Cα streching	Wagner & Weidlich, 1986
1137	Standard solution	Pigments - Chlc1+Chlc2	457.9	C-N streching and CNC bending modes	Wagner & Weidlich, 1986
1361	*Gomphonema parvulum* e *Phaeodactylum tricornutum*	Pigments - Chlc1+Chlc2	457.9	C-N breathing mode	Wagner & Weidlich, 1986
1539	Standard solution	Pigments - Chlc1+Chlc2	457.9	CB=CB streching	Wagner & Weidlich, 1986
1562	Standard solution	Pigments - Chlc1+Chlc2	457.9	CB=CB streching	Wagner & Weidlich, 1986
1622	Standard solution	Pigments - Chlc1+Chlc2	457.9	Cα=Cβ vinyl streching mode coupled with na in-plaine deformation of CβH2 groups	Wagner & Weidlich, 1986
1692	Standard solution	Pigments - Chlc1+Chlc2	457.9	C=O streching mode	Wagner & Weidlich, 1986
1158-1159	*Haslea ostrearia* and *Haslea provincialis*	Pigments - Chloroplasts	514.5	Partial light scattering of the chloroplasts	Gastineau et al. 2012

1520-1522	*Haslea ostrearia* and *Haslea provincialis*	Pigments - Chloroplasts	514.5	Partial light scattering of the chloroplasts	Gastineau et al. 2012
1555	*Cyclotella meneghiniana*	Pigments - Chls	413,1 and 457,9	C=C streching modes	Alexandre et al. 2014
1612, 1655, 1679, 1694	*Cyclotella meneghiniana*	Pigments - Chls	413,1 and 441,6	Carbonyl streching modes	Alexandre et al. 2014
1640 and 1710	*Cyclotella meneghiniana*	Pigments - Chls	457.9	Carbonyl conjugated groups stretching modes	Alexandre et al. 2014
980	*Cyclotella meneghiniana*	Pigments - Ddx/Dtx	488 and 496,5	hydrogen out of plain wagging modes	Alexandre et al. 2014
983	*Cyclotella meneghiniana*	Pigments - Ddx/Dtx	476.5	hydrogen out of plain wagging modes	Alexandre et al. 2014
965	*Cyclotella meneghiniana*	Pigments - Fx	570	hydrogen out of plain wagging modes	Alexandre et al. 2014
1530	*Cyclotella meneghiniana*	Pigments - Fx blue	413,1 and 476,5	C=C streching modes	Alexandre et al. 2014
1530.4	*Cyclotella meneghiniana*	Pigments - Fx green	514.5	C=C streching modes	Alexandre et al. 2014
1532.6	*Cyclotella meneghiniana*	Pigments - Fx green	488	C=C streching modes	Alexandre et al. 2014
1531	*Cyclotella meneghiniana*	Pigments - Fx green/blue of oligomeric FCP	441,6-488	C=C streching modes	Premvardhan et al. 2009

1532	*Cyclotella meneghiniana*	Pigments - Fx green/blue of trimeric FCP	441,6; 476,5 and 488	C=C streching modes	Premvardhan et al. 2009
1533	*Cyclotella meneghiniana*	Pigments - Fx green/blue of trimeric FCP	457.9	C=C streching modes	Premvardhan et al. 2009
960	*Cyclotella meneghiniana*	Pigments - Fx in FCP	413,7 to 570	hydrogen out of plain wagging modes	Premvardhan et al. 2009
1000 and 1020	*Cyclotella meneghiniana*	Pigments - Fx in FCP	413,7 to 570	CH3 in-plane wagging modes	Premvardhan et al. 2009
1150-1210	*Cyclotella meneghiniana*	Pigments - Fx in FCP	413,7 to 570	C-C stretching modes	Premvardhan et al. 2009
1529.5	*Cyclotella meneghiniana*	Pigments - Fx red	570	C=C streching modes	Alexandre et al. 2014
1530.7	*Cyclotella meneghiniana*	Pigments - Fx red	528.5	C=C streching modes	Alexandre et al. 2014
1655-1657	*Cyclotella meneghiniana*	Pigments - Fx red in oligomeric FCP	540-550	Carbonyl groups	Premvardhan et al. 2009
1645-1647	*Cyclotella meneghiniana*	Pigments - Fx red in trimeric FCP	540-550	Carbonyl groups	Premvardhan et al. 2009
1530.1	*Cyclotella meneghiniana*	Pigments - Fx red of oligomeric FCP	550; 560 and 570	C=C streching modes	Premvardhan et al. 2009
1531.3	*Cyclotella meneghiniana*	Pigments - Fx red of oligomeric FCP	514,5 and 528,7	C=C streching modes	Premvardhan et al. 2009

1532.3	*Cyclotella meneghiniana*	Pigments - Fx red of oligomeric FCP	540	C=C streching modes	Premvardhan et al. 2009
1529	*Cyclotella meneghiniana*	Pigments - Fx red of trimeric FCP	560 and 570	C=C streching modes	Premvardhan et al. 2009
1531	*Cyclotella meneghiniana*	Pigments - Fx red of trimeric FCP	514,5 and 540,7	C=C streching modes	Premvardhan et al. 2009
1532.6	*Cyclotella meneghiniana*	Pigments - Fx red of trimeric FCP	540 and 550	C=C streching modes	Premvardhan et al. 2009
1122-1125, 1160 and 1125	*Haslea karadagensis*	Pigments - Gametes and apices	514.5	N.A.	Gastineau et al. 2012
1300, 1442 and 1650	*Haslea ostrearia*	Pigments - Gametes and apices	514.5	N.A.	Gastineau et al. 2012
1162	*Haslea ostrearia* and *Haslea provincialis*	Pigments - Marennine and Marennine-like pigments	514.5	C-C streching modes	Gastineau et al. 2012
1527	*Haslea ostrearia* and *Haslea provincialis*	Pigments - Marennine and Marennine-like pigments	514.5	C=C streching modes	Gastineau et al. 2012
1163, 1211, 1294, 1464, 1577, 1607, 1652, 1335, 1400, 1470, 1540 and 1610	*Haslea ostrearia* and *Haslea provincialis*	Pigments - Marennine and Marennine-like pigments	514.5	N.A.	Gastineau et al. 2016

Appendix II

Table IA - Statistical correlations considering all the areas (A) of the diatom RS bands. Correlations significant at p<1x10E-15 are highlighted in red.

	Means	Std.Dev.	A867	A920	A963	A1013	A1160	A1180	A1198	A1270	A1315	A1390	A1445	A1526	A1606	A1656
A867	12081.61	9078.66	1.00	0.41	0.49	0.75	0.83	0.43	0.47	0.76	0.52	0.63	0.19	0.79	0.36	0.53
A920	4700.27	6580.65	0.41	1.00	0.29	0.38	0.46	0.23	0.17	0.40	0.31	0.38	0.06	0.41	0.13	0.26
A963	69474.84	82683.83	0.49	0.29	1.00	0.59	0.57	0.35	0.34	0.51	0.44	0.44	0.51	0.63	0.39	0.43
A1013	225760.46	169562.94	0.75	0.38	0.59	1.00	0.83	0.43	0.48	0.78	0.49	0.61	0.27	0.89	0.54	0.64
A1160	708730.87	480419.42	0.83	0.46	0.57	0.83	1.00	0.41	0.50	0.88	0.62	0.71	0.20	0.92	0.38	0.67
A1180	65746.19	84113.43	0.43	0.23	0.35	0.43	0.41	1.00	0.24	0.49	0.24	0.36	0.06	0.48	0.21	0.36
A1198	418888.75	555198.37	0.47	0.17	0.34	0.48	0.50	0.24	1.00	0.52	0.34	0.38	0.22	0.51	0.36	0.31
A1270	94857.86	71450.48	0.76	0.40	0.51	0.78	0.88	0.49	0.52	1.00	0.53	0.64	0.19	0.87	0.38	0.65
A1315	5374.89	6184.77	0.52	0.31	0.44	0.49	0.62	0.24	0.34	0.53	1.00	0.45	0.25	0.56	0.23	0.40
A1390	11719.17	12734.10	0.63	0.38	0.44	0.61	0.71	0.36	0.38	0.64	0.45	1.00	0.13	0.68	0.30	0.52
A1445	53524.27	130230.59	0.19	0.06	0.51	0.27	0.20	0.06	0.22	0.19	0.25	0.13	1.00	0.30	0.42	0.17
A1526	894517.18	627583.75	0.79	0.41	0.63	0.89	0.92	0.48	0.51	0.87	0.56	0.68	0.30	1.00	0.46	0.70
A1606	50841.94	72392.87	0.36	0.13	0.39	0.54	0.38	0.21	0.36	0.38	0.23	0.30	0.42	0.46	1.00	0.28
A1656	16267.82	16717.46	0.53	0.26	0.43	0.64	0.67	0.36	0.31	0.65	0.40	0.52	0.17	0.70	0.28	1.00

Table IB - Statistical correlations considering all the areas (A) of the diatom RS bands normalized with the Area of the band 1526 cm-1. Correlations significant at p<1x10E-15 are highlighted in red.

	Means	Std.Dev.	A867	A920	A963	A1013	A1160	A1180	A1390	A1315	A1198	A1270	A1445	A1606	A1656
A867	0.01	0.01	1.00	0.05	-0.02	0.05	0.17	0.02	0.11	0.09	0.01	0.11	-0.04	-0.07	-0.07
A920	0.01	0.01	0.05	1.00	-0.01	0.02	0.10	0.02	0.06	0.05	-0.02	0.06	-0.06	-0.05	-0.08
A963	0.08	0.06	-0.02	-0.01	1.00	0.39	-0.12	-0.02	0.03	0.10	0.04	-0.13	0.33	0.28	-0.04
A1013	0.26	0.11	0.05	0.02	0.39	1.00	-0.04	0.00	-0.01	-0.01	-0.01	-0.03	0.03	0.09	0.02
A1160	0.81	0.11	0.17	0.10	-0.12	-0.04	1.00	-0.03	0.03	0.11	0.01	0.16	-0.28	-0.29	-0.03
A1180	0.08	0.06	0.02	0.02	-0.02	0.00	-0.03	1.00	-0.05	-0.08	-0.08	-0.04	-0.11	-0.05	-0.04
A1390	0.01	0.01	0.11	0.06	0.03	-0.01	0.03	-0.05	1.00	0.14	-0.01	0.06	-0.06	-0.03	0.01
A1315	0.01	0.01	0.09	0.05	0.10	-0.01	0.11	-0.08	0.14	1.00	0.03	0.05	0.00	-0.03	-0.03
A1198	0.47	0.32	0.01	-0.02	0.04	-0.01	0.01	-0.08	-0.01	0.03	1.00	0.23	0.11	0.14	-0.03
A1270	0.11	0.03	0.11	0.06	-0.13	-0.03	0.16	-0.04	0.06	0.05	0.23	1.00	-0.11	-0.09	0.07
A1445	0.06	0.12	-0.04	-0.06	0.33	0.03	-0.28	-0.11	-0.06	0.00	0.11	-0.11	1.00	0.52	-0.04
A1606	0.06	0.07	-0.07	-0.05	0.28	0.09	-0.29	-0.05	-0.03	-0.03	0.14	-0.09	0.52	1.00	-0.05
A1656	0.02	0.02	-0.07	-0.08	-0.04	0.02	-0.03	-0.04	0.01	-0.03	-0.03	0.07	-0.04	-0.05	1.00

Table IC - Statistical Correlations considering all the width (W) of the diatom RS bands. Correlations significant at p<1x10E-15 are highlighted in red.

	Means	Std.Dev.	W867	W920	W963	W1013	W1160	W1180	W1198	W1270	W1315	W1390	W1445	W1526	W1606	W1656
W867	12.56	6.00	1.00	0.01	-0.01	0.04	0.09	-0.10	0.04	0.16	0.09	0.09	0.09	0.00	0.02	0.09
W920	7.34	5.65	0.01	1.00	-0.06	0.01	0.14	-0.03	-0.06	0.04	0.16	0.10	-0.10	-0.02	-0.09	0.06
W963	21.25	6.85	-0.01	-0.06	1.00	0.31	-0.10	0.01	0.15	-0.04	-0.03	0.05	0.38	0.32	0.33	0.00
W1013	24.43	4.39	0.04	0.01	0.31	1.00	0.04	-0.10	0.21	0.08	-0.03	0.05	0.30	0.45	0.38	0.10
W1160	20.46	1.35	0.09	0.14	-0.10	0.04	1.00	-0.34	0.00	-0.09	0.20	0.15	-0.08	0.01	-0.08	0.21
W1180	8.72	2.97	-0.10	-0.03	0.01	-0.10	-0.34	1.00	-0.13	-0.07	-0.16	-0.08	-0.09	0.03	0.01	-0.09
W1198	31.37	7.16	0.04	-0.06	0.15	0.21	0.00	-0.13	1.00	0.09	-0.04	-0.02	0.30	0.16	0.30	-0.06
W1270	16.52	3.31	0.16	0.04	-0.04	0.08	-0.09	-0.07	0.09	1.00	0.02	0.06	0.05	0.16	0.02	0.14
W1315	5.09	3.44	0.09	0.16	-0.03	-0.03	0.20	-0.16	-0.04	0.02	1.00	0.19	-0.03	-0.06	-0.06	0.16
W1390	10.43	6.36	0.09	0.10	0.05	0.05	0.15	-0.08	-0.02	0.06	0.19	1.00	0.02	0.14	0.02	0.28
W1445	19.61	13.82	0.09	-0.10	0.38	0.30	-0.08	-0.09	0.30	0.05	-0.03	0.02	1.00	0.27	0.50	0.01
W1526	20.52	0.86	0.00	-0.02	0.32	0.45	0.01	0.03	0.16	0.16	-0.06	0.14	0.27	1.00	0.22	0.07
W1606	14.19	7.47	0.02	-0.09	0.33	0.38	-0.08	0.01	0.30	0.02	-0.06	0.02	0.50	0.22	1.00	-0.04
W1656	10.75	5.83	0.09	0.06	0.00	0.10	0.21	-0.09	-0.06	0.14	0.16	0.28	0.01	0.07	-0.04	1.00

Table ID - Statistical correlations considering all the frequencies (F) of the diatom RS bands. Correlations significant at p<1x10E-15 are highlighted in red.

	Means	Std.Dev.	F867	F920	F963	F1013	F1160	F1180	F1198	F1270	F1315	F1390	F1445	F1526	F1606	F1656
F867	827.11	183.59	1.00	0.14	-0.01	-0.03	0.02	-0.01	-0.19	0.32	0.18	0.23	0.21	0.04	0.20	0.24
F920	794.96	315.09	0.14	1.00	0.09	0.03	0.09	0.09	-0.19	0.18	0.26	0.13	0.09	0.16	0.24	0.24
F963	961.26	34.35	-0.01	0.09	1.00	0.26	0.19	-0.01	-0.24	0.40	0.08	0.14	-0.01	0.19	0.33	0.10
F1013	1012.69	1.17	-0.03	0.03	0.26	1.00	0.54	-0.11	0.00	0.04	0.04	0.02	-0.04	0.54	0.30	0.05
F1160	1160.54	0.87	0.02	0.09	0.19	0.54	1.00	0.02	-0.18	0.21	0.18	0.08	-0.01	0.76	0.28	0.17
F1180	1178.43	42.03	-0.01	0.09	-0.01	-0.11	0.02	1.00	-0.03	-0.01	0.08	0.13	0.00	0.05	-0.01	0.10
F1198	1197.82	1.95	-0.19	-0.19	-0.24	0.00	-0.18	-0.03	1.00	-0.34	-0.21	-0.16	0.00	-0.24	-0.37	-0.35
F1270	1260.73	110.44	0.32	0.18	0.40	0.04	0.21	-0.01	-0.34	1.00	0.16	0.15	0.10	0.18	0.40	0.26
F1315	1103.06	481.18	0.18	0.26	0.08	0.04	0.18	0.08	-0.21	0.16	1.00	0.17	0.13	0.20	0.08	0.31
F1390	1298.57	348.70	0.23	0.13	0.14	0.02	0.08	0.13	-0.16	0.15	0.17	1.00	0.08	0.03	0.16	0.21
F1445	1423.17	184.35	0.21	0.09	-0.01	-0.04	-0.01	0.00	0.00	0.10	0.13	0.08	1.00	0.01	0.08	0.15
F1526	1527.04	0.83	0.04	0.16	0.19	0.54	0.76	0.05	-0.24	0.18	0.20	0.03	0.01	1.00	0.36	0.21
F1606	1588.44	170.74	0.20	0.24	0.33	0.30	0.28	-0.01	-0.37	0.40	0.08	0.16	0.08	0.36	1.00	0.28
F1656	1485.55	502.80	0.24	0.24	0.10	0.05	0.17	0.10	-0.35	0.26	0.31	0.21	0.15	0.21	0.28	1.00

Table II- Diatom valve counts and valve percentage found in the three lakes of Oporto City Park. Species >1% abundance in at least one lake are highlighted in yellow.

	Valve count				Valve percentage (%)			
Species	L 1	L2	L 3	Total	L1	L 2	L 3	Total
Achnanthidium exiguum (Grunow) Czarnecki 1994	0	2	6	8	0.0	0.5	1.5	0.6
Achnanthidium minutissimum (Kützing) Czarnecki 1994	4	105	10	119	1.0	24.6	2.5	9.6
Achnanthidium straubianum (Lange-Bertalot) Lange-Bertalot 1999	0	0	6	6	0.0	0.0	1.5	0.5
Achnanthidium subhudsonis (Hustedt) H.Kobayasi 2006	2	0	1	3	0.5	0.0	0.2	0.2
Amphora minutissima W.Smith 1853	0	0	2	2	0.0	0.0	0.5	0.2
Amphora ovalis (Kützing) Kützing 1844	0	4	0	4	0.0	0.9	0.0	0.3
Amphora pediculus (Kützing) Grunow 1875	6	8	96	110	1.5	1.9	23.5	8.9
Amphora veneta Kützing 1844	0	0	5	5	0.0	0.0	1.2	0.4
Cocconeis placentula var. euglypta (Ehrenberg) Grunow 1884	2	0	0	2	0.5	0.0	0.0	0.2
Cocconeis placentula var. lineata (Ehrenberg) Van Heurck 1885	0	2	0	2	0.0	0.5	0.0	0.2
Ctenophora pulchella (Ralfs ex Kützing) D.M.Williams & Round 1986	0	22	6	28	0.0	5.2	1.5	2.3
Cyclotella meneghiniana Kützing 1844	0	2	3	5	0.0	0.5	0.7	0.4
Cyclotella stelligera (Cleve & Grunow) Van Heurck 1882	0	0	27	27	0.0	0.0	6.6	2.2
Cymbella tumida (Brébisson) Van Heurck 1880	0	9	0	9	0.0	2.1	0.0	0.7
Eolimna minima (Grunow) Lange-Bertalot in Moser & al. 1998	4	0	21	25	1.0	0.0	5.1	2.0

Fragilaria crotonensis Kitton 1868	65	4	24	93	15.9	0.9	5.9	7.5
Pseudostaurosira brevistriata (Grunow) D.M.Williams & Round 1988	0	4	12	16	0.0	0.9	2.9	1.3
Fragilaria parva (Grunow) A.Tuji & D.M.Williams 2008	2	0	4	6	0.5	0.0	1.0	0.5
Fragilaria vaucheriae (Kützing) J.B.Petersen 1938	0	0	20	20	0.0	0.0	4.9	1.6
Gomphonema acuminatum Ehrenberg 1832	0	0	4	4	0.0	0.0	1.0	0.3
Gomphonema affine Kützing 1844	4	7	0	11	1.0	1.6	0.0	0.9
Gomphonema clavatum Ehrenberg 1832	2	0	2	4	0.5	0.0	0.5	0.3
Gomphonema exilissimum (Grunow) Lange-Bertalot & E.Reichardt 1996	9	0	0	9	2.2	0.0	0.0	0.7
Gomphonema gracile Ehrenberg 1838	0	0	2	2	0.0	0.0	0.5	0.2
Gomphonema lagenula Kützing 1844	31	0	0	31	7.6	0.0	0.0	2.5
Gomphonema olivaceum (Hornemann) Ehrenberg 1838	0	4	0	4	0.0	0.9	0.0	0.3
Gomphonema parvulum (Kützing) Kützing 1849	128	6	6	140	31.4	1.4	1.5	11.3
Mayamaea permitis (Hustedt) K.Bruder & Medlin 2008	1	1	0	2	0.2	0.2	0.0	0.2
Melosira varians C.Agardh 1827	89	23	7	119	21.8	5.4	1.7	9.6
Navicula cryptocephala Kützing 1844	21	2	7	30	5.1	0.5	1.7	2.4
Navicula cryptotenella Lange-Bertalot 1985	5	1	0	6	1.2	0.2	0.0	0.5
Luticola goeppertiana (Bleisch) D.G.Mann ex J.Rarick, S.Wu, S.S.Lee & Edlund 2017	0	4	0	4	0.0	0.9	0.0	0.3
Navicula gregaria Donkin 1861	12	6	24	42	2.9	1.4	5.9	3.4

Navicula notha J.H.Wallace 1971	0	0	21	21	0.0	0.0	5.1	1.7
Navicula radiosa Kützing 1844	0	0	2	2	0.0	0.0	0.5	0.2
Nitzschia amphibia Grunow 1862	0	16	14	30	0.0	3.8	3.4	2.4
Nitzschia communis Rabenhorst 1860	0	4	0	4	0.0	0.9	0.0	0.3
Nitzschia fonticola (Grunow) Grunow 1881	0	13	0	13	0.0	3.1	0.0	1.0
Nitzschia inconspicua Grunow 1862	1	6	10	17	0.2	1.4	2.5	1.4
Nitzschia palea (Kützing) W.Smith 1856	14	29	8	51	3.4	6.8	2.0	4.1
Nitzschia subcapitellata Hustedt 1939	6	0	2	8	1.5	0.0	0.5	0.6
Pinnularia microstauron (Ehrenberg) Cleve 1891	0	0	2	2	0.0	0.0	0.5	0.2
Planothidium frequentissimum (Lange-Bertalot) Lange-Bertalot 1999	0	1	12	13	0.0	0.2	2.9	1.0
Tabularia tabulata (C.Agardh) Snoeijs 1992	0	123	40	163	0.0	28.9	9.8	13.1
Ulnaria ulna (Nitzsch) Compère 2001	0	18	2	20	0.0	4.2	0.5	1.6
Total	408	426	408	1242	100.0	100.0	100.0	100.0

Table III - Raman continuous input and train, test and validation accuracy of the Artificial Neural Network models determined to predict the different diatom Species, Genus, Family, Order and Subclass. The network architecture used was Multilayer-Perceptron. A – Area, F – Frequency, W – Width; A1526NN – Non-normalized area of the band 1526 cm^{-1}.

Categorical target	Species											Genus		Family		Order		SubClass								
Raman continuous imput	All	W; F.; A1526 NN	A; F	W; F	A; W	F; A1526 NN	W; A1526 NN	A	W	F	All	W; F.; A1526 NN	All	W; F.; A1526 NN	All	W; F.; A1526 NN	all	W; F; A1526 NN	A; F	W; F	A; W	F;A1526NN	W;A1526NN	A	W	F
Train Accuracy (%)	59.4	48.2	46.8	50.9	42.5	38.1	37.9	25.8	35.2	41.6	70.1	55.7	74.0	71.0	84.2	69.2	92.9	78.3	83.8	83.1	83.1	81.7	79.7	78.5	76.9	81.3
Test Accuracy (%)	33.1	34.3	32.0	32.6	26.3	32.0	25.7	20.0	25.7	32.0	52.6	51.4	54.9	55.4	58.3	54.3	76.6	78.9	76.0	75.4	75.4	78.3	76.0	75.4	74.3	76.6
Validation Accuracy (%)	32.0	32.0	31.3	33.7	25.0	31.8	24.4	19.8	25.6	29.0	52.0	49.1	52.6	51.4	53.1	52.0	74.9	76.0	75.6	76.0	75.6	74.4	73.9	74.6	73.9	73.9

Table IV - Sensitivity of the Artificial Neural Network models with the best validation accuracy in predicting the different diatom Subclasses, Orders and Species. The network architecture used was Multilayer-Perceptron. The predictive Subclass and Order models considered all the Raman variables as a continuous input; the predictive species model considered all the frequencies and width of the Raman bands as input data.

Subclass	Sens.	Order	Sens.	Species	Sens.
				Nitzschia amphibia	0%
Bacillariophycidae	80%	*Bacillariales*	50%	*Nitzschia fonticola*	0%
				Nitzschia inconspicua	50%

				Nitzschia palea	29%
				Nitzschia subcapitellata	20%
		Cocconeidales	50%	*Achnanthidium exiguum*	80%
				Achnanthidium minutissimum	100%
				Achnanthidium straubianum	0%
				Planothidium frequentissimum	0%
		Cymbellales	50%	*Cymbella tumida*	0%
				Gomphonema affine	67%
				Gomphonema exilissimum	67%
				Gomphonema lagenula	0%
				Gomphonema parvulum	50%
		Naviculales	50%	*Navicula cryptocephala*	16%
				Navicula cryptotenella	33%
				Navicula gregaria	20%
				Navicula notha	10%
				Eolimna minima	0%
		Thalassiophysales	60%	*Amphora pediculus*	38%
				Amphora veneta	66%
Fragilariophycidae	70%	*Fragilariales*	80%	*Fragilaria crotonensis*	54%
				Fragilaria vaucheriae	0%

				Pseudostaurosira brevistriata	33%
		Licmophorales	50%	*Ctenophora pulchella*	50%
				Tabularia tabulata	25%
				Ulnaria Ulna	40%
Melosirophycidae	50%	*Melosirales*	70%	*Melosira varians*	42%
Thalassiosirophycidae	100%	*Stephanodiscales*	100%	*Cyclotella stelligera*	67%