Analysis of Secondary Structure of proteins by Vibrational Spectroscopy and Self-Organizing Maps

by

Marco Antonio Pinto Corujo

Academic Supervisors: Prof Alison Rodger and Dr Nikola Chmel

Industry Supervisors: Dr Vivian Lindo and Dr Maurizio Muroni

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Analytical Science

University of Warwick, Department of Chemistry

September 2019

Biophysical Chemistry Group

AstraZeneca

EPSRC
Engineering and Physical Sciences Research Council

MAS CDT
Molecular Analytical Science Centre for Doctoral Training

WARWICK
THE UNIVERSITY OF WARWICK

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Dedicated to the memory of my grandfather A. Corujo

ACKNOWLEDGEMENTS

I would like to thank my academic and industry supervisors Alison Rodger, Nikola Chmel, Vivian Lindo and Maurizio Muroni for their guidance and constant support. Also, to the administration personnel, advisory committee members and academics of the CDT Naomi Grew, Christina Forbes, Vasilios Stavros, Jozef Lewandowski, Pat Unwin and Steven Brown for all their help and advice in all the matters surrounding the PhD journey. I would also like to show my gratitude to the sponsors of my project Astrazeneca and the EPSRC.

To my former and new colleagues from the Biophysical Chemistry group and collaborators Meropi Sklepari, Dale Ang, Praveen Amarasinghe, Maria Lizio, Tamara Smidlehner, Daniela Lobo, James Tehan, Claire Broughton, Glen Dorrington, Alan Wemyss, Rhiannon Brooks, Christine Lockey, Alevtina Mikhaylina, Ciaran Guy, Agnieszka Mierek-Adamska, James Coverdale, Monika Kumar, Rahaf Hilouneh, Lisa Igel, Kirsty Alsop, Katie Wood, Pavel Michal and Josef Kapitan.

I would also like to thank to all my friends and people who had to bear me in the past 4 years Lavrentis Galanopoulos, Jhih Ci Yang, Javier Fernández, Antonio Esposito, Ángela García , Maria Muñoz, Neus García, Daniel Pesmed and Cynthia Valdivia, and in special to P. Marco Turano, P. Attaulla Sheagel and F.Clemence Chaudron.

Finally, to my grandfather, uncle, aunt, cousins, mother and sisters.

DECLARATION

To the best of my knowledge, the material contained in this thesis is my own work except where otherwise indicated, cited, or commonly known. The SOM code used in this worked was written by Dr Dale Ang[1] from a previous version coded by Dr Vincent Hall[2], and is based on the algorithm developed by Prof Kohonen[3]. The algorithm is detailly explained in chapter 3. The material in this thesis is submitted to the University of Warwick in partial fulfilment for the degree of Doctor of Philosophy in Biophysical Chemistry, as described in the Graduate School regulations. It has not been submitted to any other university or for any other degree.

ABSTRACT

Antibodies are proteins produced by the immune system and one of the top biopharmaceutical market types due to their applications in oncology therapy among others[4]. As any other protein, their functionality depends on the preservation of their native form which, under certain stressing conditions, can undergo changes at different structural levels and thus loss of their activity[5]. Although mass spectrometry is a powerful technique for primary structure determination, it often fails to give information at higher order levels. In this project we explored the possibilities of vibrational spectroscopic techniques as a tool kit to help ensure the integrity and batch to batch reproducibility in antibody manufacture.

Infrared (IR) and Raman spectra are well known to contain bands (Amide I, II and III) with shapes that correlate to secondary structure (SS)[6,7]. Unlike Circular Dichroism (CD) (the most well-established technique for secondary structure analysis[8]), IR and Raman spectroscopy allow much wider ranges of optical density which makes the analysis of complex pharmaceutical samples more feasible. However, the data processing and extraction of this information are ambiguous and, in many cases, limited by spectral noise and water absorption in IR and fluorescence in Raman.

In this work, data sets of proteins with known SS were collected in both solid and aqueous state by Raman, IR and Raman Optical Activity and used along a neural network algorithm called Self-Organizing Maps (SOMspec) for SS prediction of proteins. It was found that Raman spectroscopy provides the best predictions followed by IR based on the shape of the amide I band. Although the ROA amide I of proteins has also been reported in the literature to correlate to SS content, we failed to predict SS structure by ROA-SOM based on this band. More work needs to be carried out in the future to attempt to predict SS based on the ROA amides II and III instead.

LIST OF ABBREVIATIONS

Ab Antibody

ATR Attenuated Total Reflectance

BMU Best Matching Unit

BSA Bovine Serum Albumin

CCD Charge-Coupled Device

CD Circular Dichroism

CID Collision Induced Dissociation

CPV Cauchy Principal Value

DNA Deoxyribonucleic Acid

FSD Fourier Self-Deconvolution

FTIR Fourier Transform Infrared

Ig Immunoglobulin

IR Infrared

mAb Monoclonal Antibody

MCT Mercury Cadmium Telluride

MRW Mean Residue Weight

NRMSD Normalized Root Mean Square Deviation

PLS Partial Least Squares

PTM Post Translational Modification

RNA Ribonucleic Acid

ROA Raman Optical Activity

RR Raman Resonance

S/N Signal to Noise ratio

SOM Self-Organizing Maps

SOP Standard Operational Procedure

SS Secondary Structure

SSNN Secondary Structure Neural Network

SVD Singular Vector Decomposition

TDC Transition Dipole Coupling

TGS Triglycine Sulphate

UV-Vis Ultraviolet-Visible

1 STRUCTURE, CHARACTERIZATION AND PRODUCTION OF PROTEINS

1.1 INTRODUCTION

Proteins are biomolecules with characteristic 3D shapes that render them different vital functions, e.g., structural, immune, enzymatic, regulatory[9]**.** In the last 20 years, there was a growing interest by biopharmaceutical industry in proteins (specially antibodies) as therapeutic agents[10]. Biopharmaceutical companies research and develop new protein-based drugs with applications in oncology among others[11]. They are generally expressed in cell lines which implies the need for separation from the other cell components[11,12]. The separation and purification procedures involve the use of conditions that might induce structural changes at different levels. For a protein to be functional or active, it needs to be in its native conformation which makes mandatory implementing methods and techniques that ensure the native form of the protein is always preserved throughout its manufacturing[10]. In this chapter we will review the structure of proteins, the existing methods for protein production (expression, extraction and purification) and the existing techniques for protein characterization.

1.2 STRUCTURE OF PROTEINS

1.2.1 General structure of proteins

Proteins are synthesized by ribosomes from RNA molecules (mRNA) in a process during which triplets of ribonucleotides are translated into specific amino acids (translation)[9,13]. Proteins consist of long chains of covalently bound amino acids (primary structure) with groups that interact through hydrogen bonding and intermolecular forces causing the chains to fold into higher order structures referred to as secondary and tertiary[9,14]. These structures can further form oligomers to yield what it is known as quaternary structures[9,15]. An amino acid is a molecule with a central carbon bound to a carboxylic acid, an amine, a hydrogen and a variable R group (Figure 1.1)[9,14,15]. There are 20 different R groups and thus 20 amino acids in nature[9] (Appendices D, Table 0-3). Although all of them -with exemption of

Glycine- exist in two possible enantiomeric configurations (L and R), only L is present in proteins and mammalian metabolism[15].

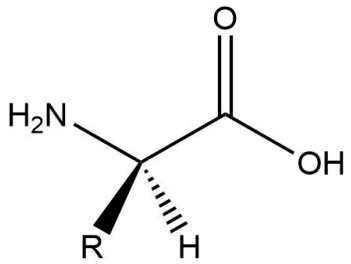

Figure 1.1. General structure of a L – amino acid.

1.2.1.1 Primary structure

The primary structure refers strictly to the sequence of amino acids which are bound covalently through the carbon of an amino acid and the N of the contiguous[14]. This bond is called peptide bond and results from the reaction between -COOH and H_2N- of two neighbour units (Figure 1.2)[14]. These sequences are grown from the terminal -NH_2 to the terminal -COOH and so the terminal -NH_2 is the first amino acid in the sequence and the -COOH the last[15].

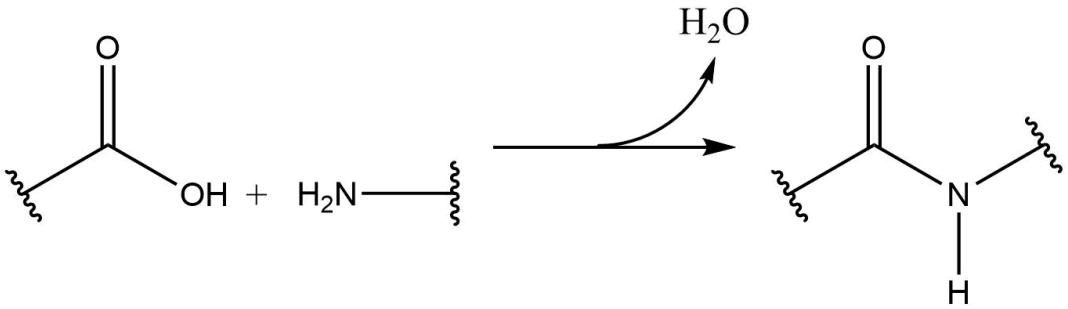

Figure 1.2. Peptide bond formation. Note the carboxylic acid and amine join to give place to an amide group.

Due to mesomeric effects, the order of the C-N bond is between 1 and 2[15]. This partial double bond character is the reason why rotation about the C-N bond is energetically forbidden meaning the substituents cannot interconvert between cis and trans conformations being trans the most favourable one because its less steric hindrance[15]. Generally, the chains are linear but interchain disulphide bridges might form resulting in a branched polypeptide[14]. The primary structure determines the higher order ones and thus the conformation of the protein and it is specified by

genes[9,13]. A change of even one single amino acid in the original sequence could lead to changes in the folding of the protein and its activity as a result[14].

1.2.1.2 Secondary structure

Because of the capability of the bonds N-C and C-C to rotate, the peptide chain can adopt a variety of spatial orientations which are defined by Ramachandran angles φ and ψ (Figure 1.3)[9,13,14]. These angles range from -180° to 180° where 0° is arbitrarily assigned to angles that would result in collision of atoms from one amino acid with atoms from the contiguous[15]. The values for the angles and thus different conformations are determined by hydrogen bond patterns between the carbonyl oxygen and the backbone nitrogen[9,15]. This patterns or motifs are better known as secondary structure[9]. Although different criteria exist to define the secondary structure, we will only use hydrogen bonding pattern-based definition or better known as Dictionary of Secondary Structure of Proteins (DSSP) which divides secondary structure into 8 major classes abbreviated as follows: 3_{10}-helix (G), α-helix (H), π-helix (I), β-sheet (E), β-bridge (B), turn (T), bend (S) and coil (C)[16].

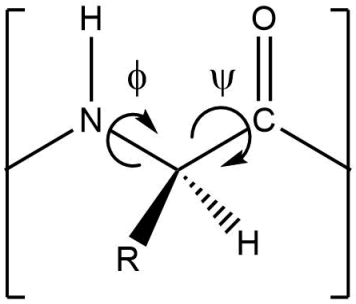

Figure 1.3. Ramachandran angles representation.

Helix is a type of structure where the polypeptide chain turns about an imaginary axis in spiral[14,15]. Three different type of helices occur in depends on the existing number of amino acids per turn: 3_{10}-helix (3), α-helix (3.6), π-helix (4.4)[16]. Helices are mostly counter clock-wise (also called right-handed) and α-helix is the most stable and common one of them[14,15]. Helical structures are usually rich in alanine, leucine, glutamine, glutamine, arginine, methionine and lysine[17].

β-sheet is a pleat-like structure where the polypeptide chain is nearly fully extended, and the H bonding occurs between amino acids further apart in the sequence

between different segments of the chain[13–15]. Beta sheet can exist in two different forms: parallel (Figure 1.5), where all the strands have the same direction and antiparallel (Figure 1.4), where they run in opposite[13–15]. Amino acids between adjacent strands are 3.5 Å apart whereas the distance between adjacent amino acids in helices is 1.5 Å[13]. β-bridge is the same type of structure but consisting only of two hydrogen bonded residues located in different strands[16]. The amino acids prone to form beta strands are valine, isoleucine, tyrosine, phenylalanine, threonine and tryptophan[17].

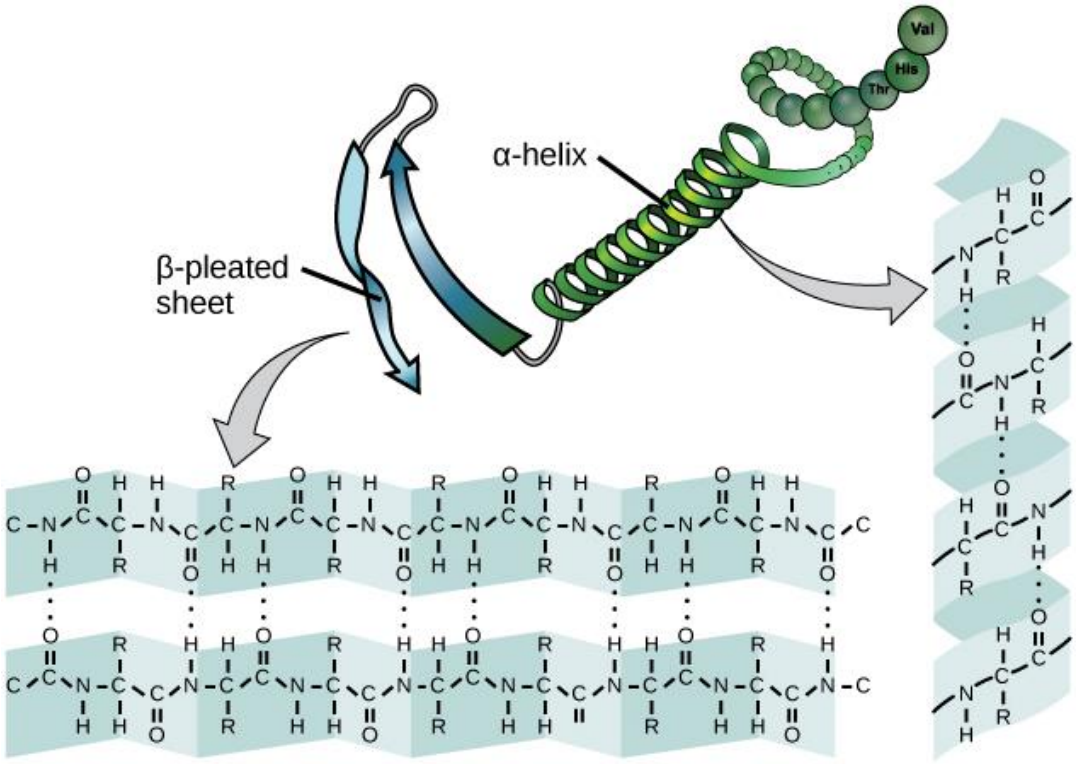

Figure 1.4. α-helix and antiparallel β-sheet motifs. Image by M.A Clark, M. Douglas, J. Choi. OpenStax Biology 2nd Edition.

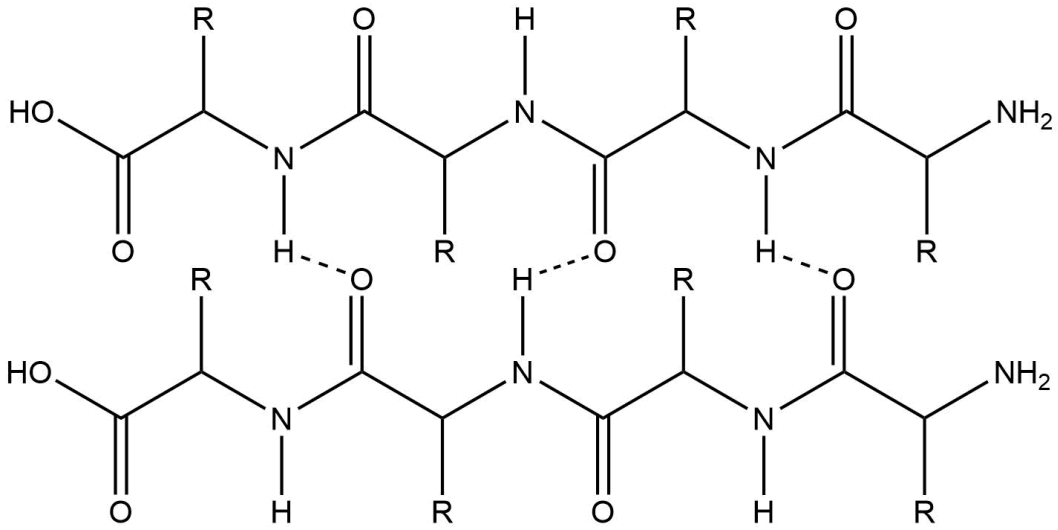

Figure 1.5. Parallel β-sheet structure.

Turns (Figure 1.6) are changes in the direction of the polypeptide chain that often connect antiparallel β-sheet[13,15]. The structure of a turn consists of n = 3, 4, 5 amino acids with two hydrogen bonds between the *i* and the *i* + *n* residue being n = 4 the most common[13,16]. Turns usually are made up of glycine because of its small size and proline due to its small Ramachandran angles[15,18].

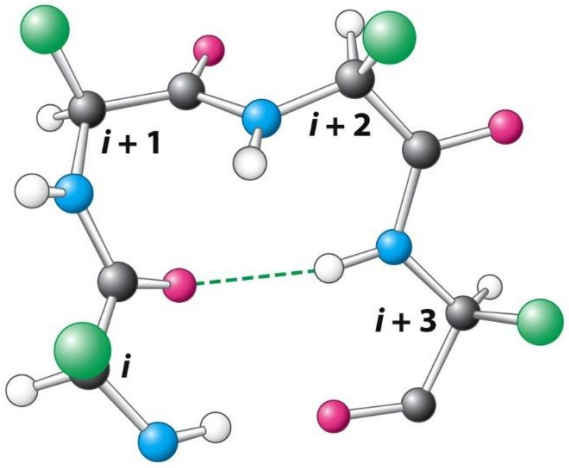

Figure 1.6. β-turn structure. Image by J.M Berg, J.L Tymoczko, L. Stryer. Biochemistry. 7th Ed, New York, W.H Freeman and company, 2012.

Bends are curvatures in the polypeptide chain[16]. For an amino acid *i*, the curvature is expressed by the relative angle of the two contiguous amino acids on the left and right (*i* - 2 and *i* + 2) which needs to be bigger than 70° to be considered as such[16].

Coil, also known as irregular, random, disordered and unordered in the literature[19–22], are structures with no define pattern, instead the polypeptide chain traces out irregular shapes in space that cannot be fit in any of the previous categories[15]. In this text, it will be preferentially -but not only- referred to as irregular. The sequences yielding this type of structure are usually richer in proline, glycine, aspartic acid, serine and aparagine[17].

Some literature considers one more class of structure in between secondary and tertiary called super-secondary structure which consist of recurrent structural patterns combining helices and sheets: α, β, β-α-β and β-α barrel[9]. Although the secondary structure elements can have the same order in the sequence, the overall 3D structure will always be unique because the way they are wired up is different from protein to protein[9]. The interactions between helices and sheets are based on Van der Walls forces, coulombic and hydrogen bonding[9].

In this work, we use for the SS of the protein reference sets the annotations found in http://2struc.cryst.bbk.ac.uk, which are based on the DSSP definitions described above, and group them into the following major classes:

Helix (G+H+I) =: 3_{10}-helix + α-helix + π-helix

Sheet (E+B) = β-sheet + β-bridge

Turn (T)

Bend (S)

Coil (C)

1.2.1.3 Tertiary and quaternary structure

The tertiary structure refers to the overall 3D structure (globe) of a protein. It is based on Van der Walls, coulombic, hydrogen bonding and hydrophobic interactions, and S-S bridges between side chains of groups very far apart in the polypeptide chain that become close in space when the protein folds[14,15]. These globular structures can further interact through S-S bridges, hydrogen bonding and Van der Walls, coulombic and hydrophobic forces to yield oligomers of identical or different polypeptide chains or subunits[14,15].

1.3 CHARACTERIZATION OF PROTEINS

1.3.1 Primary structure determination

The first thing to determine in a protein is its sequence of amino acids. There are different techniques to do so such as Edman sequencing, where the residues are removed one by one from the peptide chain and analysed by chromatography for their identification or DNA sequencing, where the genes that codify for the protein are cloned and sequence, what allows to infer the sequence of the protein by reading the nucleotide sequence[13]. Another way to determine the sequence, and probably the most popular one, is by mass spectrometry. Mass spectrometry (MS) is a technique based on fragmentation where the analytes are fragmented by different means (electro impact or assisted laser desorption/ionization) and passed through mass analysers that separate them in terms of their mass to charge [9,13]. One of the most well-established sequencing techniques in the field of proteomics is mapping + MS/MS[23]. The protein is digested with proteases that cleave the protein in specific sites and then passed through a column that separates the different fragments[23]. Upon elution from the column, they are introduced one by one into the Mass Spectrometer where they are firstly passed through a mass analyser (quadrupole filter) to further isolate and purify the ion of interest and later accelerated into a collision dissociation chamber (CID) to break up the ion into multiple fragments - ideally all possible fragments with one residue difference- that are sent into a second mass analyser (orbitrap or time of flight)[9,23,24]. The sequence is determined by computing the differences in mass between the peaks in the mass spectrogram[9,25].

1.3.2 Secondary and tertiary structure determination

Although many bioinformatic methods and neural network algorithms exist to model higher order structures from amino acid sequences, prediction of secondary and tertiary structure remains a challenge[9]. That along with the fact that proteins can unfold under certain unfavourable conditions, creates a need for techniques capable of providing conformational “pictures”. In this section, a set of spectroscopic techniques and their fundamentals and applications on protein characterization will be discussed.

Although Nuclear Magnetic Resonance spectroscopy (NMR) and X-Ray crystallography provide accurate structural information at the secondary, tertiary and quaternary levels, some limitations exist on the use of these techniques that make them no feasible for the characterization of aqueous globular proteins in routine quality control analysis[9,26,27]. The 2D varieties of NMR (COSY, NOESY and TOCSY) can predict secondary structures and their position in the sequence but they are traditionally limited to between 25-35 KDa[28,29]. Crystallography is based on diffraction of X-Ray by crystal structures but requires the samples to be in crystalline form and thus makes it unsuitable for aqueous samples[26,27].

1.3.2.1 Ultra-violet absorption spectroscopy

Ultra-violet (UV) absorption is the field in spectroscopy that studies electronic transitions when matter absorbs electromagnetic radiation with frequencies in the UV region[30]. The frequency and intensity of this transitions provide structural information on chemical groups but in the case of proteins, it is mostly used for protein concentration determination. Aromatic side chains absorb between 260-280 nm (phenylalanine ~260 nm, tyrosine ~275 nm and tryptophan ~280 nm)[31,32] and their extinction coefficients (ξ) can be found in databases (e.g., Uniprot) for λ=280 nm[33]. This means that, by introducing the pathlength and absorbance too in Equation 1.1 (Beer-Lambert law), it is possible to work out the concentration of a protein in solution[34,35]. Although the backbone transitions (180-240 nm) could potentially be used to determine protein concentration, all the extinction coefficients available in the aforementioned database are based on the aromatic side chains.

$$I = I_o \cdot 10^{-A} = I_o \cdot 10^{-\xi \cdot l \cdot C}$$
$$A = -\log \frac{I}{I_o} = \xi \cdot C \cdot l \qquad (1.1)$$

where I and I_o mean intensity (energy flux in $J.cm^{-2}.s^{-1}$) hitting the detector in presence and absence of absorbent respectively, ξ means molar extinction coefficient in $(M.cm)^{-1}$, l means optical pathlength in cm and C means concentration in M.

1.3.2.2 Circular dichroism spectroscopy

Circular dichroism (CD) is a spectroscopic technique based on absorption of radiation in the UV region but unlike UV spectroscopy, is uses circular polarized light for excitation[8,36]. The two polarizations right and left are absorbed to different extents by chiroptical samples and the difference between them is the CD signal, conventionally reported as ellipticity in mdeg[8,36]. The spectra need to be converted to mean residue weight differential absorptivity to make it comparable for the extraction of the SS components (Equation 1.2)[8]

$$\Delta A = A_L - A_R = \Delta\xi \cdot l \cdot C$$
$$\Delta\xi = \frac{0.1 \cdot \theta \cdot MRW}{l \cdot Conc \cdot 3298} \quad (1.2)$$

where ΔA is the difference in absorption between both polarizations, $\Delta\xi$ is the difference between the molar extinction coefficients of both polarizations, *C* is the concentration in *M*, *MRW=MW/(n-1)* in *Da.residue*$^{-1}$ is the mean residue weight, *l* the optical pathlength in *cm*, *Conc* the concentration in *mg.ml*$^{-1}$ and Θ is the ellipticity in *mdeg*.

A CD spectrum show characteristic spectral signatures for the different SS contents and it is possible to extract their fractions from any superposition of them by expressing the spectra as a linear combination of a reference set of proteins with known SS[37]. There are two relevant spectral regions: *far UV-CD*, from 180 to 240 nm where the peptide transitions take place and, *near UV*, between 250 and 300 nm, where the transitions from the aromatic side chains occur[36]. The alpha helix is the major secondary structure in most of proteins. Its spectrum consists of two negative minima at ~ 207 and ~ 222 nm corresponding to π-π* and n-π* transitions respectively, and a positive maximum at ~ 190 nm that corresponds to a π-π* transition orthogonal to the 207 nm one, being both counterparts of an exciton splitting of the π-π* [36,37].

β-sheet protein spectra are more variable in relative and absolute intensity than that of α-helix[37]. The two beta sheet forms parallel and antiparallel are slightly different in position and intensity. They show a positive band at ~ 195 nm and a smaller negative

peak at ~ 217 nm that corresponds to the π-π* and n-π* transitions respectively[37]. Because of the structural variability associated with it, the level of accuracy in the characterization of β-sheet is not as high as for α-helix and it is also affected by the presence of random coil whose spectral features happen to have their maxima at similar wavelengths but opposite sign which could result in wrong estimations[36,37].

The main two types of β-turn (I and II) show a similar signature to that of beta sheet but red shifted by 5-10 nm, weak negative band at ~ 225 nm and a strong positive band at ~ 200-205 nm[37].

Irregular structures often show a weak positive band at ~ 217 nm and an intense negative one at ~ 197 nm corresponding to n-π* and π-π* transitions respectively, which has a strong resemblance with that of polyprolineII[37].

Near UV-CD is related to the aromatic side chains (phenylalanine, tyrosine and tryptophan) and it can be used as a fingerprint for identification of proteins and probe for conformational changes due to its sensitivity to tertiary structure[37].

The SS predictions are carried out by means of a variety of methods, e.g., ridge regression, Singular Value Decomposition (SVD), Neural Network (NN), Partial Least Squares (PLS), that pursue expressing the CD spectrum as a linear combination of a reference set of proteins of known SS (X-Ray annotations) where the SS contents can be inferred from the weights of the sum[37,38].

1.3.2.3 Fluorescence spectroscopy

Fluorescence is a spectroscopic technique where a chromophore is excited with light of wavelength λ_{ex} and emits light of a longer wavelength λ_{em} that is recorded for structural studies[39,40]. Not all chromophores display fluorescence, but it is mostly typical of aromatic or highly conjugated molecules[41]. In proteins, fluorescence is due to tyrosine, phenylalanine and tryptophan side chains mainly[39,42]. As far as characterization is concerned, the emission maximum and intensity of these residues are sensitive to the polarity of the environment and energy transfer phenomena, and can be used to obtain information on their location and changes on the overall conformation of the protein[39,43].

1.3.2.4 IR spectroscopy

Infrared (IR) spectroscopy is a technique where the absorbance of radiation in the Infrared region by a sample is measured[41,44]. IR absorption occurs when the frequency of an electromagnetic wave is in resonance with that of the vibrational motion of different groups in the molecule and thus an IR spectrum contains structural information that can be used for qualitative analysis[41,44]. In the Mid IR region (~4000-400 cm^{-1}) it is possible to distinguish bands at characteristic wavelengths that correspond to chemical groups that hence can be identified by looking at the intensity and wavelength on the spectra[45]. In proteins, abundant literature exists on SS determination by band fitting (Appendices-B.2) of the amide C=O stretch band (amide I)[6,46]. This band is in the region from 1600 to 1700 cm^{-1} and consists of overlapped bands whose positions have been assigned to different SS: 1620-1640 cm^{-1} β-sheet, 1650-1656 cm^{-1} α-helix, 1670-1685 cm^{-1} turns and 1640-1650 cm^{-1} other[6,46].

A huge drawback that IR presents is that H_2O absorbs in the amide I region, which makes the analysis of the spectra tedious or impossible in some cases[47]. Alternatively, H_2O can be replaced by D_2O, but since D_2O exchanges H for D, it is not advisable for the study of proteins[19]. Even though water can be subtracted successfully, the optical density due to the water absorption is so high that no spectra can be recorded without reducing significantly the pathlength[48]. Attenuated Total Reflectance (ATR) is an IR mode based on evanescent waves[47]. These waves penetrate only few microns into the sample which reduces the pathlength by several orders of magnitude compared to transmission, allowing for the measurement of aqueous samples[47]. Another big inconvenience in IR is the interference of water vapour in the region of the amide I which can be supressed to certain extent by purging the instrument with N_2 and also be subtracted[47].

1.3.2.5 Raman spectroscopy

Raman spectroscopy is a technique based on scattering of radiation when light passes through a material[49]. The scattered radiation might be of the same frequency as the excitation beam (Raleigh scattering) or red shifted (Raman scattering)[49]. In the latter, these differences in wavelength correspond to transitions between vibrational levels thus the spectra will be in the IR region[49]. Although it shares some peaks with

IR, some vibrations present in IR are not in Raman and vice-verse (mutual exclusion principle)[49].

Raman spectroscopy can be used to determine SS and to monitor changes in the primary structure of proteins (PTM)[7,47]. As in IR, the amide I band (1620–1720 cm^{-1}) is a contribution of several peaks that represent different conformations that can be band fitted to give an estimation on the fractions of the different SS contents[7]. Although the amide III band (1230–1300 cm^{-1}) is also related to the SS, it is usually overlapped with bands from side-chain groups which makes the analysis difficult[7]. Because of this, the amide I is the preferred band for SS analysis. The most recurrent positions for the constituents of the amide I band are ~1656, ~1672 and ~1668 cm^{-1}, assigned to α-helix, β-sheet and irregular respectively[7,52].

Because water scatters weakly in the region of the amide I band, Raman spectroscopy allows for the analysis of proteins in solution with small data manipulation compared to IR[7].

The most important disadvantage of Raman Spectroscopy is the background fluorescence[53]. Fluorescence causes curvatures in the baseline and degradation of the signal to noise ratio which makes harder the spectra interpretation[53]. Different methods exist to supress it (differential Raman, time resolved Raman, etc) but the most common one is photodegradation or photobleaching[54,55]. Also, there are some postprocessing techniques available to correct the curvature in the baseline based on fitting and subtraction of the baseline[56].

On a separate note, the dependence of the Raman intensity on the concentration and pathlength is given by an equation analogous to Beer-Lambert law (Equation 1.3)[57]

$$I = I_o \cdot e^{-N \cdot \sigma \cdot z} \tag{1.3}$$

where σ is the Raman scattering cross section in $cm^2/molecule$, N is the concentration of the sample in $molecules/cm^3$ and z is the pathlength in cm. Since in Raman spectroscopy the intensity is small, Equation 1.3 can be approximated by the linear term of its Taylor expansion

$$I = I_o(1 + N \cdot \sigma \cdot z)$$

$$I \propto I_o \cdot N \cdot \sigma \cdot z \tag{1.4}$$

Furthermore, if the beam is focused on the sample the signal becomes independent from the pathlength and Equation 1.4 can be written as

$$I \propto I_o \cdot N \cdot \sigma \tag{1.5}$$

1.3.2.6 Raman Optical Activity spectroscopy

Raman Optical Activity (ROA) is the equivalent to CD in scattering[58]. It measures the difference between scattered right and left circular polarized light (Equation 1.6) which as in CD, it is of a very small magnitude, so more powerful lasers, time exposures and larger number of accumulations need to be used[58]. Most of the times and despite all the efforts listed before, the spectra need to be smoothed. Detailed SS assignments have been established for both amide I and III[59,60]. Amide I consist of a couplet with a negative peak about 1635 cm^{-1} and a positive one about 1655 cm^{-1} for both helical and sheet proteins[59,60]. In the amide III region, helical proteins show three distinctive peaks at ~1300, 1342 and 1270 cm^{-1} with positive, positive and negative sign respectively[59,60]. β-sheet shows a characteristic negative band at ~1245 cm^{-1} [59,60]. Although ROA is used to study SS and conformational changes, no quantitative post- processing approaches seem to exist to date for estimation of SS.

$$ROA = (I_R - I_L) \tag{1.6}$$

1.4 PRODUCTION OF PROTEINS IN PHARMACEUTICAL INDUSTRY

In pharmaceutical industry, proteins are produced by means of biological expression systems (mammalian cell lines), among which Chinese Hamster Cells (CHO) are the most popular ones[61]. These expression systems are based on recombinant DNA and cell culture technology which consists of incorporating (transfection)[12] fragments of DNA (genes) that codify for a protein of interest into the DNA of another organism (host) for it to express it as its own[61,62]. CHO are the most widely used mammalian cells because of their ability to grow in serum-free cultures and because of their

ability to resist human viral infections[12]. The development of a stable and high productive cell lines at large scale remains a big challenge in biopharmaceutical industry[63]. A previous transient expression step is required to assess the feasibility of the recombinant proteins in terms of manufacturability and clinical efficacy[12]. Once the protein is validated, the recombinant DNA needs to be transferred into the cell and amplified[12]. Because of random insertion into the plasmid, clonal variations occur and, since every clone reacts differently to different culture media, it is necessary to select those that are suitable for the large scale production[12]. This is achieved by limiting dilution, a clone screening method in which the parental clones are grown separately[12]. Finally, the selected clone is put into a culture suspension for large scale protein expression in bioreactors[12]. After their production, proteins need to be separated and purified from the components in the culture suspension which consist of host cells, DNA, other proteins and a variety of small molecules[64]. The first step in the purification is the clarification by centrifuging or filtration to remove all those thick components that are incompatible with chromatography[15,65]. If the protein is not secreted by the host cell, the protein needs to be extracted from the cell by detergent lysis, sonication or salting out before[66,67]. Once the clarification has been performed, the mixture is put through affinity chromatography and ion exchange for further purification[64,65]. Finally, structural analyses are performed to assess the quality of the protein by mass spectrometry to determine the intact mass, sequence and posttranslational modifications[12].

2 ATTENUATED TOTAL INTERNAL REFLECTION AND ANOMALOUS DISPERSION

2.1 INTRODUCTION

Infra-red (IR) spectroscopy is a technique used for molecular structure determination. IR spectra of proteins are known in the literature to provide information on secondary structure content[6,20,22]. In this chapter, a neural network approach called Self-Organizing-Maps (SOM), previously developed for circular dichroism spectroscopy[2], was implemented for secondary structure (SS) determination of proteins. SOM requires a reference set of protein spectra of well-known structure in terms of which the spectra of test proteins can be expressed. Attenuated Total Reflectance (ATR) is a technique based on evanescent waves and the effective depth of penetration of these waves -the distance the wave reaches above the crystal- changes as a function of the wavenumber (Figure 2.1). This results in relative band intensity changing across the spectrum and also position shifts that make ATR spectra no longer comparable to their transmission analogous[68,69]. Because of this, it was necessary to convert ATR spectra into transmission spectra prior to putting them through SOM.

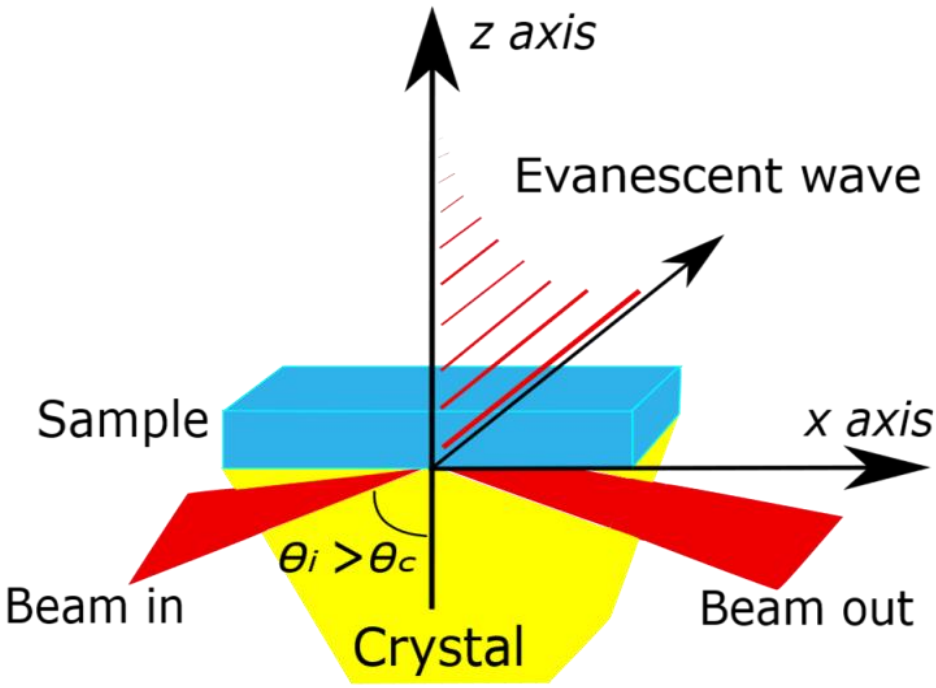

Figure 2.1. Diagram of an ATR experiment.

The conversion mentioned above required a deep look into the phenomenon called "anomalous dispersion". We found in the literature numerous references to the book of the pioneer of the technique Harrick (Internal Reflection Spectroscopy)[70] with multiple equations that, just like with the book itself, we failed to understand as a whole due to the gaps in the presented derivations. Because of that, we decided to derive our own equations following the principles of electromagnetic waves in dielectrics and dispersion theory.

This chapter intends to review the theory of ATR and anomalous dispersion; and establish a semiempirical method to calculate the refractive index of an absorbent for the sake of the conversion between ATR and transmission.

2.2 MATERIALS AND METHODS

2.2.1 Theoretical background review

2.2.1.1 Evanescent wave

When light travels between two media of different refractive index, the beam changes direction according to Snell's law (Equation 2.3). When the angle of incidence is above the critical one, no light is transmitted anymore but reflected in its totality in a phenom known as "total reflection". By examining Maxwell's equations at boundaries for dielectrics, it can be seen that the electric and magnetic fields of an electromagnetic wave penetrate the second medium but fades away as one moves away from the interphase. This is called "evanescent wave", a standing wave with an electric field amplitude that decays exponentially with distance (Figures 2.2-2.3) resulting in a penetration (effective pathlength) of about a micron although it depends on the refractive index of the crystal used. When an absorbent is placed on top of the ATR crystal (Figure 2.1), the evanescent wave interacts with it in what the community refers to as "attenuated total reflection" yielding spectra with intensities of several orders of magnitude less than those of conventional transmission[68,71,72].

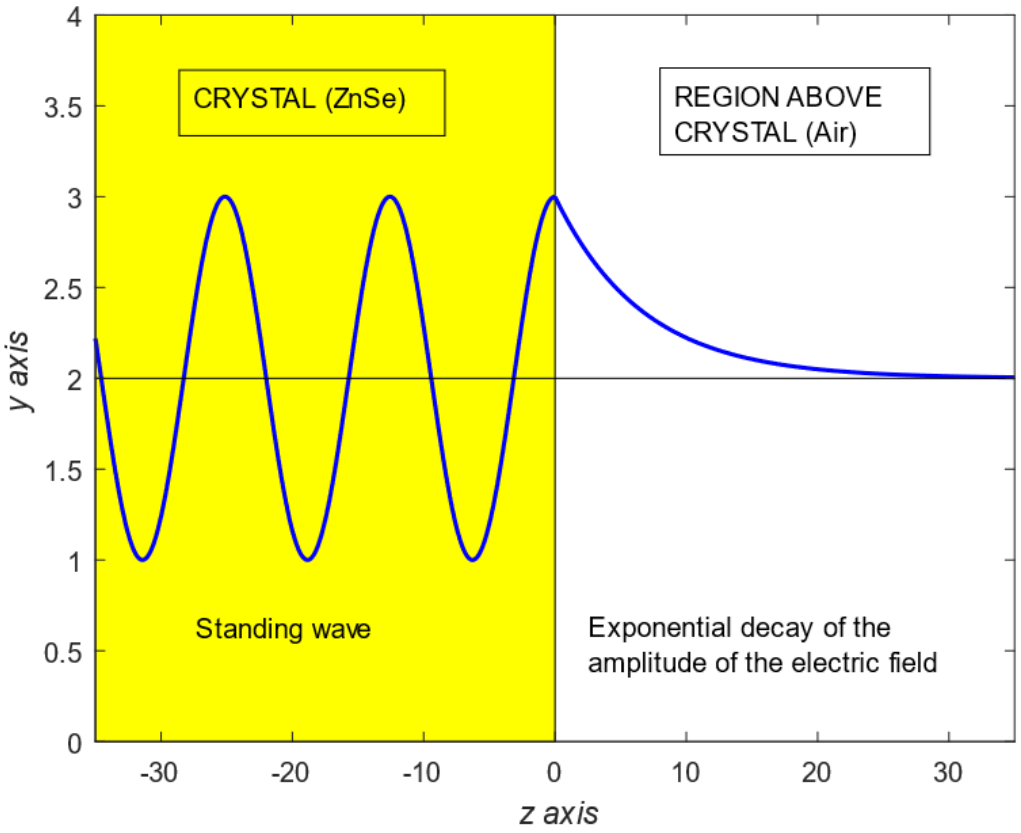

Figure 2.2. Side view of an evanescent wave below and above the boundary.

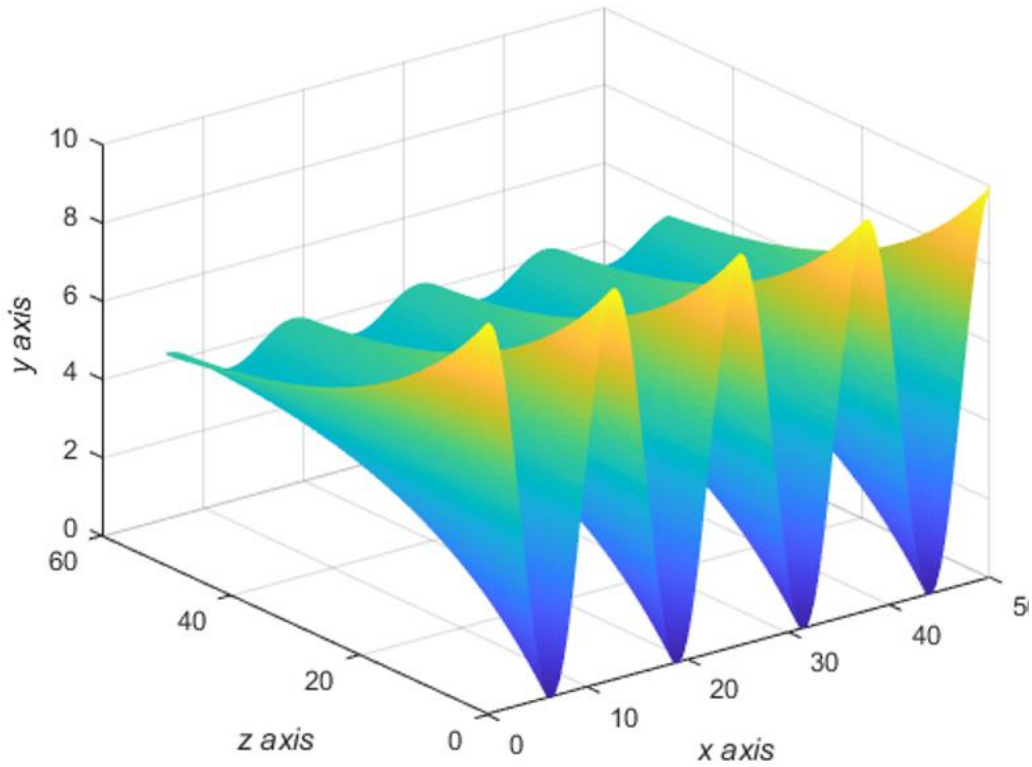

Figure 2.3. Evanescent wave propagating in the *x* direction and decaying exponentially above the surface of the crystal (*z* axis).

When light moves from one medium to another its frequency ν remains the same but its speed c and hence wavelength $\lambda_\circ$ change[72]

$$\nu = \frac{c}{\lambda_\circ} = \frac{v_1}{\lambda_1} = \frac{v_2}{\lambda_2} \tag{2.1}$$

$$\overrightarrow{k_2} = \vec{k}\frac{c}{v_2} = \vec{k} \cdot n_2 \tag{2.2}$$

where $\vec{k} = \frac{2\pi}{\overrightarrow{\lambda_\circ}}$ is the wavenumber in vacuum, n means refractive index (ratio of speed of light in a vacuum to that in the considered medium) and the sub-indexes *1* and *2* represent the crystal and sample respectively.

Consider an incident beam in a dense medium (crystal) of refractive index n_1 propagating towards a flat surface (Figure 2.1) at angle θ_1 to the normal of the surface then Snell's law for its refraction by the surface is

$$n_1 \cdot sin\,\theta_1 = n_2 \cdot sin\,\theta_2 \tag{2.3}$$

where the subscript *2* refers to the sample above the surface. The critical angle θ_c, is the θ_1 for which $\theta_2 = 90°$. From Equation 2.3, the critical angle is[73]

$$\frac{n_1}{n_2} = \frac{1}{sin\,\theta\,c} = \frac{sin\,\theta_2}{sin\,\theta_1} \tag{2.4}$$

As light varies harmonically with frequency$\omega = 2\pi \cdot \nu$

$$\vec{E} = Re[\,E_\circ \cdot e^{i(\vec{k}\cdot\vec{r} - w\cdot t)}] \cdot \overrightarrow{\varphi_\circ} \tag{2.5}$$

where $\vec{\boldsymbol{r}} = (\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ is the position vector in the space, $\vec{k}$ is the wavevector, *w* is the angular frequency, *t* is the time, $\overrightarrow{\varphi_o}$ is the unitary vector that represents the direction of the electric field and $E_\circ$ is the magnitude of the electric field on the surface. If we

consider now the wave above the crystal, and call *z* to the direction up from the ATR surface and consider *x* the direction of propagation

$$\overrightarrow{k_2} \cdot \vec{r} = x \cdot k_{2x} \cdot sin\,\theta_2 + z \cdot k_{2z} \cdot cos\,\theta_2$$

(2.6)

It can be seen in Equation 2.6 that above the critical angle $sin^2\,\theta_2 > 1$ the second term goes imaginary and so it is more convenient to use

$$\cos\theta_2 = \sqrt{(1-\sin^2\theta_2)} = i\sqrt{(\sin^2\theta_2 - 1)}$$

(2.7)

where *i* means imaginary. Combining Equations 2.5, 2.6 and 2.7 the expression for the magnitude of the electric field of an evanescent wave is obtained

$$E = Re\left[E_\circ \cdot e^{i(\vec{x}\cdot\vec{k}_2\cdot sin\,\theta_2 - w\cdot t)} e^{-\vec{z}\cdot\vec{k}_2\sqrt{\sin^2\theta_2 - 1}}\right] = Re\left[E_o \cdot e^{i(\vec{x}\cdot\vec{k}_2\cdot sin\,\theta_2 - w\cdot t)} e^{-\vec{z}\cdot\vec{\gamma}}\right]$$

(2.8)

where

$$\vec{\gamma} = \vec{k}_2\sqrt{\sin^2\theta_2 - 1}$$

(2.9)

That means there is no transmitted wave along the *z* axis anymore but just the so-called evanescent wave which propagates along the *x* axis, decays exponentially from the surface and has electric field intensity components in the *x*, *y* and *z* directions (Figure 2.3)[74].

2.2.1.2 Depth of penetration

The amount of energy that interacts with the sample depends on the distance the wave penetrates into the sample. This distance is known as depth of penetration and is arbitrarily defined by Harrick[70] as the distance from the ATR crystal surface required for the amplitude of the electric field to drop to *1/e* of its value at the surface[68,71,74].

$$e^{-\gamma \cdot dp} = \frac{1}{e}$$

(2.10)

Combining Equation 2.9 with Equation 2.10 and expressing the wavevector in terms of the refractive index and the wavenumber in vacuum

$$dp = \frac{1}{\gamma} = \frac{1}{k_2 \cdot \sqrt{sin^2\,\theta_2 - 1}} = \frac{\lambda_o}{2\pi \cdot n_1 \sqrt{sin^2\,\theta_1 - \left(\frac{n_2}{n_1}\right)^2}}$$

(2.11)

where $\lambda_\circ$ is the wavelength of light of a given frequency in vacuum. This means that, for example, for $\theta_1 = 45^\circ$ with a ZnSe ATR-crystal, $n_1 = 2.4$ and $n_2 \simeq 1.334$ (refractive index of water at 25°C and infinite wavenumber when there is no absorbance), $d_p \simeq 1\mu m$.

When the medium absorbs, Equation (2.11) still applies but the sample refractive index can no longer be approximated by its value at infinite wavenumber, but its dispersive nature must be considered (see next section).

2.2.1.3 Anomalous dispersion

The speed of light in vacuum *c* is constant and the maximum predicted by the relativity theory for any electromagnetic wave. When light enters a dielectric material, the electric field induces a local one as it passes through which in turn interacts with the first causing the delay of the wave and resulting in a decrease of the propagation speed. This is expressed in terms of the refractive index. Moreover, for most of known materials, the refractive index is not independent of the incident wavelength, but it changes with it in a phenomenon known as normal dispersion. The most important fact about dispersion is that the refractive index decreases with wavelength with approximately $1/\lambda^2$ fashion or, in a more general way, $\partial n/\partial \lambda < 0$[72]. However, when there is absorption, the refractive index shows in the region of the oscillator natural frequency (frequency of maximum absorption) a fashion that resembles a positive and negative asymptote at lower and higher wavenumbers respectively with an abrupt change in magnitude close to the natural frequency in which the refractive index changes positively with wavelength ($\partial n/\partial \lambda > 0$) (Figure

2.4). This phenomenon is known as anomalous dispersion and it is responsible for changes in shape, intensity and position of the bands of ATR spectra compared to those of transmission.

The effects of anomalous dispersion can be explained by having a quick look at the depth of penetration (Equation 2.11) which can be regarded as the equivalent pathlength of the transmission experiment consider that the absorbance is proportional to it in accordance with Beer-Lambert law. On the left side of the band in Figure 2.4, the refractive index grows as one approaches the frequency of maximum absorbance which translates in a higher critical angle (Equation 2.4) and so larger depth of penetration. On the right side, the refractive index goes smaller as one gets closer to the natural frequency of the vibration making the depth of penetration smaller too. This causes the bands to shift left and change in shape. The other distinctive characteristic of ATR spectra is due to the linear dependence of the depth of penetration on the wavelength which makes the absorbance drop as the wavenumber increases resulting in less intense bands (Figure 2.5).

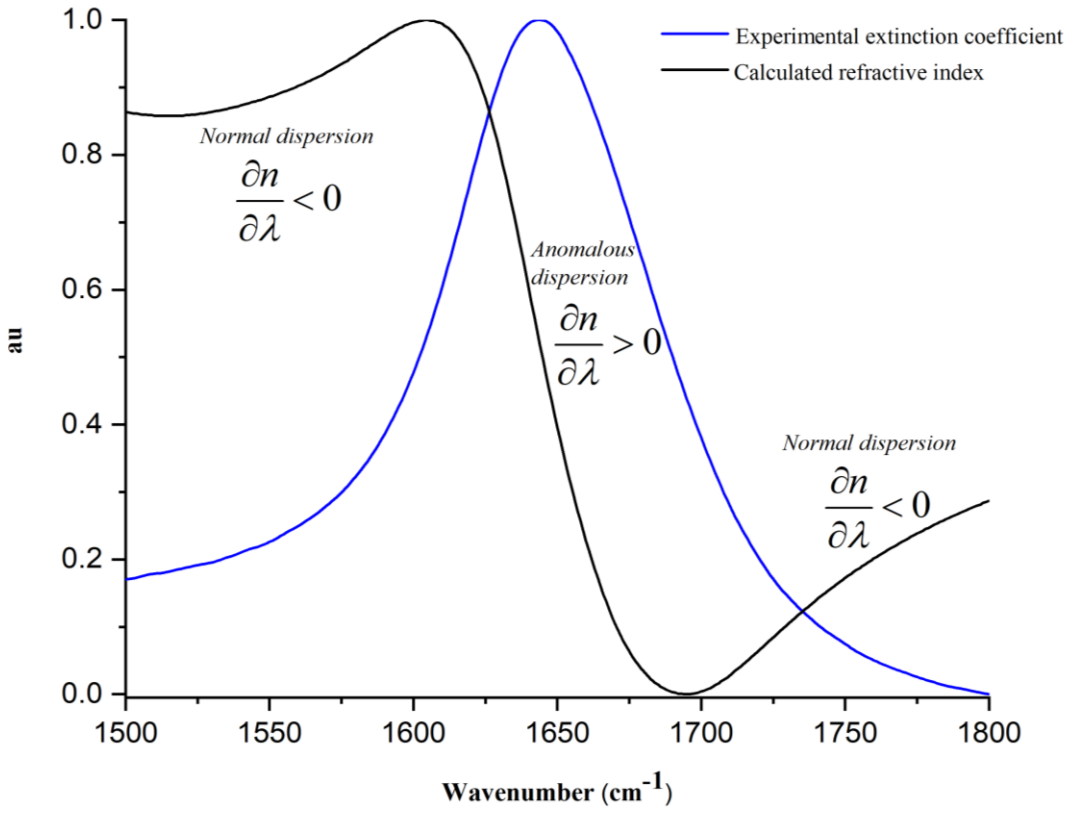

Figure 2.4. Calculated refractive index of water and its correspondent extinction coefficient within the region of the OH deformation mode. Both the refractive index and the absorbance were normalized by the interval method.

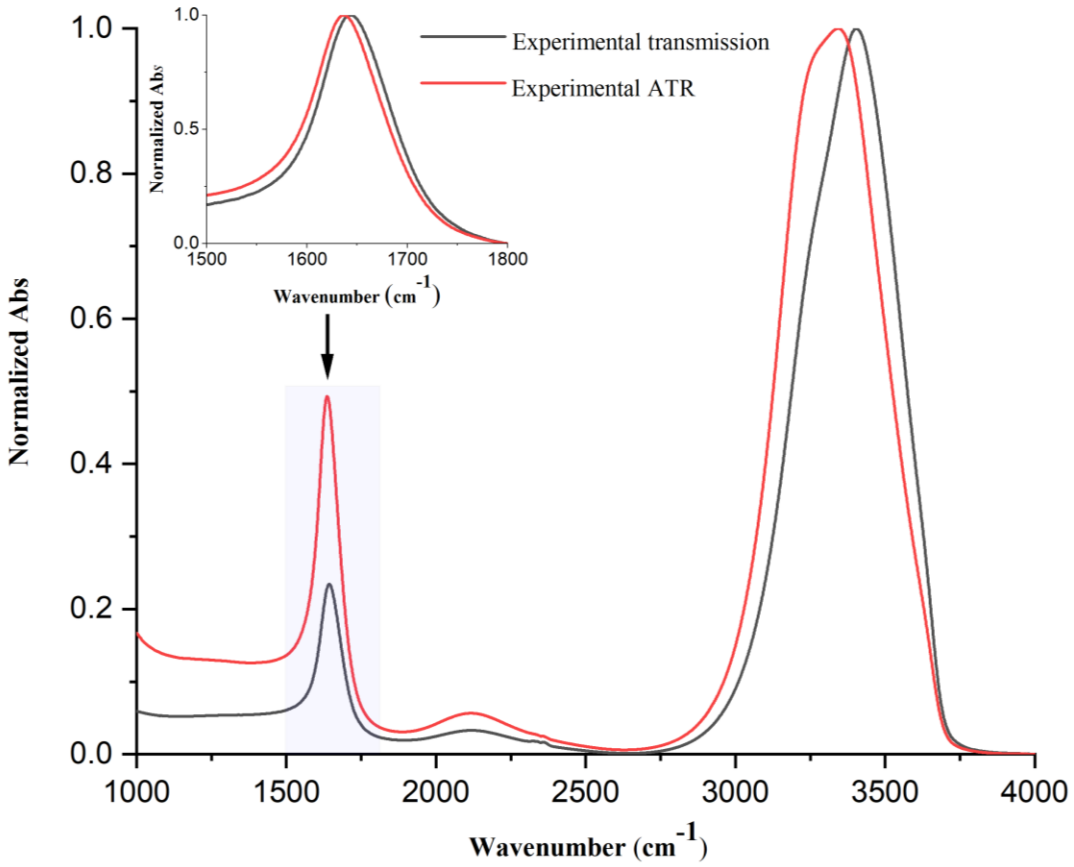

Figure 2.5. Experimental transmission and ATR spectra of water. Both the inset figure and main one, were normalized by the interval method.

2.2.1.4 Refractive index as a function of wavenumber

Beer-Lambert law states that the flux of energy penetrating an absorbing medium reduces with distance in an exponential fashion (Equation 1.1). In this section the complex nature of the refractive index will be proved by means of the classical damped harmonic oscillator model[75].

Ampere-Lenz and Faraday's laws in a dielectric read as follows[75]

$$\vec{\nabla} \wedge \vec{E} = -\frac{\partial \vec{B}}{\partial t} = -\mu_{\circ}\left[\frac{\partial \vec{H}}{\partial t} + \frac{\partial \vec{M}}{\partial t}\right] \tag{2.12}$$

where $\vec{E}$ is the electric field, $\vec{B}$ is the induction magnetic field, $\vec{H}$ is the magnetic field, $\vec{M}$ the magnetization, t means time and μ_o is the magnetic permittivity in vacuum

$$\vec{\nabla} \wedge \vec{H} = -\frac{\partial \vec{D}}{\partial t} + \vec{J} = -\left[\varepsilon_{\circ} \frac{\partial \vec{E}}{\partial t} + \frac{\partial \vec{P}}{\partial t}\right] + \vec{J}$$

(2.13)

and $\vec{D}$ is the induction electric field or displacement, $\vec{P}$ is the polarization, $\vec{J}$represents electric currents and ε_o is the electric permittivity in vacuum. In a non-conducting material $\vec{J} \sim 0$ and so Equation 2.13 can be rewritten as

$$\vec{\nabla} \wedge \vec{H} = -\frac{\partial \vec{D}}{\partial t} = -\left[\varepsilon_{\circ} \frac{\partial \vec{E}}{\partial t} + \frac{\partial \vec{P}}{\partial t}\right]$$

(2.14)

In a nonmagnetic material Equation 2.12 gets simplified to

$$\vec{\nabla} \wedge \vec{E} = -\mu_{\circ} \frac{\partial \vec{H}}{\partial t}$$

(2.15)

By taking the curl of Equation 2.15, and using Equation 2.14 and the identity $\vec{\nabla} \wedge (\vec{\nabla} \wedge \vec{E}) = -\nabla^2 \vec{E} + \vec{\nabla}(\vec{\nabla} \cdot \vec{E})$ (where $\vec{\nabla} \cdot \vec{E} = 0$ for materials with no net charge) we obtain

$$-\frac{\partial^2 \vec{E}}{\partial x^2} = -\mu_{\circ}\varepsilon_{\circ} \frac{\partial^2 \vec{E}}{\partial t^2} - \mu_{\circ} \frac{\partial^2 \vec{P}}{\partial t^2}$$

(2.16)

According to propagation of light in dielectrics classic theory[75], when a system of bound electrons is exposed to an electric field, they get displaced from their equilibrium position following the "classical damped harmonic model"

$$m \cdot \frac{d^2 \vec{x}}{dt^2} + m \cdot \gamma \cdot \frac{d\vec{x}}{dt} + K \cdot \vec{x} = -_{-} q \cdot E_{\circ} \cdot e^{-i \cdot w \cdot t} \cdot \overrightarrow{\varphi_{\circ}}$$

(2.17)

where $\vec{x}$ stands for the displacement, m means mass of the electron, γ is a frictional damping factor, K is the spring model constant, E_o is the amplitude of the electric field, w is the angular frequency, t means time, q is the charge of the electron and $\overrightarrow{\varphi_\circ}$ is an unitary vector. The first term means acceleration force, the second frictional damping force, the third is Hooke's law and the fourth the force exerted by an electric field.

If the electric field oscillates harmonically and assuming the electrons move around their equilibrium positions with the same time dependence, $\vec{x} = x_\circ \cdot e^{-i \cdot w \cdot t} \cdot \vec{\varphi}_\circ$ we get

$$(-m \cdot w^2 - i \cdot w \cdot m \cdot \gamma + K) \cdot x_\circ = -_- q \cdot E_\circ \tag{2.18}$$

The polarization is defined as the electric dipole moment per unit of volume

$$\vec{P} = -\frac{\sum_{i=1}^{N} \vec{p}_i}{V} = \frac{\sum_{i=1}^{N} -q_i \cdot \vec{x}}{V} = \frac{N \cdot_- q}{V} \cdot \vec{x} \tag{2.19}$$

producing a net polarization in the material that can be expressed in terms of the electric field by rearranging Equation 2.18 for $\vec{x}$ and replacing in Equation 2.19

$$\vec{P} = \frac{N}{V} \cdot_- q \cdot \left(\frac{-q \cdot \overrightarrow{E}}{-m \cdot w^2 - i \cdot w \cdot m \cdot \gamma + K} \right) = \frac{N \cdot_- q^2}{V \cdot (-m \cdot w^2 - i \cdot w \cdot m \cdot \gamma + K)} \cdot \vec{E} \tag{2.20}$$

By substituting the polarization into Equation 2.16 we get

$$-\frac{\partial^2 \vec{E}}{\partial x^2} = -\mu_\circ \cdot \varepsilon_\circ \left[1 + \frac{N \cdot_- q^2}{V \cdot \varepsilon_\circ} \cdot \frac{1}{(-m \cdot w^2 - i \cdot w \cdot m \cdot \gamma + K)} \right] \frac{\partial^2 \vec{E}}{\partial t^2} \tag{2.21}$$

The wave equation for an electromagnetic wave in a medium is

$$-\frac{\partial^2\vec{E}}{\partial x^2} = -\mu \cdot \varepsilon \frac{\partial^2\vec{E}}{\partial t^2} = -\frac{1}{v^2}\frac{\partial^2\vec{E}}{\partial t^2} = -\left(\frac{n}{c}\right)^2 \frac{\partial^2\vec{E}}{\partial t^2} = -\mu_o \cdot \varepsilon_o \cdot n^2 \frac{\partial^2\vec{E}}{\partial t^2}$$

(2.22)

It can be seen by comparing Equation 2.21 with Equation 2.22 that the term inside the squared brackets in Equation 2.21 is the same as the refractive index of the medium. Multiplying the second term inside the squared brackets above and below by the conjugate of the expression in round brackets and expressing the force constant in terms of the natural frequency $K = w_\circ^2 \cdot m$ results in

$$\eta^2 = 1 + \frac{N \cdot_- q^2}{V \cdot \varepsilon_\circ \cdot m} \cdot \left(\frac{(w_\circ^2 - w^2)}{(w_\circ^2 - w^2)^2 + (\gamma \cdot w)^2}\right) + \frac{N \cdot_- q^2}{V \cdot \varepsilon_\circ \cdot m} \cdot \left(\frac{i \cdot \gamma \cdot w}{(w_\circ^2 - w^2)^2 + (\gamma \cdot w)^2}\right)$$

(2.23)

Equation 2.23 is a complex quantity and can be expressed as

$$\eta = n + i \cdot \kappa$$

(2.24)

where κ relates to the Eulerian extinction coefficient as follows

$$\kappa = \frac{ln\,1\,0 \cdot \xi \cdot C \cdot \lambda_\circ}{4\pi}$$

(2.25)

where ξ is the standard decadic extinction coefficient, C means concentration and λ_o is the wavelength of an electromagnetic wave in vacuum.

Since the refractive index is related to the wavenumber by the equation $n = k \cdot c/w$, the wavenumber can be then expressed as a complex number too

$$\vec{K} = \vec{k} + i \cdot \vec{\alpha}$$

(2.26)

where $\vec{\alpha}$ is the coefficient of absorption

By substituting Equation 2.26 into Equation 2.5 we get

$$\vec{E} = Re[\, E_\circ \cdot e^{i((\vec{k}+i\cdot\vec{\alpha})\vec{r}-w\cdot t)}] \cdot \overrightarrow{\varphi_\circ} = Re[\, E_\circ \cdot e^{-\vec{\alpha}\cdot\vec{r}} \cdot e^{i(\vec{k}\cdot\vec{r}-w\cdot t)}] \cdot \overrightarrow{\varphi_\circ}$$

(2.27)

which means the electric field gets attenuated with exponential fashion along the axis of propagation. This is the foundation of Beer-Lambert law (Equation 1.1).

Absorbance and refractive index at a given wavenumber are related by the Kramers - Kronig transformation[48,76,77]:

$$n_2(\tilde{v}_i) = n_{2\infty} + \frac{2}{\pi} \wp \int_0^\infty \frac{\tilde{v} \cdot \kappa(\tilde{v})}{\tilde{v}^2 - \tilde{v}_i^2} \cdot d\tilde{v} = n_{2\infty} + \frac{ln\,10}{2\pi^2} \wp \int_0^\infty \frac{\xi(\tilde{v}) \cdot C}{(\tilde{v}^2 - \tilde{v}_i^2)} \cdot d\tilde{v}$$

(2.28)

where $\tilde{v} = 1/\lambda_\circ$, $n_{2\infty}$ is the sample refractive index at infinite wavenumber and $\wp$ denotes Cauchy Principle Value (CPV). A CPV integral is a type of improper integral in which the singularity is approached at the same rate from both sides[78].

The conversion between ATR and transmission requires knowing the extinction coefficients and concentrations of both protein and water which can be determined experimentally. Because of the discontinuity of the function, it was decided to solve the integral analytically, for which it was necessary to express the spectrum as an integrable function. The experimental transmission spectrum of the protein in water was collected with a demountable transmission cell with CaF_2 windows and no spacer in between. After, the spectra were vapour subtracted, zeroed between 4400 and 4500 cm^{-1}, divided by the pathlength to express it as $A/l = \xi \cdot C$ and fitted with a linear combination of Lorentzian curves (Equation 2.29) between 1500 and 1780 cm^{-1} through a Matlab[79] routine that used as initial values those from a prior fitting of water performed in Origin[80]. The pathlength was estimated by comparing the absorbance within the region 1860 - 2650 cm^{-1} (due to water only) to that of a water spectrum of known thickness (100 μm). A more detailed description of the measurement and processing of transmission spectra will be given in chapter 3.

$$Abs = y_o + \sum_{i=1}^{n} 1/\pi \frac{2 \cdot A_i \cdot w_i}{4 \cdot \left(\tilde{\nu} - \tilde{\nu}_{o_i}\right)^2 + w_i^2}$$

(2.29)

Where A_i, w_i and $\tilde{\nu}_i$ are the intensity, the bandwidth and the centre of the i-th peak respectively and $\tilde{\nu}$ the wavenumber in cm^{-1}.

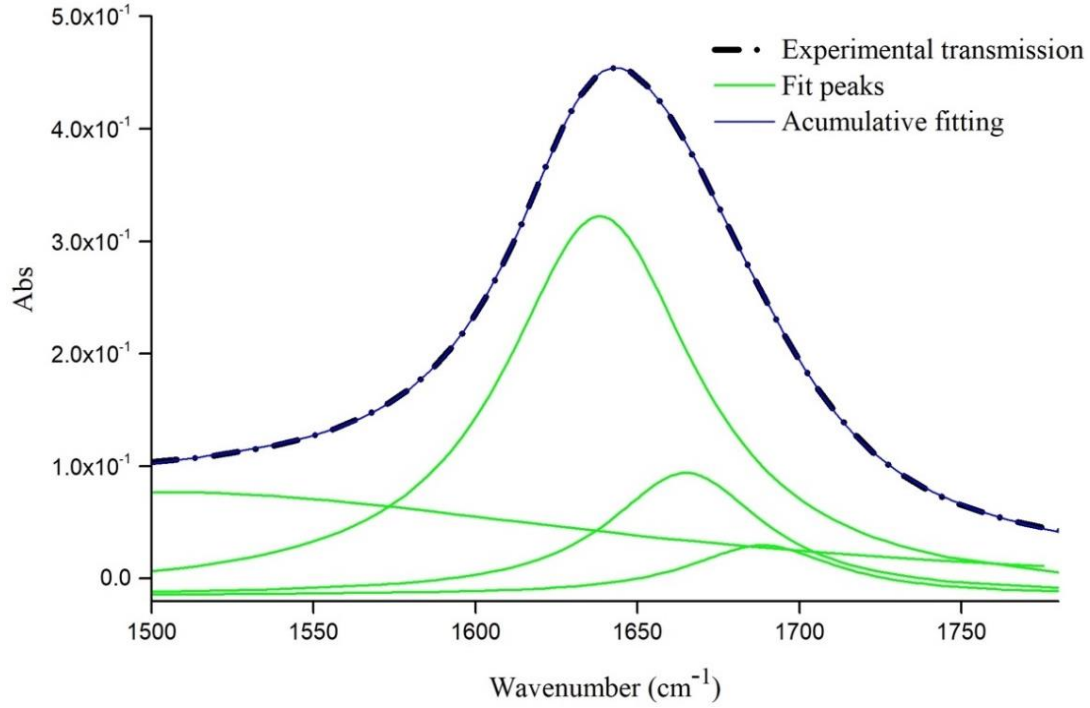

Figure 2.6. Fitting of the band corresponding to the OH deformation mode of a water transmission spectrum. The fitting was done in Origin[80] with Lorentzian curves. The initial position and number of peaks used were chosen by adding peaks one by one from the centre of gravity of the band until no significant improvement in the R^2 value could be achieved, the residuals were normally distributed across the spectrum and no visual distinction between experimental and accumulative fit could be easily made.

2.2.1.5 Evanescent wave in presence of an absorbent

In equation 2.9 -in presence of an absorbent- the refractive index and wavevector go complex (Equations 2.24 and 2.26)

$$\gamma = K_2\sqrt{sin^2\,\theta_2 - 1} = \eta_2 \cdot k\sqrt{sin^2\,\theta_1 \cdot \left(\frac{n_1}{\eta_2}\right)^2 - 1} = k\sqrt{sin^2\,\theta_1 \cdot n_1^2 - \eta_2^2}$$

$$k\sqrt{sin^2\,\theta_1 \cdot n_1^2 - (n_2 + i \cdot \kappa)^2} = k\sqrt{sin^2\,\theta_1 \cdot n_1^2 - (n_2^2 + 2n_2 \cdot i \cdot \kappa - \kappa^2)}$$
$$= k(a + i \cdot b)$$

(2.30)

By squaring the terms on both sides and equating real and imaginary parts we obtain a system whose solutions -neglecting κ^2 ($\kappa^2 \sim 0$ for very small values of absorption)- are as follows

$$n_1^2 \cdot sin^2\,\theta_1 - n_2^2 - 2n_2 \cdot i \cdot \kappa = a^2 + 2a \cdot i \cdot b - b^2$$
$$\left.\begin{matrix} a^2 - b^2 = n_1^2 \cdot sin^2\,\theta_1 - n_2^2 \\ a \cdot b = -n_2 \cdot \kappa \end{matrix}\right\}$$
$$a = \sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}$$
$$b = \frac{-n_2 \cdot \kappa}{\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}}$$

(2.31)

Combining now the solutions *a* and *b* with the complex definition of γ given in Equation 2.30 and the equation of the evanescent wave (Equation 2.8) yields

$$E = Re\left[E_\circ \cdot e^{i\left(x \cdot k_1 \cdot sin\,\theta_1 - w \cdot t + z\frac{n_2 \cdot k \cdot \kappa}{\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}}\right)} e^{-z \cdot k\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}}\right]$$

(2.32)

This is now a wave that propagates in the *x* and *z* directions with an electric field amplitude that decreases exponentially above the surface.

The speed at which the wave propagates across the boundary is given by the refractive index implicit in the exponential term of Equation 2.32

$$\frac{n_2 \cdot \kappa}{\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}} k \cdot z = n_{2'} \cdot k \cdot z = k_{2'} \cdot z$$

(2.33)

and, by means of Equation 2.11, the refractive index $n_{2'}$ can be expressed as

$$n_{2'} = n_2 \cdot k \cdot \kappa \cdot d_p$$

(2.34)

The wave can then be thought of as a transverse wave that leaks above the surface of the crystal at a rate given by the fictitious refractive index which depends on the extinction coefficient of the sample.

2.2.1.6 Ratio of Intensity of light reaching the detector

The flux of electromagnetic energy is given by the Poynting vector $\vec{S}$[57,81]

$$\vec{S} = \upsilon^2 \cdot \varepsilon \cdot \vec{E} \wedge \vec{B}$$

(2.35)

where υ is the speed of light in any considered material medium, ε is the electric permittivity of the material, $\vec{E}$ the electric field and $\vec{B}$ the induction magnetic field. By means of some vector-algebra it comes out that the intensity of the light beam can be expressed as a function of the amplitude of the electric field as follows[57,81].

$$\langle \vec{S} \rangle = \vec{v} \cdot \varepsilon \cdot |E|^2$$

(2.36)

where $\langle \vec{S} \rangle$ is the average of the Poynting vector over a cycle and its units are energy per unity of cross-section and time.

The intensity of light that reaches the detector is the incident light minus the amount of energy that leaks through the boundary. In order to connect the three of them and considering the transmitted light will have a different cross section, it is necessary to express the flux of energy in Energy per unit of time

$$I = \langle \vec{S} \rangle \cdot \vec{\mathbb{C}}$$

(2.37)

where $\vec{\mathbb{C}}$ means beam cross-section and I is Energy per unit of time. The absorbance is defined in terms of the ratio reflected intensity with and without absorption[30,57]

$$A = -\log R = -\log \frac{I^r}{I_o^r}$$

(2.38)

where I^r stands for the reflected intensity when there is absorption and I_o^r means reflected intensity when there is no absorption.

The magnitude of the reflected intensity relates to the incident intensity I^i by means of the Fresnel coefficient for reflection *r* and the magnitude of the transmitted one I^t by means of the Fresnel coefficient for transmission *t* [72].

In the absence of absorption, the amount of light reflected equals the incident (Equation 2.39).

$$I_o^r = I^i$$

(2.39)

where I^i is

$$I^i = \varepsilon_1 \cdot \upsilon_1 \cdot E_o^2 \cdot \mathbb{C}_1$$

(2.40)

and the balance of the total light is

$$I^r = I^i - I^t$$
$$R = \frac{I^r}{I_o^r} = 1 - \frac{I^t}{I^i} = 1 - \frac{\varepsilon_{2'} \cdot \upsilon_{2'} \cdot E_o^2 \cdot |t|^2 \cdot \mathbb{C}_{2'}}{\varepsilon_1 \cdot \upsilon_1 \cdot E_o^2 \cdot \mathbb{C}_1} = 1 - \frac{n_{2'}}{n_1} \frac{|t|^2}{\cos \theta_1}$$

(2.41)

where the indexes *r*, *i*, and *t* stand for reflected, incident, transmitted respectively, *2'* means the medium above the crystal with refractive index $n_{2'}$ and the ratio of the cross-sections was replaced using the following relation

$$\frac{\mathbb{C}_{2'}}{\mathbb{C}_1} = \frac{\pi \frac{d \cdot h}{4}}{\pi \left(\frac{d}{2}\right)^2} = \frac{1}{\cos \theta_1}$$

(2.42)

Using now Equation 2.34 to replace $n_{2'}$ and expressing the absorption index κ in terms of the decadic one by means of Equation 2.25 yields the equation introduced by Milosevic[74,82]

$$R = 1 - \frac{n_{2'}}{n_1}\frac{|t|^2}{cos\,\theta_1} = 1 - \frac{n_2 \cdot K \cdot \kappa}{n_1}\frac{|t|^2}{cos\,\theta_1}d_p$$
$$= 1 - \frac{n_2}{n_1}\frac{|t|^2}{cos\,\theta_1}d_p\frac{ln(10)\cdot\xi\cdot C}{2}$$

(2.43)

Note any reference to the relative cross-section of the beam above and below the boundary is now expressed in terms of the cosine of the angle of incidence. Grouping up terms it comes out a new depth of penetration that in the literature is referred to as effective depth of penetration.

$$deff = \frac{n_2 \cdot |t|^2}{n_1 \cdot cos\,\theta_1}\frac{dp}{2}$$

(2.44)

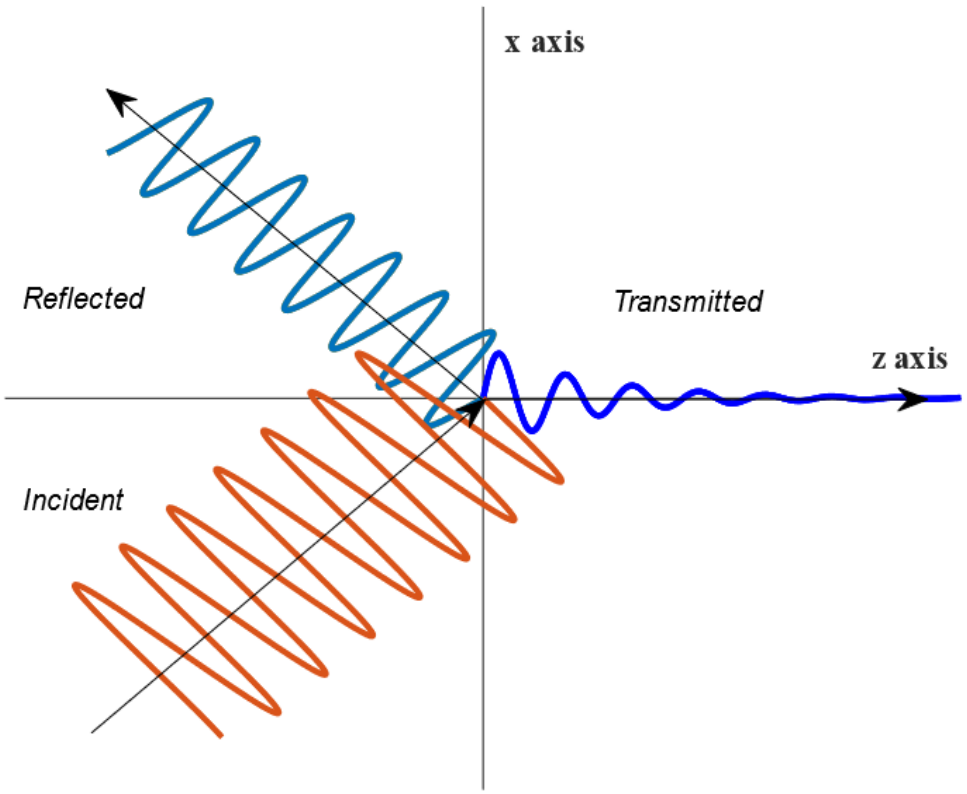

Figure 2.7. Pictorial representation of the electromagnetic wave at the boundary between the crystal and the sample.

By means of a Taylor expansion Equation 2.43 gives

$$R = 1 - ln(10) \cdot \xi \cdot C \cdot d_{eff} \simeq e^{-ln(10)\cdot\xi\cdot C\cdot deff}$$

(2.45)

And so we get the expression found in the literature referencing the pioneer of the technique (Harrick)[70].

$$A = \xi \cdot C \cdot deff = \xi \cdot C \cdot \frac{n_2 \cdot |t|^2}{n_1 \cdot cos\,\theta_1} \frac{dp}{2}$$

(2.46)

A different approach to derive Equations 2.43 and 2.46 is by thinking of the problem as a transverse wave that propagates in the *z* direction and whose amplitude decays exponentially due to the nature of the evanescent wave and further because of absorption (Figure 2.8)

$$E \quad = Re\left[e^{i\cdot\phi} \cdot E_o \cdot t \quad \cdot e^{-\kappa\cdot k\cdot z} \cdot e^{-z\cdot\gamma}\right]$$

(2.47)

where ϕ summarizes the argument of the complex part, which goes away when squaring the electric field for the sake of the flux of energy as it will be seen later in this section.

The intensity of light at any position z considered will be given by

$$I^t = \varepsilon_2 \cdot v_2 \cdot E_o^2 \cdot |t|^2 \cdot \mathbb{C}_2(e^{-2\gamma\cdot z} \cdot e^{-2\kappa\cdot k\cdot z})$$

(2.48)

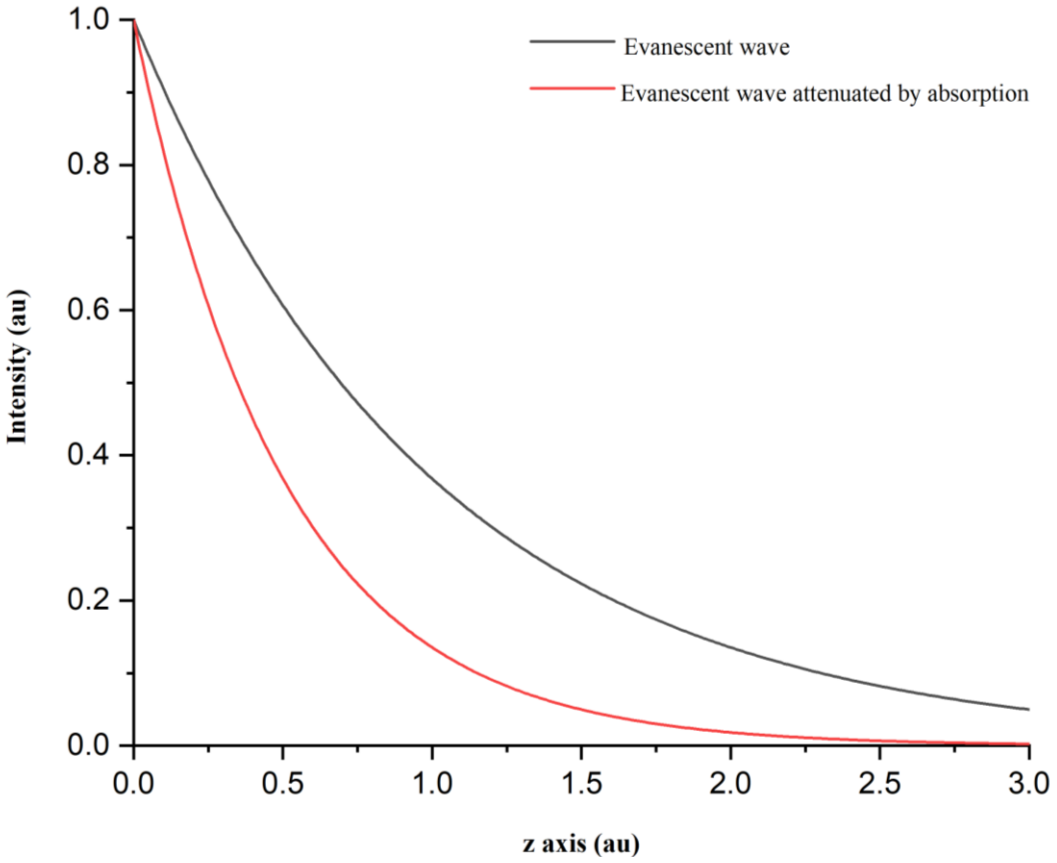

Figure 2.8. Graphical representation of the decay of intensity of the evanescent wave above the surface of the crystal with (red) and without (black) absorbent.

However, the total pathlength an evanescent wave goes is infinite and thus it might be more convenient from a mathematical point of view to convert the problem from a wave whose amplitude decays exponentially and travels to infinite into a wave with constant amplitude that goes a limited and well defined distance which is what happens in transmission. In order to do that, the total energy was considered to be contained into a cylinder of cross section $\mathbb{C}_2$ and limited length z (Figure 2.9).

By integrating the density distribution of the evanescent wave in all the volume that occupies it is possible to get a pathlength equivalent to that of transmission

$$U = \int_0^\infty u \cdot dV = \varepsilon_2 \cdot E_{o2}^2 \cdot \mathbb{C}_2 \cdot \int_0^\infty e^{-2 \cdot z \cdot \gamma} dz = \varepsilon_2 \cdot E_{o2}^2 \cdot \mathbb{C}_2 \frac{dp}{2} \tag{2.49}$$

where $z = d_p/2$

where U is the total energy in the infinite volume and E_{o2} means $E_o \cdot t$.

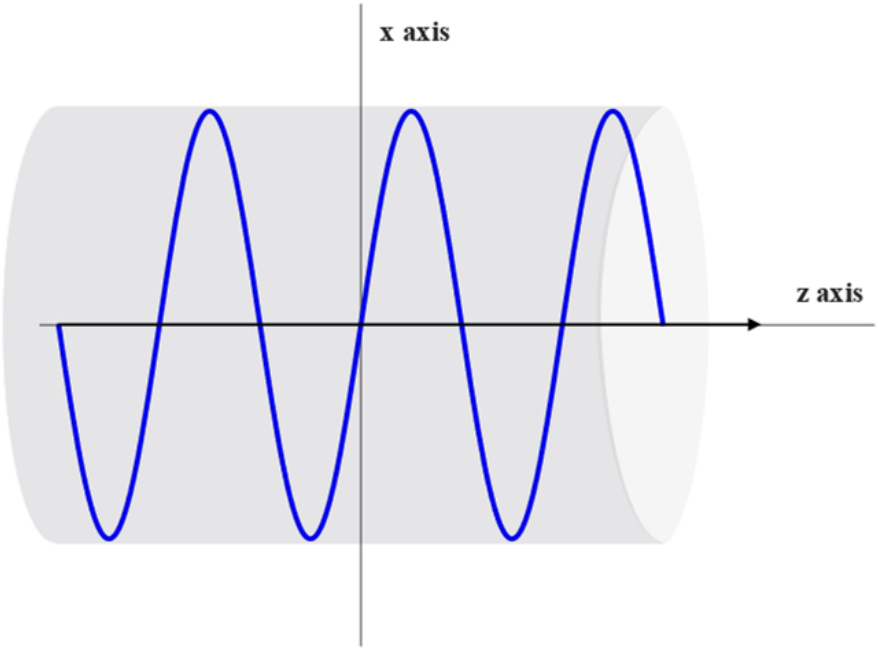

Figure 2.9. Pictorial representation of the density energy of the evanescent wave in a volume of finite length.

The absorbed light I^t above the surface then becomes

$$I^t = \langle S_{o2} \rangle \cdot \mathbb{C}_2 \left(1 - e^{-2\kappa \cdot k \cdot dp/2}\right) \tag{2.50}$$

$$\langle S_{o2} \rangle = \varepsilon_2 \cdot \upsilon_2 \cdot E_o^2 \cdot |t|^2 \tag{2.51}$$

where *t* is the Fresnel term that corrects the amplitude of the electric field above the crystal.

$$R = 1 - \frac{I^t}{I_o^r} = 1 - \frac{|t|^2 \cdot \upsilon_2 \cdot \varepsilon_2 \cdot \mathbb{C}_2 (1 - e^{-2\kappa \cdot k \cdot z})}{\upsilon_1 \cdot \varepsilon_1 \cdot \mathbb{C}_1}$$
$$= 1 - |t|^2 \frac{n_2}{n_1} \frac{\mathbb{C}_2}{\mathbb{C}_1} (1 - e^{-2\kappa \cdot k \cdot z}) \tag{2.52}$$

The Fresnel coefficient for the reflected light when no absorption takes place is 1 (total internal reflection). The expression inside the round brackets in Equation 2.52 can then be approximated by the linear term of a Taylor expansion[74]

$$R = 1 - \frac{n_2 \cdot |t|^2}{n_1 \cdot cos\,\theta_1} \cdot \kappa \cdot k \cdot z$$

(2.53)

which is the same as Equation 2.43. By means of another Taylor expansion and Equation 2.25 to replace the index of absorption κ yields

$$R = 1 - ln\,1\,0 \cdot \xi \cdot C \cdot d_{eff} \simeq e^{-\ln 10 \cdot \xi \cdot C \cdot deff}$$

(2.54)

which is the same as Equation 2.45. The absorbance is then

$$A = \xi \cdot C \cdot deff = \xi \cdot C \cdot \frac{n_2 \cdot |t|^2}{n_1 \cdot cos\,\theta_1} \frac{dp}{2} = \xi \cdot C \frac{n_2 \left(|t_x|^2 + \left|t_y\right|^2 + |t_z|^2\right)}{n_1 \cdot cos\,\theta_1} \frac{dp}{2}$$

(2.55)

where the term *t* can be expressed in terms of its cartesian components *x*, *y* and *z*.

Although the approximations -based on Taylor expansions- of Equation 2.52 lead to the same expressions found by the first approach, its derivation was to some extent based on intuition and its accuracy has not yet been tested on highly absorbing media. Thus, caution needs to be taken when applying it to spectra with higher absorption values *(A~1)*.

Now, it is necessary to derive the Fresnel coefficients for the case of incident light above the critical angle and in the presence of an absorbent or which is the same, with a complex refractive index.

Fresnel coefficients are as follows[72,81,83]

$$t_s = \frac{2n_1 \cdot cos\,\theta_1}{n_1 \cdot cos\,\theta_1 + \sqrt{\eta_2^2 - n_1^2 \cdot sin^2\,\theta_1}}$$
$$t_p = \frac{2n_1 \cdot \eta_2 \cdot cos\,\theta_1}{\eta_2^2 \cdot cos\,\theta_1 + n_1\sqrt{\eta_2^2 - n_1^2 \cdot sin^2\,\theta_1}}$$

(2.56)

where *s* and *p* mean, perpendicular and parallel components of the electric field relative to the plane of incidence (*zx*) respectively.

Above the critical angle, Equations 2.56 can be written as

$$t_s = \frac{2n_1 \cdot cos\,\theta_1}{n_1 \cdot cos\,\theta_1 + i\sqrt{n_1^2 \cdot sin^2\,\theta_1 - \eta_2^2}}$$
$$t_p = \frac{2n_i \cdot n_t \cdot cos\,\theta_i}{\eta_2^2 \cdot cos\,\theta_1 + i \cdot n_1\sqrt{n_1^2 \cdot sin^2\,\theta_1 - \eta_2^2}}$$

(2.57)

and when absorption occurs, the refractive index can be then expressed as in Equation 2.24 and the resulting square root terms as in Equation 2.30 subsequently

$$\sqrt{n_1^2 \cdot sin^2\,\theta_1 - (n_2 + i \cdot \kappa)^2} = \sqrt{n_1^2 \cdot sin^2\,\theta_1 - (n_2^2 - \kappa^2 + 2n_2 \cdot i \cdot \kappa)} = a + i \cdot b$$

(2.58)

where $a = \sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}$ and $b = -n_2 \cdot \kappa/\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}$ following Equation 2.31.

Now, since the *y* component of the electric field is perpendicular to the plane of incidence, it needs to be expressed in terms of *s* polarization. Replacing *a* and *b* into Equation 2.58 and 2.58 into Equation 2.57 for *s* and squaring gives

$$|t_s|^2 = \frac{4n_1^2 \cdot cos^2\,\theta_1}{\left(n_1 \cdot cos\,\theta_1 + \frac{n_2 \cdot \kappa}{\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}}\right)^2 + n_1^2 \cdot sin^2\,\theta_1 - n_2^2}$$

(2.59)

A further approximation is required in order to achieve the expression in the literature[70,82]. That is neglecting any term multiplied by κ (since it is assumed to be small) which yields

$$|t_s|^2 = \frac{4n_1^2 \cdot cos^2\,\theta_1}{n_1^2 \cdot cos^2\,\theta_1 + n_1^2 \cdot sin^2\,\theta_1 - n_2^2} = \frac{4n_1^2 \cdot cos^2\,\theta_1}{n_1^2 - n_2^2} = |t_y|^2$$

(2.60)

This corresponds to the *y* component of the Fresnel's term. Although it is more compact than the expanded version, it is necessary to take into consideration that this

expression might lead to significant deviations for the case of samples in water where the absorption index is not small.

Now it is necessary to derive the Fresnel terms for the *x* and *z* components. Since the amplitude of the transmitted electric field is oblique to the plane of incidence it is necessary to work out their cartesian projections first and express them in terms of the incidence angle by means of Snell's law.

$$\overrightarrow{E_{\circ_{2x}}} = -\overrightarrow{E_{\circ_2}} \cdot cos\,\theta_2 = -\overrightarrow{E_\circ} \cdot t_p \sqrt{1 - \left(\frac{n_1}{\eta_2} sin\,\theta_1\right)^2}$$

$$\overrightarrow{E_{\circ_{2z}}} = \overrightarrow{E_{\circ_2}} \cdot sin\,\theta_2 = \overrightarrow{E_\circ} \cdot t_p \frac{n_1}{\eta_2} sin\,\theta_1$$

(2.61)

Substituting now t_p using Equation 2.57 for *p* polarization but neglecting the contribution from κ in $\eta_2 = n_2 + i \cdot \kappa$ again yields

$$\overrightarrow{E_{\circ_{2x}}} = -\overrightarrow{E_\circ} \frac{2n_1 \cdot cos\,\theta_1}{n_1 \cdot cos\,\theta_1 + i\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}} i \sqrt{\left(\frac{n_1}{n_2} sin\,\theta_1\right)^2 - 1}$$

$$\overrightarrow{E_{\circ_{2z}}} = \overrightarrow{E_\circ} \frac{2n_1 \cdot cos\,\theta_1}{n_1 \cdot cos\,\theta_1 + i\sqrt{n_1^2 \cdot sin^2\,\theta_1 - n_2^2}} \frac{n_1}{n_2} sin\,\theta_1$$

(2.62)

and adding up the *x* and *z* components we get

$$|t_p|^2 = \frac{|E_{o2x}|^2 + |E_{o2z}|^2}{|E_o|^2} = \frac{4n_1^2 \cdot cos^2\,\theta_1\,[(n_1^2 \cdot sin^2\,\theta_1 - n_2^2) + n_1^2 \cdot sin^2\,\theta_1]}{\left(n_2^4(1 - sin^2\,\theta_1)\right) + n_1^2[n_1^2(n_1^2 \cdot sin^2\,\theta_1 - n_2^2)]}$$

$$= \frac{4n_1^2\,cos^2\,\theta_1\,[2n_1^2 \cdot sin^2\,\theta_1 - n_2^2]}{(n_1^2 - n_2^2)[sin^2\,\theta_1\,(n_1^2 + n_2^2) - n_2^2]}$$

(2.63)

Finally, for non-oriented samples and non-polarised light, the total transmitted coefficient $|t|^2$is the average of the *s* and *p* polarizations[73,84].

$$|t|^2 = \frac{|t_s|^2 + |t_p|^2}{2}$$

(2.64)

2.2.2 Anomalous dispersion correction SOP

2.2.2.1 Equations

In the previous section, the three different approximated equations that relate ATR and transmission were derived. Those equations, expressed in the form that corresponds to the conversion from ATR absorbance to transmission in $(M \cdot cm)^{-1}$ are

$$\xi = \frac{2A_{atr}}{C\frac{n_2}{n_1}d_p\frac{1}{cos\,\theta_1}|t|^2}$$

(2.65)

$$\xi = \frac{2(1-10^{-A_{atr}})}{log(10)\cdot C\frac{n_2}{n_1}d_p\left(\frac{1}{cos\,\theta_1}\right)|t|^2}$$

(2.66)

$$\xi = -\frac{2}{ln(10)\cdot C\cdot d_p}ln\left[1-\frac{1-10^{-A_{atr}}}{|t|^2\frac{n_2}{n_1}\frac{1}{cos\,\theta_1}}\right]$$

(2.67)

where ξ is the transmission extinction coefficient, A_{atr} means ATR absorbance, d_p is depth of penetration (Equation 2.69), n_2 means sample refractive index, n_1 stands for crystal refractive index, C is the sample concentration, Θ_1 means angle of incidence and $|t|^2$is the transmission Fresnel term (Equation 2.68), and 2.65 corresponds to the most approximated version of Equation 2.67 followed by Equation 2.66.

$$|t|^2 = \frac{1}{2}\left[\frac{4n_1^2\,cos^2\,\theta_1\,[2n_1^2\cdot sin^2\,\theta_1 - n_2^2]}{(n_1^2-n_2^2)[sin^2\,\theta_1\,(n_1^2+n_2^2)-n_2^2]} + \frac{4n_1^2\cdot cos^2\,\theta_1}{n_1^2-n_2^2}\right]$$

(2.68)

$$dp = \frac{1}{\tilde{\nu} \cdot 2\pi \cdot n_1 \sqrt{sin^2\,\theta_1 - \left(\frac{n_2}{n_1}\right)^2}} \tag{2.69}$$

The only ingredient missing here is the refractive index which is

$$n_2(\tilde{\nu}_i) = n_{2\infty} + \frac{ln\,1\,0}{2\pi^2} \wp \int_0^\infty \frac{\xi(\tilde{\nu}) \cdot C}{\left(\tilde{\nu}^2 - \tilde{\nu}_i^2\right)} \cdot d\tilde{\nu} \tag{2.70}$$

where *i* is the i-th element of wavenumber for which the refractive index needs to be known. To work out what the effect of the anomalous dispersion is over the range 1500-1800 with a 1 cm^{-1} step the integral would have to be performed over all spectrum for each value of wavenumber, 301 in the former considered range.

2.2.2.2 Correction refractive index calculated from experimental transmission spectrum

The integral can be calculated analytically by measuring first the experimental extinction coefficient of the sample (water + protein) and fitting it with a function of choice which in our case was a linear combination of Lorentzian functions. The whole procedure was performed through a MATLAB routine (appendices A.1). This routine consisted of several functions:

1. Importing, trimming and converting into a user defined internal vector length the experimental ATR spectrum of the considered protein.
2. Importing and trimming the experimental transmission spectrum of the protein of interest.
3. Importing and trimming the experimental transmission spectrum of the problem protein in water (water + protein).
4. Iterative fitting of the water + protein spectrum with a sum of Lorentz functions (Equation 2.71) with initial values between 20-40 for both *w* and *A*. The initial values used for the positions *u* were those found with the peak finder toolbox. Several fittings were performed with systematic variations of the initial positions resulting in different goodness of fit. The sum of the residuals of each of them

was computed and the smallest one chosen as the optimal fit to the experimental curve.

5. Calculation of the refractive index using the optimal fit from the step before in Equation 2.70. First, the pathlength is determined to convert the fit of the experimental transmission spectrum to extinction coefficient using the small water band at ~2150 cm^{-1} as a reference and a prior measured water spectrum of known pathlength (0.1 mm spacer) and then the integral is calculated.

6. Conversion from experimental ATR to transmission in $(M.cm)^{-1}$ using Equations 2.65-2.69.

$$Abs = y_o + \sum_{i=1}^{N} \frac{A \cdot w}{(\tilde{v} - u_i)^2 + w^2} \tag{2.71}$$

2.2.2.3 Correction with refractive index calculated from experimental ATR spectrum

In the previous section, the correction was performed based on the experimental refractive index measured with transmission. The use of ATR responds to our desire to avoid using transmission due to the experimental complexity it means (to be discussed in chapter 3) so there is no point in measuring transmission to correct for anomalous dispersion since our goal is to avoid measuring transmission in the first place. To solve this, an iterative method for the estimation of the refractive index from ATR water was implemented in MATLAB with similar functions as the ones described in the previous section but using the experimental ATR spectrum of the protein in water instead. The code performs several cycles of the functions [fitting-refractive index calculation-conversion-fitting-refractive index calculation-conversion-…] until the value of the corrected ATR and refractive index do not change significantly anymore or what is the same, until self-consistence. See codes (Appendices A.2) for further information.

2.3 RESULTS

2.3.1 Correction with refractive index calculated from transmission spectrum

Three transmission spectra replicates of a 50 $mg.ml^{-1}$ Lysozyme solution were first zeroed, scaled, averaged and; water and vapour subtracted as it will be explained in

chapter 3. Then, they were fitted with a sum of Lorentz functions and this sum used to calculate the refractive index within the range 1500-1800 cm^{-1} (Figure 2.10).

Figure 2.11 shows the anomalous dispersion correction by means of Equation 2.65 which is the popular Harrick's expression found in the ATR spectroscopy literature. The success of this approximation requires absorbance to be very small compare to unity so that higher order terms in the expansion can be neglected[85].

Figure 2.12 shows the correction based on Equation 2.66 which also requires absorbance to be well below unity (values of absorbance in the region of the amide I by a 50 $mg.ml^{-1}$ protein solution with a 1 bounce ATR unit are ~0.15).

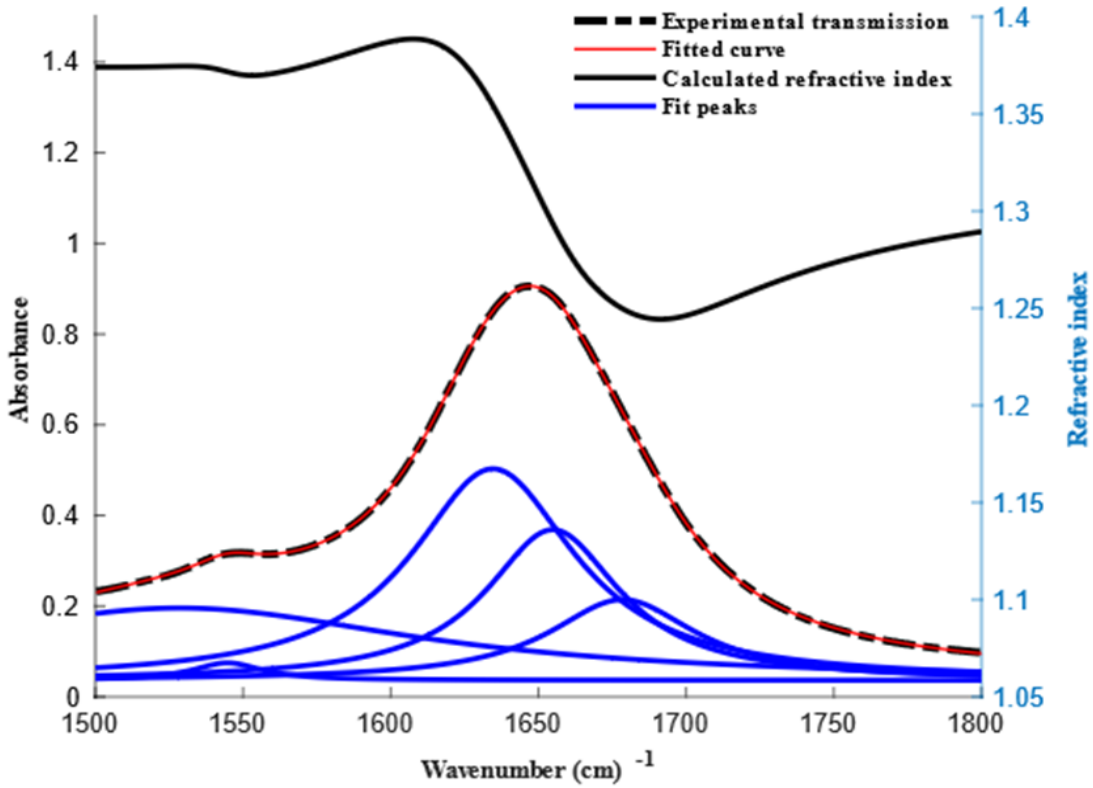

Figure 2.10. Band fitting of the transmission spectrum of a 50 mg/ml solution of Lysozyme in water and its corresponding calculated refractive index.

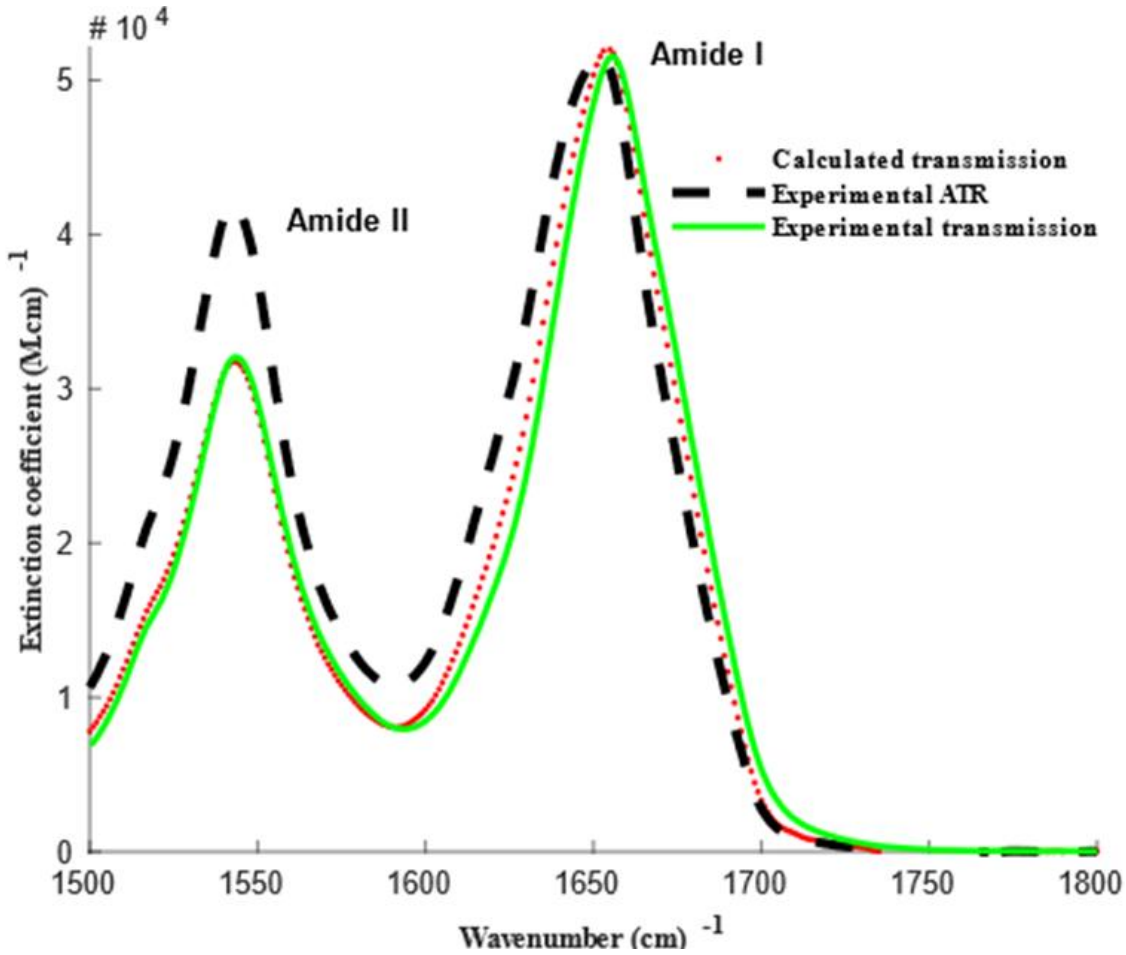

Figure 2.11. Simulated transmission spectrum of Lysozyme calculated from transmission with Equation 2.65.

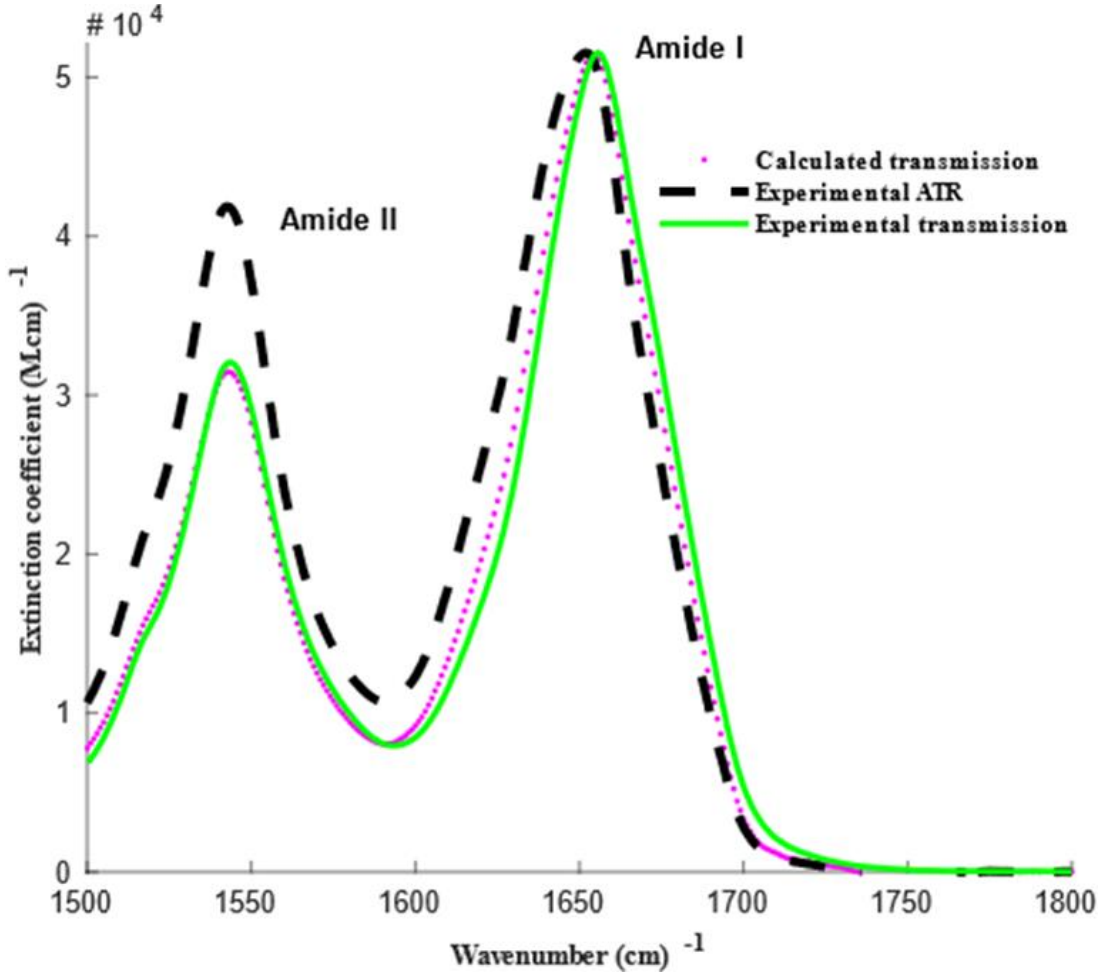

Figure 2.12. Simulated transmission spectrum of Lysozyme calculated from transmission with Equation 2.66.

Finally, the correction done with Equation 2.67 is shown in Figure 2.13.

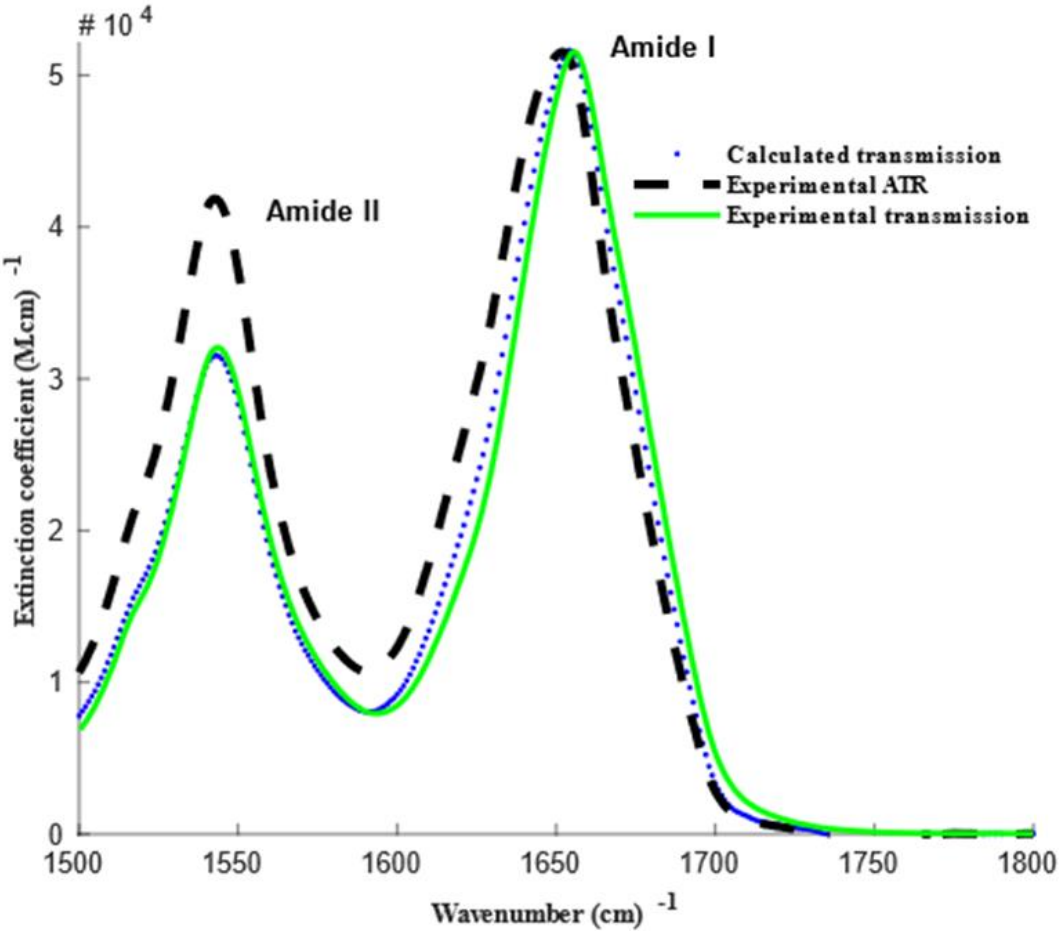

Figure 2.13. Simulated transmission spectrum of Lysozyme calculated from transmission with Equation 2.67.

As it can be seen in Figures 2.11-2.13, there is not much difference between the corrections based on Equations 2.65-2.67. Because in this case the absorbance was significantly below unity (~0.15 for a 50 mg.ml^{-1} protein in water measured with a single bounce ATR unit), Equation 2.67 could be approximated through two consecutive expansions without significant difference.

2.3.2 Correction with refractive index calculated iteratively from experimental ATR spectrum

The ATR spectra of the proteins in water were taken as an initial value to approximate their refractive index, which were then used to estimate the transmission extinction coefficient iteratively until self-consistence. The range considered was from 1500 to 2300 cm^{-1} to include the liberation + deformation combination band from water that again was used as a reference to work out the

pathlength. No significant improvement was achieved above 4 iterations (Figures 2.14-2.15).

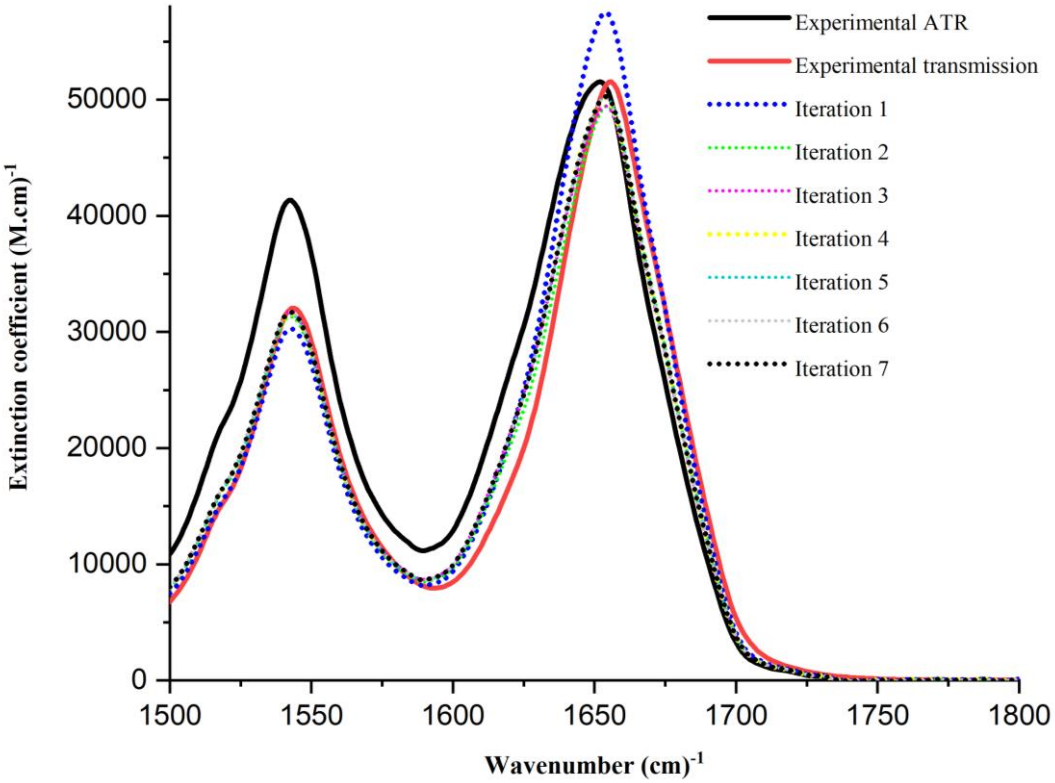

Figure 2.14. Simulated transmission spectrum calculated from ATR iteratively with equation 2.67.

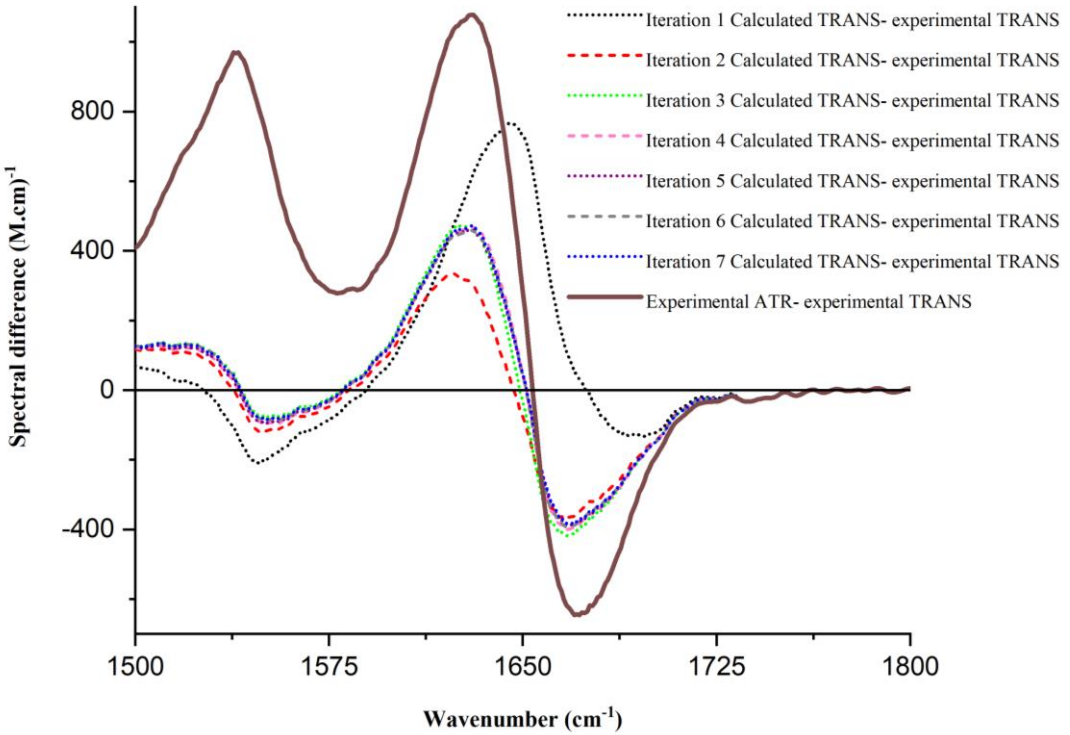

Figure 2.15. Spectral difference representation for the different iterations compared to the scaled experimental ATR.

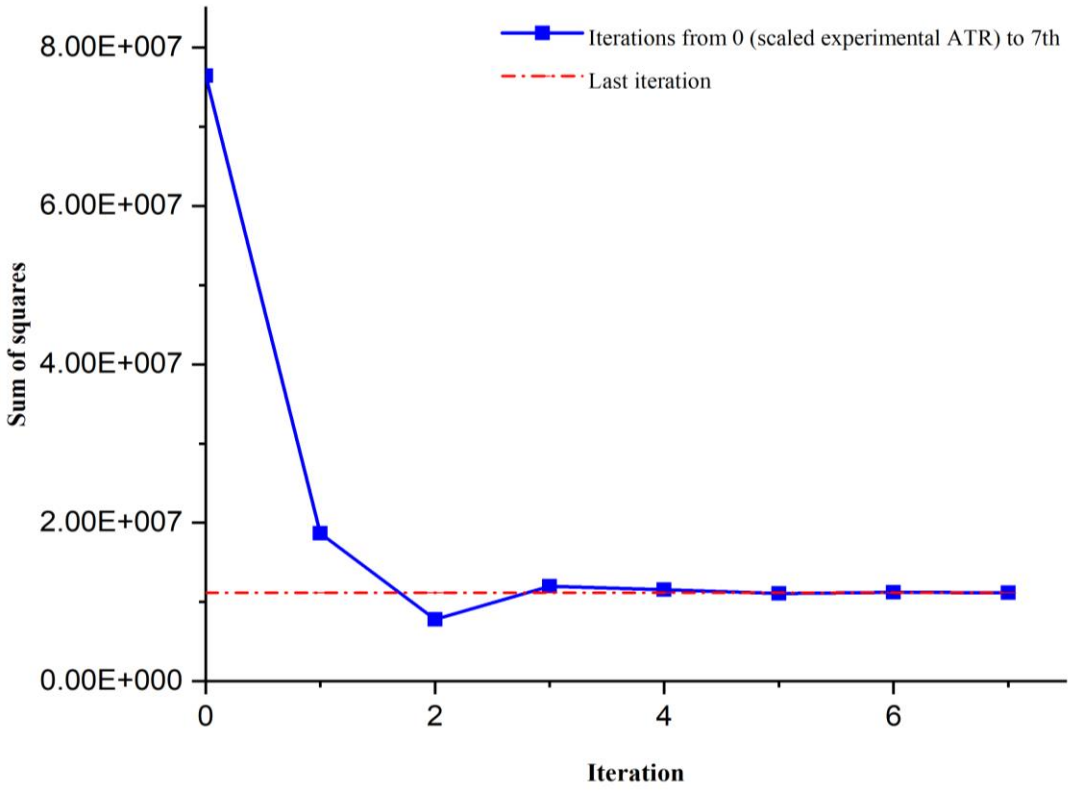

Figure 2.16. Sum of squares vs number of iterations.

The last iteration reduces the difference with respect to the experimental ATR by ~84 % (Figure 2.16).

2.4 CONCLUSIONS

Although the transmission spectrum was fairly-well reproduced from the experimental ATR in terms of relative intensities and shift by both methods, there were still some notable deviations that should be considered. These deviations might be attributed to some or a combination of the following aspects:

1. The range of the integral for the calculation of the refractive index, which was restricted for simplicity to short ranges in both cases, should actually include the totality of the spectrum.

2. The fitting of the spectra must be considered as a potential cause of error because no fit will ever be a perfect representation of the experimental band. Poorer fittings were achieved on the experimental ATR than on the transmission spectrum because of the challenging of the procedure: automated, wider spectral range and more changeable through iterations because of the accumulative corrections performed on it.

3. The angle of incidence, that seemed to be ~43º based on the correction factor applied to make the corrected ATR match the experimental transmission and confirmed by an eye examination of the mirrors inside the ATR accessory.

4. The polarization of the beam: another potential source of error could be the lack of perfectly unpolarized light which would affect the weights of the *s* and *p* components in Equation 2.68.

3 INFRARED SPECTROSCOPY OF PROTEINS AND SELF-ORGANIZING MAPS

3.1 INTRODUCTION

Infrared spectroscopy is one of the most well-established techniques for structural characterization and stability studies of proteins under stressing conditions such as pH, temperature, ionic strength and a wide range of environments e.g., organic solvents, membrane bound, etc[46,86–90]. IR spectra of proteins show characteristic bands (Amide A, B, I, II, III, IV, V, VI and VII) with the amide I being the most sensitive to the backbone conformation and popular for characterization purposes[6,20,22,46], although not the only one related to higher order structures[89]. This band consists of overlapped components that spread across the range between 1600 and 1700 cm^{-1} and whose position reveal characteristic secondary structures segments[6,20,22,46]. The most well-established technique to extract the relative contents of secondary structure is band fitting. The subjacent components of the amide I are first resolved (by Fourier Self-Deconvolution (FSD) or second derivative), then fitted with linear combinations of Gaussians and the fraction of the areas in the different secondary structure ranges computed[6,46,90]. Other approaches based on multivariate analysis (PLS and multi linear regression)[88,91] and also neural network algorithms[92] seem to be blooming in the recent years.

In this chapter, the secondary structure content of a set of proteins of well-known secondary structure (SS) is determined by Fourier Transform Infrared Spectroscopy-Attenuated Total Reflectance (FTIR-ATR) and a neural network algorithm called Self-Organizing Maps (SOM)[1] that we developed from a previous version (SSNN)[2] used to fit Circular Dichroism (CD) spectra. Moreover, we also evaluated the differences in secondary structure predictions between attenuated total reflectance and transmission in relation to the band distortions that arise from anomalous dispersion as discussed in the previous chapter and discussed the significant spectral differences between solid and aqueous state of amide I bands.

Although there are many existing experimental procedures for IR data collection of proteins in the literature, they are mostly for transmission, in D_2O or of high concentrated samples in H_2O. We found these approaches for proteins in aqueous

state particularly ambiguous in their water and vapour subtraction criterion. In this chapter we review the existing protocols for data collection, push down the sample concentrations towards the instrument limits in order to meet the top possible concentrations measurable by CD and address the need for unambiguous subtraction methods of water and vapour.

3.2 Self- Organizing Maps

A Self-Organizing Map (SOM) is a type of neural network architecture created by Teuvo Kohonen that produces a 2D representation from a higher dimensional input space and thus helps the visualisation and identification of structures. The new space is discrete and consist of nodes disposed in layers of arrays. Unlike traditional Neural Network algorithms, SOM is a type of unsupervised learning technique, which means the input data has no categorical labels[93–95].

It works in three steps: firstly, organises the spectra of a reference set of proteins clustering them in terms of spectral similarity, given by the distance between the nodes on the map; secondly, assigns the secondary structure to the nodes by sum weighing the contributions of the neighbouring proteins and third; it tests the sample spectrum against the map by identifying the best matching unit (BMU) in terms of the distance in the spectral space and then works out the SS by sum weighing the SS of the top 5 matching nodes in terms of the distance on the map[1,95].

3.2.1 Training and structure assignment

Firstly, all the input vectors -reference set and test samples- are converted into a fixed length internal vector representation, using a cubic spline-based interpolation method to match them in size, since no algebra is possible with vectors of different lengths.

In the map construction stage, a 2D map is initially populated with random input vectors representing the spectra. Then, the reference set (actual input vectors) is randomly sampled over a user defined number of iterations. At each iteration a node is identified as the BMU of the input vector under consideration, defined by the shortest Euclidian distance (Equation 3.1) between the node vector and the input one in the space defined by the spectral range of the reference set.

$$d = \sqrt{\sum_{i=0}^{N} (I_i - W_i)^2}$$

(3.1)

where I_i and W_i are the current input and node vectors respectively, d means Euclidian distance between both vectors, the index i indicates the element in the vector (wavelength) and N is the number of elements in the vector. Once the BMU is identified, it and all other nodes within the current neighbourhood are changed according to Equation 3.2

$$W(t+1) = W(t) + \theta(t) \cdot L(t)[I(t) - W(t)]$$

(3.2)

here, t is the current iteration and $(t+1)$ the one that follows, θ is the radial bias function (Equation 3.3) which determines the influence of the current input vector in its proximities -also known as neighbourhood- and L is the learning rate and means the extent of learning for each iteration (the default learning rate used by us was 0.1 but is user configurable).

$$\theta(t) = exp\left[\frac{-\big(D(t)\big)^2}{2\big(R(t)\big)^2}\right]$$

(3.3)

where D is the radial distance from the BMU and R the distance at which the value of the function is $1/\sqrt{e}$ and thus the parameter that defines how rapidly the influence of the node drops across the map. Both the radius and learning rates decay with exponential fashion over iterations following Equation 3.4

$$Z(t+1) = Z(t) \cdot exp\left[\frac{-t}{c}\right]$$

(3.4)

where Z represents both the learning rate and the radius. The time constant c determines the rate at which Z drops throughout iterations and it is calculated from the total number of iterations (*Iter*) and initial radius (R_o) by Equation 3.5

$$c = \frac{Iter}{ln(R_o)}$$

(3.5)

Once the predefined number of iterations is reached the second stage of training takes place with the sequential assignment to the map of the proteins in the reference set and their respective properties (SS contents). The input vectors corresponding to the proteins in the set are sequentially sampled against the map and a BMU based on a minimised NRMSD calculation (Equation 3.6) is identified for each of them. The protein and its SS content are assigned to it and the surrounding nodes changed using Equation 3.3 with the initial values for the bias radial function and learning rate

$$NRMSD = \frac{\sqrt{\frac{\sum(Y-X)^2}{N}}}{M-m}$$

(3.6)

where *Y* and *X* are the predicted and experimental values respectively, *N* is the total number of elements in the vector and *M-m* the degrees of freedom.

Finally, the test spectrum is put through the algorithm in order to determine its SS. A predefined number of BMU (usually 5) identified by the means previously described (based on minimised NRMSD) are ranked by similarity and the property contribution of each calculated using a distance dependent weighting in accordance to Equation 3.3. An output of the trained map with the BMUs (in red) and the reference set (in blue) is generated for visualization along with the spectrum fitted with a weighed sum of the BMUs and its corresponding NRMSD value (Figure 3.1).

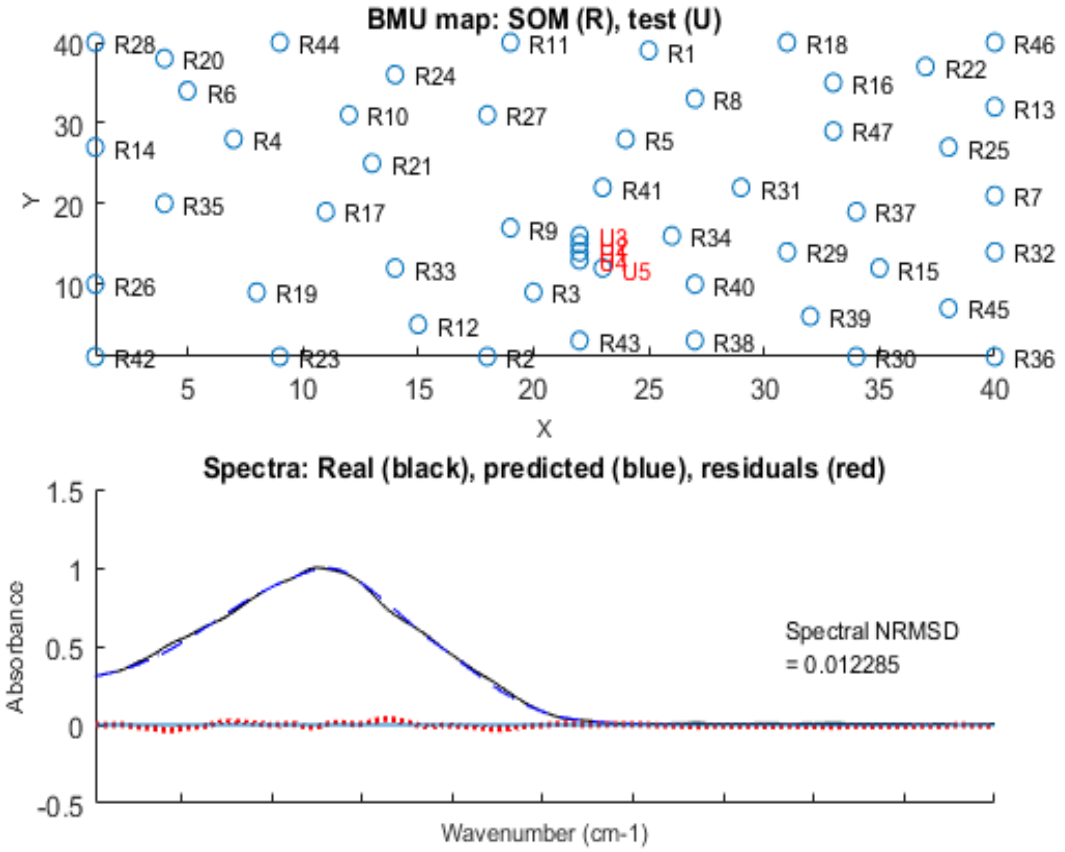

Figure 3.1. BMU positions and predicted spectrum output for a secondary structure prediction from a proteins IR spectrum. The top BMUs contributing to the prediction are displayed in red. The predicted spectrum (blue) overlays the input spectrum (black).

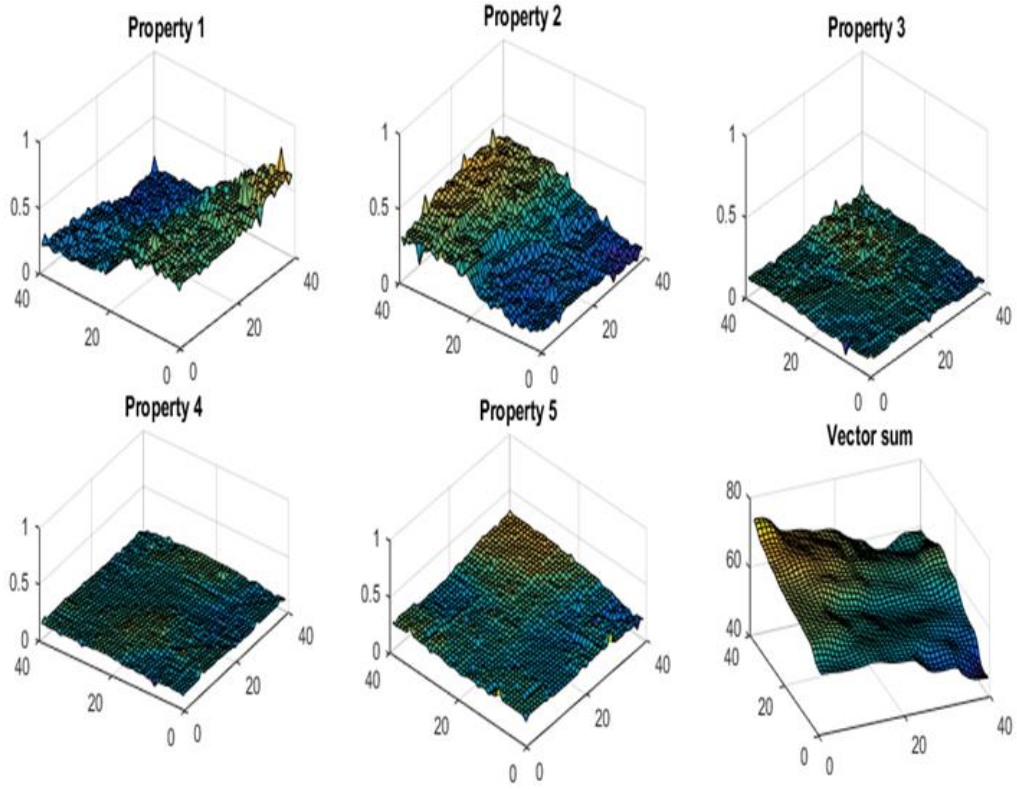

Figure 3.2. Distribution of the properties across the map.

3.3 BAND ASSIGNMENT AND SECONDARY STRUCTURE OF PROTEINS

3.3.1 Overview

IR spectroscopy is one of the most versatile characterization techniques not only due to the wide range of samples and variety of conditions it can be applied to as discussed above but also because of its ability to be used to predict primary structure and higher order structures as well. In IR absorption, light excites different structural groups when it is in resonance with their natural oscillation frequencies. The intensity of the bands is measured in terms of the extinction coefficient, which depends on the likelihood of the transition (transition dipole moment) and relates to the polarity of the bonds. The positions instead, are related to the architecture of the molecule and they are characteristic of the different functional groups which in a broad manner can be put in terms of the constant force and reduced mass of the groups (Equation 3.7)[22,41,46].

$$\tilde{\nu} = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}} \tag{3.7}$$

Furthermore, their position might be affected by their neighbouring groups by inductive and resonant effects and environment e.g., solvent polarity, hydrogen bonding, temperature, pH, etc[20,22].

Protein spectra consist of 9 characteristic bands named amide A, amide B, Amide I-VII from higher to lower frequency[20] (Figure 3.3). Miyazawa, Shimanouchi and Muzushima published in 1958 a detailed discussion on the structural nature of the amide I, II and III bands of N-Methylacetamide and its deuterated compound by means of normal coordinate calculations[96]. They assigned the amide I (about 1650 cm^{-1}) to the C=O stretch (~80%) with the remaining contribution coming from the in-plane N-H bending (~10%) and C-N stretch (~10%). The amide II (about 1550 cm^{-1}) was attributed to both the N-H in plane bending (~%60) and the C-N stretch vibration (~40%) and the amide III (~1300 cm^{-1}) was assigned to the other phase of the coupling of the C-N stretch with the N-H in plane bending. The assignments of

these bands collected from more recent literature reviews are summarized in Table 3-1[6,20,22].

Table 3-1. Characteristic IR bands of proteins.

Amide	Frequency (cm^{-1})	Assignment
A	~3300	NH stretch in resonance with 1st amide II overtone
B	~3100	
I	1600-1700	C=O stretch
II	1480-1575	N-H in plane bending coupled with C-N stretch
III	1200-1320	
IV	625-765	O-C-N bending, with contribution of other modes
V	640-800	Out of plane N-H bending
VI	535-605	Out of plane C=O bending
VII	~200	Backbone torsion

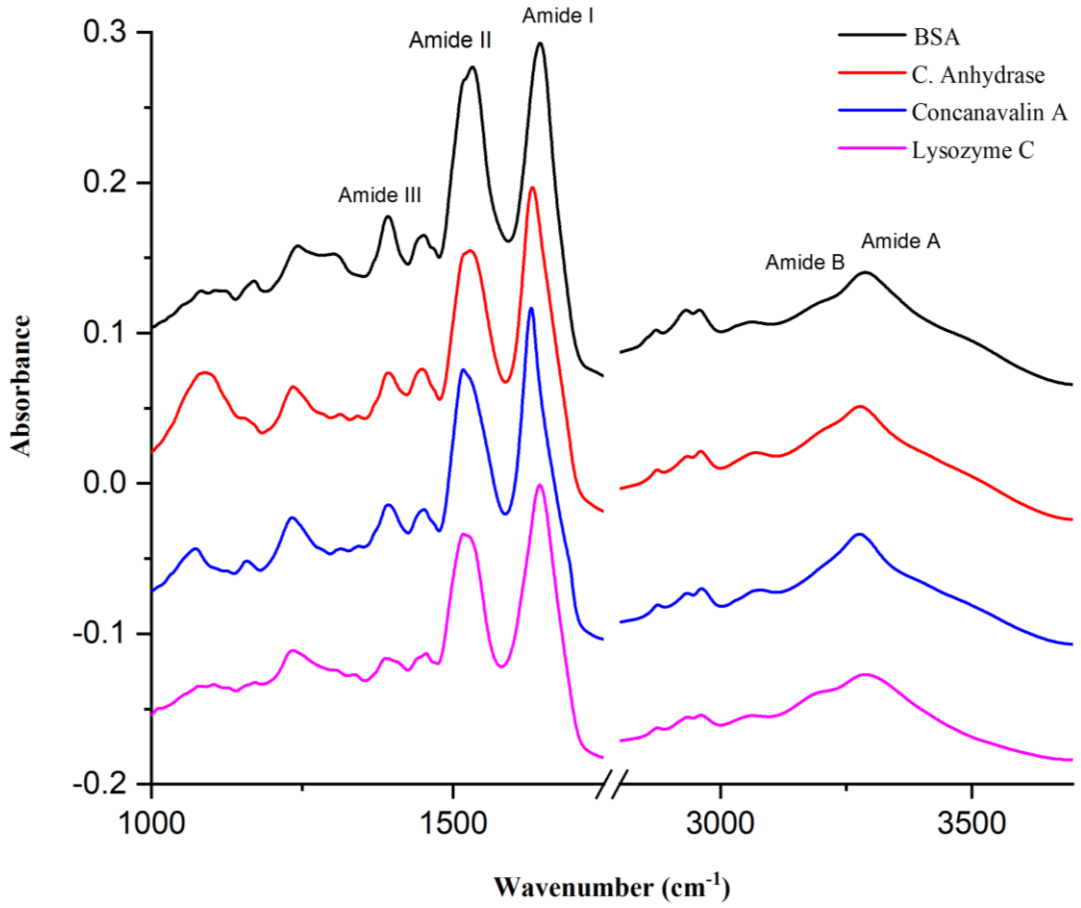

Figure 3.3. IR spectra of Lysozyme, Concanavalin, BSA and C. Anhydrase in solid state. The data was collected with a single bounce ATR plate, 4 cm^{-1} resolution and averaged over 63 scans. Only the most prominent amides (A, B, I, II and III) are shown.

3.3.2 Amide I and secondary structure

As early as in 1950, Elliot and Ambrose discovered there is a close relationship between secondary structure of synthetic polypeptides and the exact position of their carbonyl stretch band in the region between 1600 and 1700 cm^{-1} [97]. They found that synthetic Poly-L-glutamic benzyl ester forms helical structures, Poly-L-glutamic methyl ester both helical and sheet when crystalized from m-cresol and only sheet when done from formic acid. Furthermore, they also measured their IR spectra and found components at 1629 and 1659 cm^{-1} corresponding to the C=O stretch that they assigned to sheet and helical contents respectively by comparison to X-Ray data. Later in 1960, Miyazawa applied a first order perturbation treatment to the amides I and II vibrations of polypeptide chains in order to explain how different conformations give rise to the experimental frequencies [98]. He found that inter and intramolecular hydrogen bonding interactions in the different configurations play an important role in the shift of amides I and II. More complex analysis based on

Miyazawa's approach were later performed on globular proteins in fairly good agreement with the previous results found for synthetic and fibrous proteins[19,99–101]. In the years after, Gaussian fitting methods were also applied in conjunction with X-Ray and CD characterization for a refinement on the assignment of the amides I and II and semiquantitative analysis of some globular proteins[90,102]. All this information kept being updated over the years by experimental and theoretical means and compiled in recent reviews that are used in next section for a more detailed and discussion on the assignment of the amide I band.

In 2007 Barth compiled in a review all the information to the date on the effects that modify the amide I frequency[22]. The known dominant causes for the different characteristic frequencies of the different conformations are: through-bond coupling, hydrogen bonding and transition dipole coupling. Vibrations of the amide I and II cause small displacements of the C^{α} that allows a small through-bond coupling of nearest neighbour vibrations of the same kind although the effect in position is mild. However, hydrogen bonding to C=O and N-H groups of the peptide back-bone were proved to significantly lower the frequency of the Amide I. But none of the former mentioned phenomena explained the observed splitting for beta sheet bands. The introduction of transition dipole moment coupling[100] was necessary to explain the splitting of antiparallel beta sheet in a main band about 1630 cm^{-1} and a weaker one about 1690 cm^{-1}. Over the years some rules based on computation and experimentation have been established to read the backbone structure of proteins from absorption of amide I bands. Kong and Yu collected in a 2007 review[6] all the characteristic bands and their assignments based on previous studies by Byler, Susi, Dong et al[21,90,103,104] (Table 3-2).

In general, helices appear at about 1655 cm^{-1} and their position shifts down with increasing helix length, when it is bent or solvent exposed. Antiparallel sheets show a strong band between 1620 and 1630 cm^{-1} and a weak one between 1685 and 1695 cm^{-1}. The sheet position is hardly affected by the number of residues but by the number of strands instead. Sheets with larger number of strands have a lower frequency of the main band. With twisting, the main band shifts upwards and the magnitude of the splitting is smaller. The main bands of parallel sheets absorb at higher frequencies than antiparallel as a general rule (~4 cm^{-1}) and they have a much smaller splitting than antiparallel too. Differences between parallel and antiparallel

are less obvious when antiparallel sheets have fewer strands or are twisted and parallel instead have greater number of strands and no twists.

Table 3-2. Assignment of SS to amide I band in H_2O.

Mean frequencies (cm^{-1})	Assignment
1624±2	β- sheet
1627±2	β- sheet
1633±2	β- sheet
1642±1	β- sheet
1648±2	Random
1656±2	α- helix
1663±3	3_{10} helix
1667±1	β- turn
1675±1	β- turn
1680±2	β- turn
1685±2	β- turn
1691±2	β- sheet
1696±2	β- sheet

3.4 MATERIALS AND METHODS

3.4.1 Samples and reagents

All the proteins used were purchased from Sigma Aldrich and any dilution and blank measurement carried out with Milli-Q (ultra-pure) water. The proteins used for the aqueous and solid-state experiments were 21 and 31 respectively (Appendices-D). The powders were grounded with KBr (1-10 %w).

3.4.2 Instrumentation

Centrifuge sigma D-37520 14k, Millex-GV 0.22 µm syringe filter disks, sringe, Ika vortex genius 3, marble mortar, Jasco 660 UV-Vis and a Jasco 4200 FT-IR spectrometer equipped with ZnSe ATR units: a PIKE MIRAcle 1-bounce ATR unit and a Specac 6-bounce ATR unit.

3.4.2.1 Infrared Instrument set up

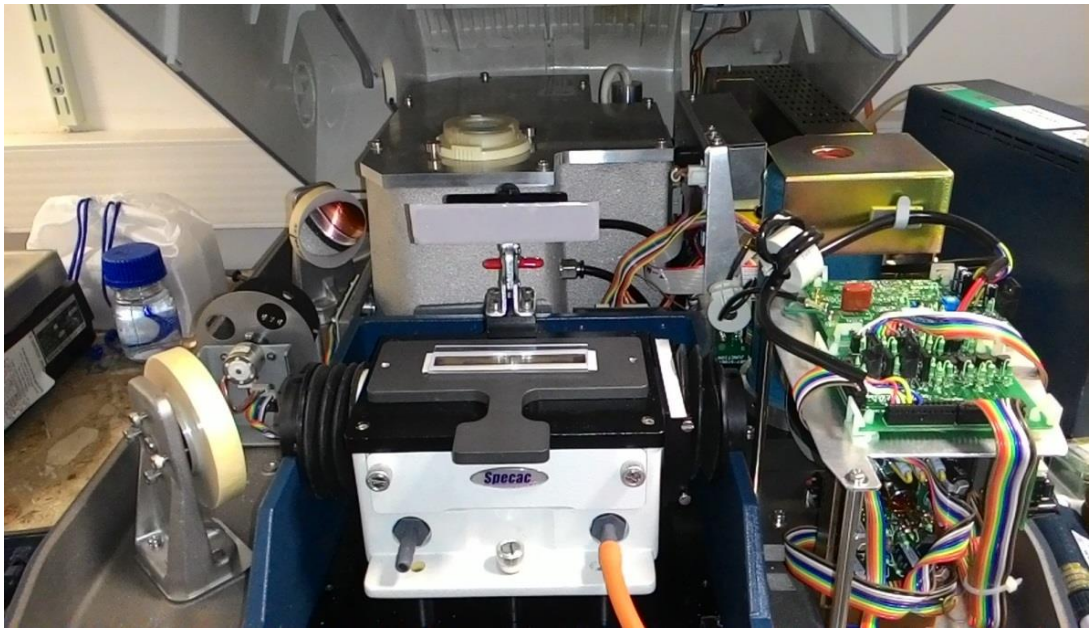

Figure 3.4. Inside of Jasco J-4200 FTIR instrument equipped with a Specac multibound ZnSe ATR unit in position.

The instrument used for these experiments was a Jasco J-4200 spectrometer that consists of the following elements:

- A resistance heating element source (ceramic rod) that produces radiation in the region between 7000 and 400 cm^{-1} when heated by a voltage.
- An additional monochromatic laser beam to calibrate the position of the movable mirror.
- A Michaelson Interferometer as a frequency modulator.
- A PIKE MIRAcle single bounce ZnSe ATR unit, a Specac six bounce ZnSe ATR unit and a Specac transmission cell equipped with CaF_2 windows and Teflon spacers.
- A pyroelectric (TGS-Deuterated Triglycine Sulfate) and photonic (MCT-Mercury, Cadmium, Telluride) detectors.

Because the source takes a long time to warm up, the baseline drifts upwards over the first couple of hours after the instrument is turned on. The instrument was therefore usually turned on at least 3 h before the start of an experiment to prevent variation in the baseline.

The instrument is fitted with an in and outlet valve for the interferometer and an additional one for the sample chamber as a part of an inner purge system that helps keep humidity (water vapour) at the lowest possible level when N_2 is passed through it.

Although transmission IR is deemed to be the definitive spectrum, given the high absorbance of water and the difficulty of reproducibly assembling few microns path-length demountable cells, we chose to use preferentially attenuated total reflectance sample holders (Figures 3.6 and 3.7) over transmission ones. However, we decided to collect the reference set in transmission-mode to avoid biasing the data with anomalous dispersion as any correction is time consuming to perform and will never be as good as data measured by transmission.

For the six bounce ATR we had to create a Teflon mask to produce a reproducible separation between the lid -a microscope slide glass instead of the clamp in the accessory- and the sample (Figure 3.7). This proved to improve the reproducibility of the baseline in the region between 1750 and 2500 cm^{-1} which is important for our ability to accurately subtract the buffer meaning less artefacts and smaller factors to be applied. Besides helping the baseline with a homogeneous thickness of the sample across the whole crystal surface, the glass cover also avoids introduction of air bubbles and evaporation during data collection. The presumption is that, since the evanescent wave decays exponentially from the surface of the crystal, the closer the lid is the more significant will be the optical phenomena that occur in it and the less reproducible the baseline. A simple experiment was designed to determine in a rough way the depth of the evanescent wave. The crystal plate was taped with different layers and the intensity of the C-H stretch vibrations between 3000 and 3100 cm^{-1} measured. It was found that 2 layers of tape (~0.25 mm) were sufficient to deem the intensity of the electric field has decayed to zero (Figure 3.5). A home-made Teflon mask was then glued to the ATR plate, surrounding the crystal to create a cavity with the depth found above (~0.25 mm).

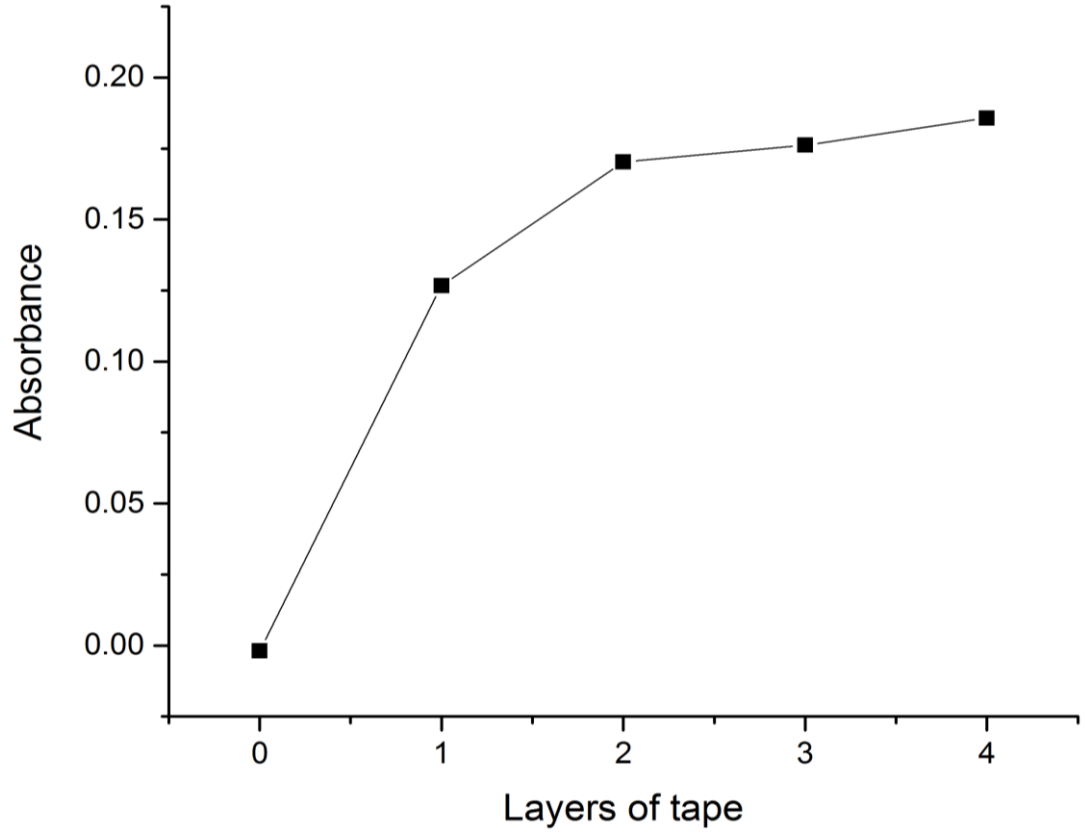

Figure 3.5. Absorbance of C-H stretch mode from tape (3100 cm^{-1}) as a function of the layers of black tape used to cover the crystal.

As for the water vapour, a tube from the outside of the instrument was connected directly to the sample holder as shown in Figure 3.4 so that a positive pressure of an inert gas (N_2) reduces the water vapour in the system. An additional flow (from the inside) was used to purge the atmosphere in the sample chamber and so remove the humidity that enters upon opening the lid.

Moreover, a routine based on a single bounce ATR unit (Figure 3.6) was also implemented in order to reduce the sample volume required for the analysis.

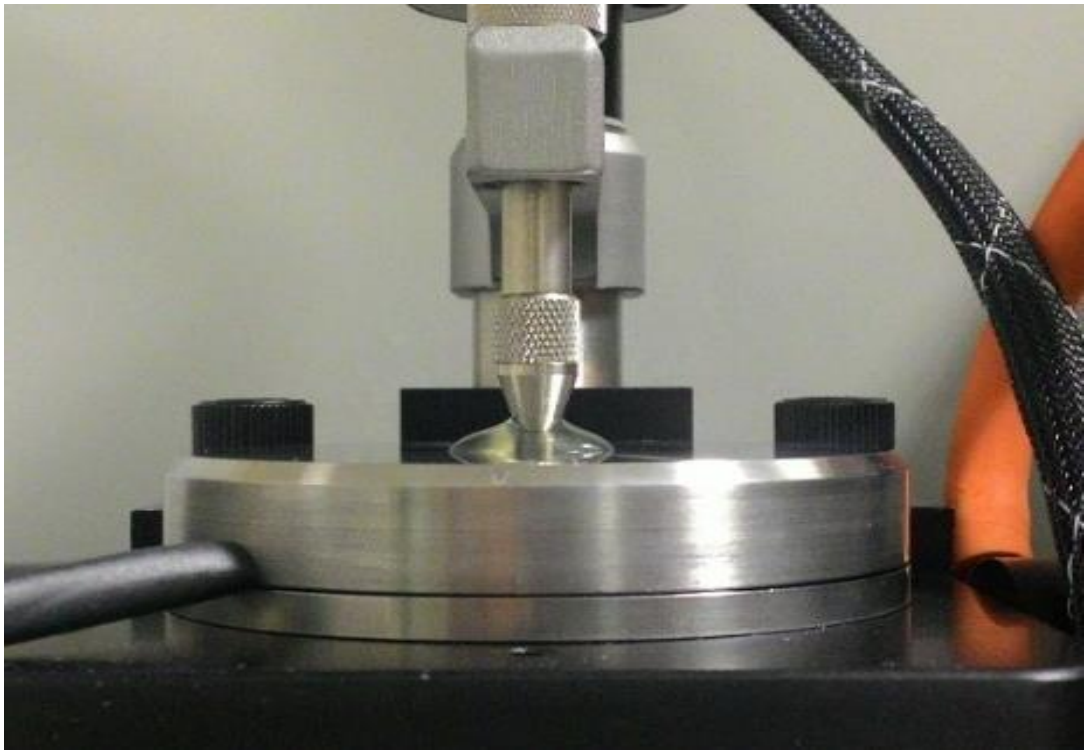

Figure 3.6. One bounce ZnSe ATR unit from Pike technologies.

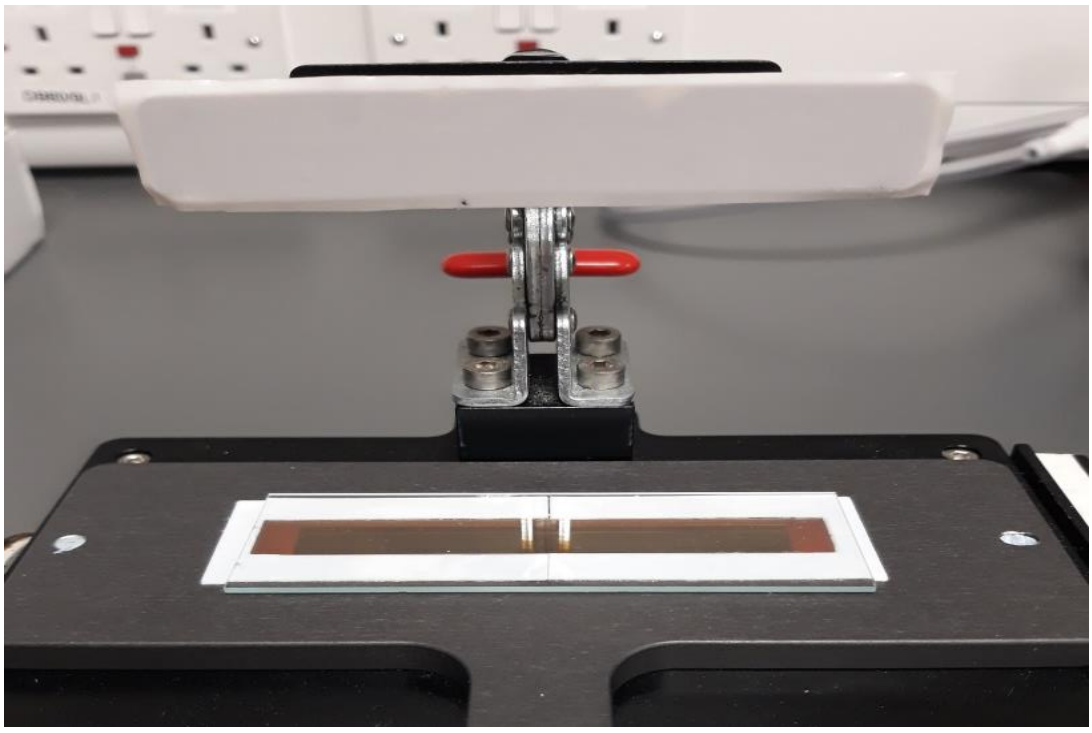

Figure 3.7. Flat surface six-bounce ZnSe ATR unit from SpecAc with home-made Teflon sample holder and cover glass.

Figure 3.8. Transmission cell from Specac and NaCl windows.

3.4.3 Experimental procedure

3.4.3.1 Instrument purge

One of the main drawbacks of IR spectroscopy is the interference of signals arising from the absorption of carbon dioxide and water vapour (Figure 3.9), with the latter being of great concern in the region of the amide I[45,46]. Although small signals can be arithmetically subtracted, variations of large amounts of vapour can induce artefacts in the spectrum as the signal does not cancel out due to lack of reproducibility between measurements. It is because of that that the instrument needs to be purged with N_2 prior to the start and over the course of the measurement to minimise vapour correction[46].

Figure 3.10 shows two series of spectra collected on two different days at regular time intervals (every 30 seconds for 9 min) with different flow rates that helped us identify when the level of humidity becomes steady enough to start recording. Once the minimum level of humidity is reached, the flow rate is decreased to about 5 $L.min^{-1}$ and kept throughout all the measurements.

When working on ATR mode, the inside of the instrument is practically insulated in its totality from the outer environment by means of the side extensions on the ATR accessory. But when working with transmission such insulation from the external

medium does not exist, as the lid must be open to change sample. For that reason and in order to re-establish the levels of humidity, it was necessary to wait a few seconds between measurements.

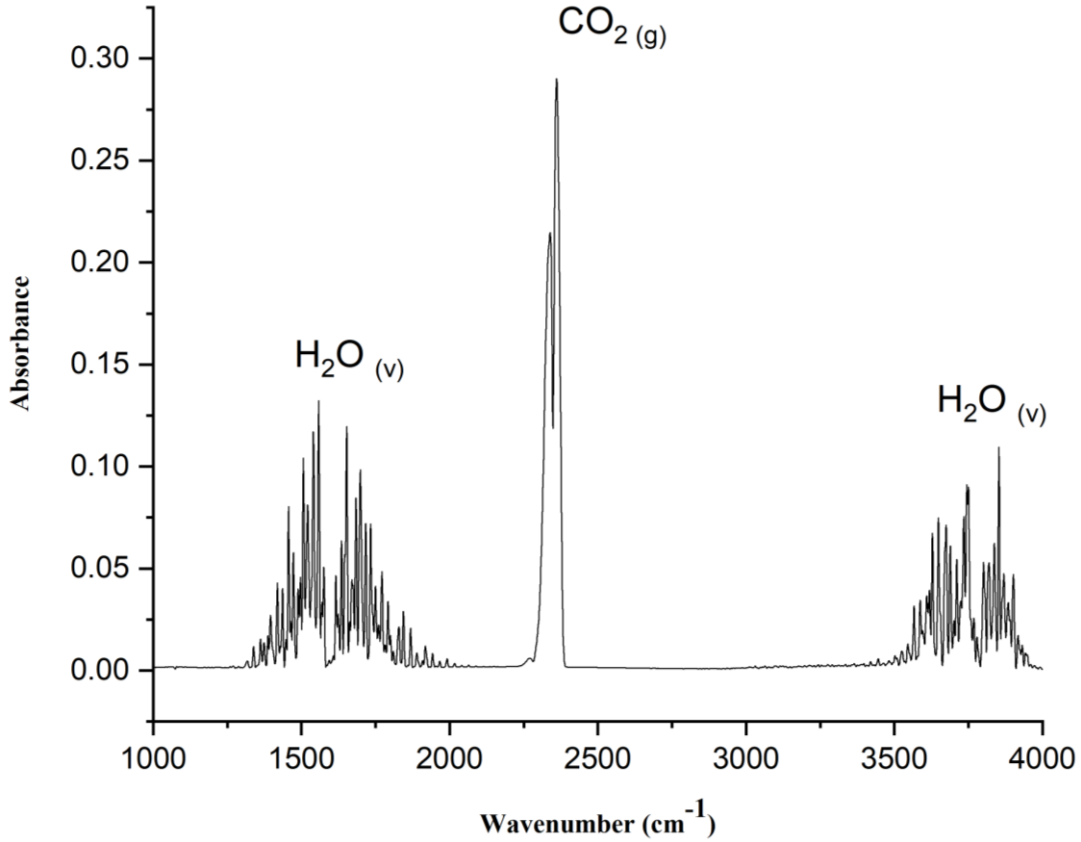

Figure 3.9. Spectra of water vapour collected with a PIKE MIRacle single bounce ATR unit.

In the light of these findings, for ATR we decided to purge the instrument in its totality (interferometer and sample chamber) before starting the data collection for about 10 min with ~30 L.min^{-1} N_2 flow and keep the purge to a minimum level during the measurements ~5 L.min^{-1} through the whole system. For transmission measurements, we decided to seal the interferometer after the initial purge in order to focus all the purging N_2 into the sample chamber and thus speed up the purge of the air coming in every time the lid is lifted.

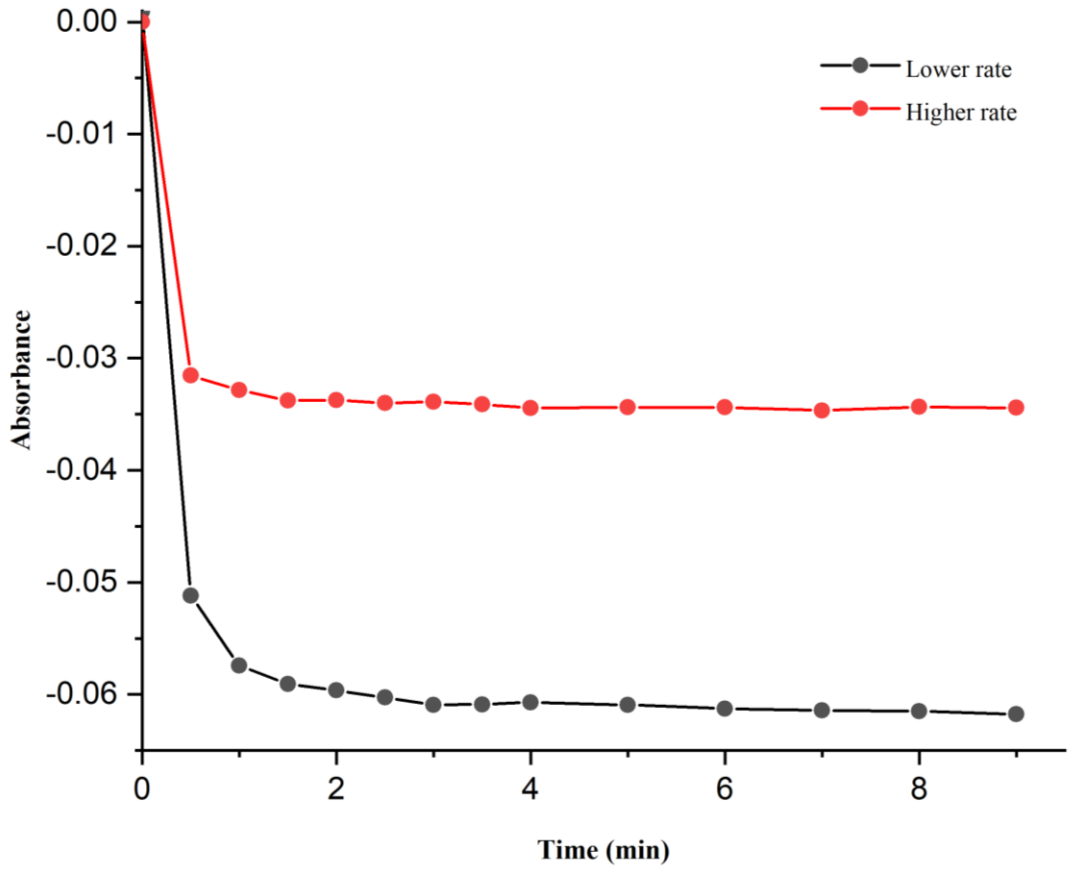

Figure 3.10. Water vapour trend over time at two different purge rates collected with a Specac 6 bounce ATR unit at 1559 cm^{-1}. The background was measured without purging the instrument and then the absorbance measured over time with the N_2 flow on. The spectra (6 accumulations each) were collected every 30 sec during the first 4 min and then every min until min 9.

3.4.3.2 Reference set data collection

3.4.3.2.1 Proteins in solid state

Between 1 and 5 mg of the protein powder was first grounded in a mortar and then mixed with separately grounded KBr to obtain a 1 to 10 % dilution. Next, the KBr and protein mixture was compressed by means of a hydraulic press (Specac pellet press) between 5 and 10 kpsi for about a min and a pellet was obtained. NaCl windows in a PIKE Technologies cell were used to hold the pellet in the beam.

The measurements were carried out with 4 cm^{-1} resolution, cosine apodization and accumulated over 100 scans in the range between 400 to 4000 cm^{-1}.

Since liquid water absorption was detected a water spectrum collected separately with CaF_2 windows was subtracted using the small water band in the region between 1800 and 2400 cm^{-1} as a reference.

3.4.3.2.2 Proteins in aqueous state

The solutions were prepared by dissolving the lyophilized powders in Milli-Q water in concentrations ranging from 40 to 80 $mg.ml^{-1}$**.** Next, the solutions were centrifuged for 5 min at 10 krpm to separate any insoluble residue. Finally, the supernatant was recovered and filtered with Teflon disk filters of 0.22 µm pore size to further eliminate any solids in suspension.

The spectra were acquired with a Specac transmission cell with CaF_2 windows and with no spacer in between making an estimated 1 µm pathlength. About 40 µl of sample were placed on one of the windows and the other was slid over it, making sure no air bubbles got trapped in the process. The spectra were collected in the range from 400 to 5000 cm^{-1} with 4 cm^{-1} resolution and cosine apodization; and averaged over 1000 scans. A water spectrum was measured before each protein in the set for subsequent subtraction (details explained in data processing Section-3.3.4). A single vapour spectrum was also collected by firstly purging the instrument for the background collection and then stopping the flow for the data acquisition. Two different approaches were taken to train the map: normalizing the spectra in terms of intensity by the interval method (values between 0 and 1) and converting the spectra into molar absorptivity (extinction coefficient ξ). The later required determining the accurate concentration of all the samples which was done by UV-Vis spectroscopy in accordance to Beer-Lambert law (Equation 1.1). The instrument used was a Jasco UV-Vis 660 equipped with a Xe and Deuterium lamp, a prism monochromator and a phototube detector. The spectra were collected in the range between 200 and 350 nm with a 3 mm quartz cuvette, 1 nm resolution, 0.1 data interval, 200 nm/min, medium response and a single accumulation. A water spectrum was collected before each sample for subsequent subtraction. The extinction coefficients used to work out the concentrations were taken from Uniprot.org.

The pathlength was determined by comparing the absorbance of the small band at ~2120 cm^{-1} which is only due to water with the value of absorbance of that band in a water spectrum previously measured with a 100 µm thick Teflon spacer**.**

3.4.3.3 Standard Operational Procedure (SOP) for IR-ATR protein collection

Although transmission measurements are possible as seen in the previous section, we found artefacts such as baseline fringes -likely derived from inconsistent assembling of the windows- that made the subtraction and reading of the path-length extremely difficult. Moreover, the changeable vapour levels and optical path between water and sample measurements sometimes made the factors required to correct for them extremely large, meaning additional spectral noise. All the mentioned reasons along with the time required for cleaning and assembling the cell pushed us into exploring the possibilities of ATR spectroscopy and leaving transmission experiments for reference proteins only.

Because of the larger optical path, the 6 bounce ATR unit was preferentially used for the lower concentrations measured (about 2 $mg.ml^{-1}$). A volume of 0.5 ml was necessary to fill the well created with the Teflon mask on the 6 bounce ATR accessory. Several drops were evenly spread across the crystal surface and then a microscope slide was slid over carefully to prevent air bubbles. The water was measured always first to avoid cross contamination and wiped gently with a tissue avoiding too much contact with the plate. The spectra were measured at room temperature with 4 cm^{-1} resolution, cosine apodization and the TGS detector. It was found that 600 scans are enough for low concentrated samples to meet the spectral quality SOM requires.

With the single bounce ATR unit, much less volume was used (50-80 µl). The samples were placed in contact with the tip and the crystal as showed in Figure 3.6. The water was always measured first so that no significant manipulation of the accessory was required and dried with a wipe tissue avoiding as much contact with the accessory as possible. It was found that for high concentrations (~50 $mg.ml^{-1}$) 250 accumulations were enough to obtain a high-quality spectrum whereas for smaller concentrations (~1 $mg.ml^{-1}$) it was necessary to accumulate about 10000 scans and thus keep adding sample over time and surround the plate with water continuously to saturate the nearby area in order to slow down evaporation of water. Although the lowest concentration measured was 1 $mg.ml^{-1}$, we believe the effective concentration could have risen significantly over the course of the measurement because of solvent evaporation.

3.4.4 Data processing

Water IR spectra show three prominent bands in the mid infrared: O-H anti and symmetric stretch at ~3400 and ~3280 respectively cm^{-1}, O-H bend and libration combination mode at ~2120 cm^{-1} and O-H bend at ~1643 cm^{-1} (Figure 3.11)**.** As stated previously, the characterization of protein secondary structure is usually based on the amide I band (C=O stretch between 1600 and 1700 cm^{-1}). Unfortunately, water absorbs strongly in that region (O-H bend vibration). Recalling Beer-Lambert law (Equation 1.1), for an aqueous protein sample in water the total absorbance is

$$A_{atr}^{pw} = \left(\varepsilon^{p} \cdot C^{p} + \varepsilon^{w} \cdot C^{w}\right) \cdot d_{eff}^{pw} \tag{3.8}$$

where *pw* stands for protein in water, *p* means protein and *w* stands for water. For just the water the absorption is

$$A_{atr}^{w} = \varepsilon^{w} \cdot C^{w} \cdot d_{eff}^{w} \tag{3.9}$$

and the protein on its own

$$A_{atr}^{p} = A_{atr}^{pw} - A_{atr}^{w} = \left(\varepsilon^{pw} \cdot c^{pw} \cdot d_{eff}^{pw} - \varepsilon^{w} c^{w} \cdot d_{eff}^{w}\right) \tag{3.10}$$

One way around the strong absorption from water is to use D_2O instead since the later mentioned vibration frequency is shifted downwards due to the larger weight of deuterium compared to that of hydrogen[90,103]. Although that works out for many protein studies, it does not for pharmaceutical related ones which require conditions close to physiological. Thus, it is necessary a protocol for accurate subtraction of water. We found that the water subtraction was possible as suggested already in the literature[46] but since the water vibrations are dependent on the chemical nature and composition of the blank, it was necessary to always match blank and sample in composition for a successful subtraction. Moreover, although all measurements were carried out with nominally the same volume of liquid and with the same instrument settings, slight differences in the magnitude of the water peak between sample and blank occurred. Because of this, we had to scale the blank with an arbitrary factor

prior to the subtraction. Due to the extremely high concentration of water compared to the protein, the Amide I band region is dominated by the absorption from water and the lower the concentration of the protein the more compromised the subtraction will be (Figure 3.12). In order to avoid biasing the choice for the scaling factor we designed mathematical criterions and performed the subtraction computationally. In agreement with the literature[46], we used the small combination band from water as a reference for the subtraction as no protein absorbs in that region and applied a series of factors until the band cancels out fully yielding a flat baseline.

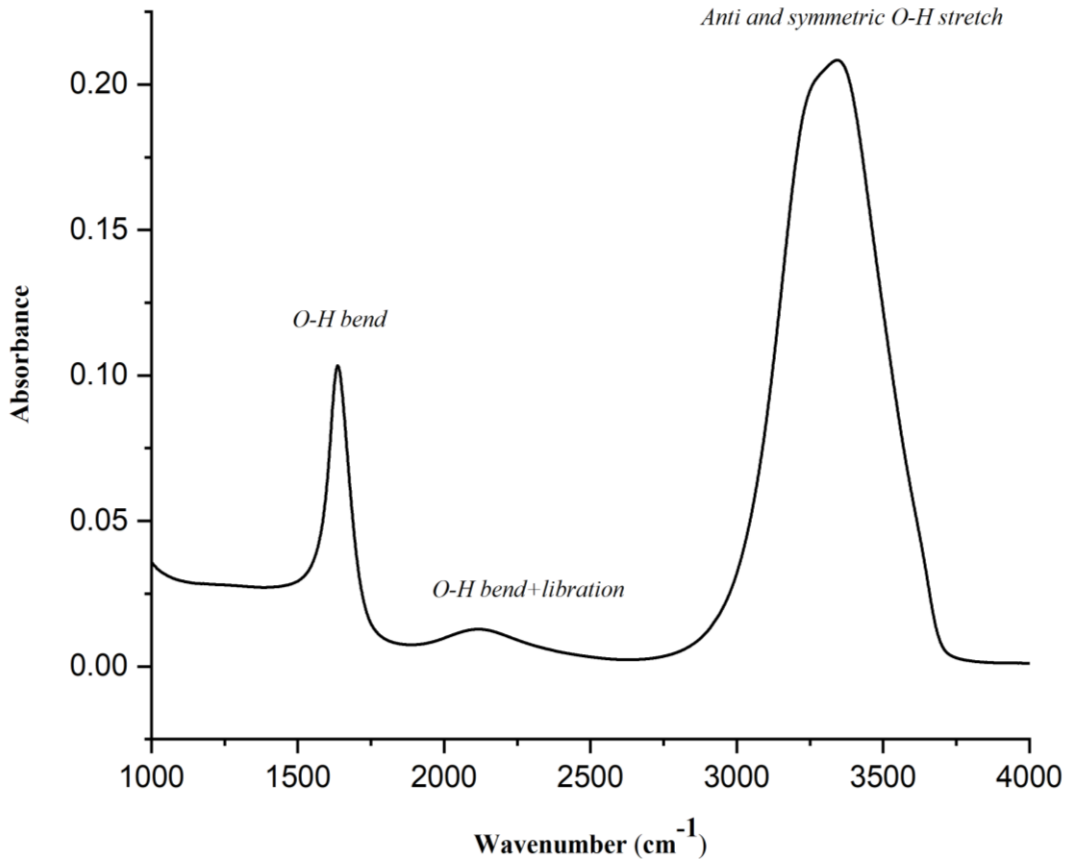

Figure 3.11. Water spectrum measured with a ZnSe 1-bounce ATR PIKE MIRAcle unit.

Furthermore, despite the instrument being continuously purged as discussed above, it was impossible to prevent the presence of some water vapour in the spectra. Because of fluctuations of humidity in the air, this rarely cancelled out when subtracting the blank from the sample (Figure 3.13), so a water vapour spectrum had to be subtracted separately too. Although the range used in the literature for the vapour subtraction is the same as the one used for water subtraction (1750-2000 cm^{-1})[46], we decided to use the signal in the region between 3800 and 3900 cm^{-1} instead.

The baseline corrections therefore consisted of the subtraction of both buffer and water vapour multiplied by a factor that compensates the differences in magnitude.

Due mainly to Mie scattering, the baseline is rarely a straight line but a curve instead[46] which makes it even harder to decide what factor is the correct one in the case of water subtraction. When it comes to vapour, noise is the limiting variable.

A MATLAB[79] routine (Appendices-A.3) was implemented to automate the iterative water and vapour subtractions. The code for IR data processing was written in and consists of integrated functions for zeroing, scaling, trimming and water and vapour subtraction. The way the code operates is as follows: the code is fed with two vectors of values evenly spread about 1 for water e.g., v = (0.5:0.1:1.5) and 0 for vapour e.g., u = (-0.5:0.1:0.5). Then, the code screens the different elements in the vectors and performs a subtraction with each of them. The results are fitted with a spline-based curve from 1850 to 2600 for water and 3800 to 3900 cm^{-1} for vapour as shown in figures 3.14 and 3.15 and their respective residual squared sums computed. The optimal factor for the subtraction is then chosen will be the one that minimizes the residuals (Equation 3.11).

$$R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.11}$$

Where n is the number of elements in the considered range, y_i is the i experimental value and $\hat{y}_i$ is the fit curve one.

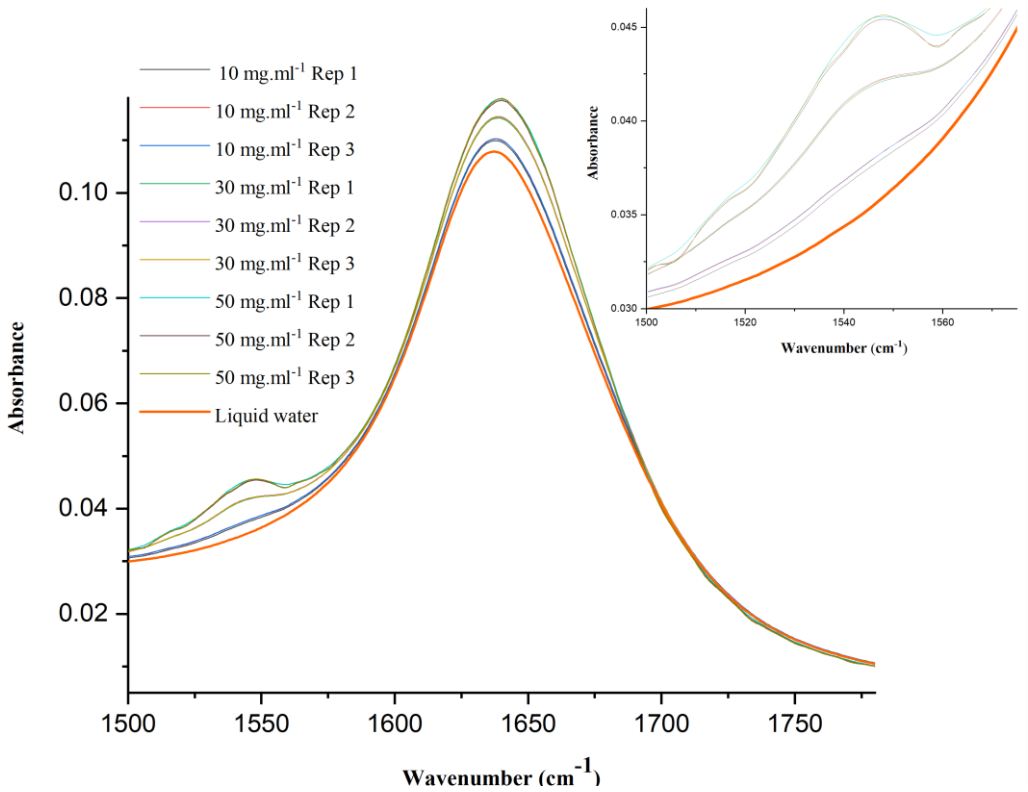

Figure 3.12. Different concentrations of Lysozyme in water with a ZnSe 1-bounce ATR PIKE MIRAcle unit.

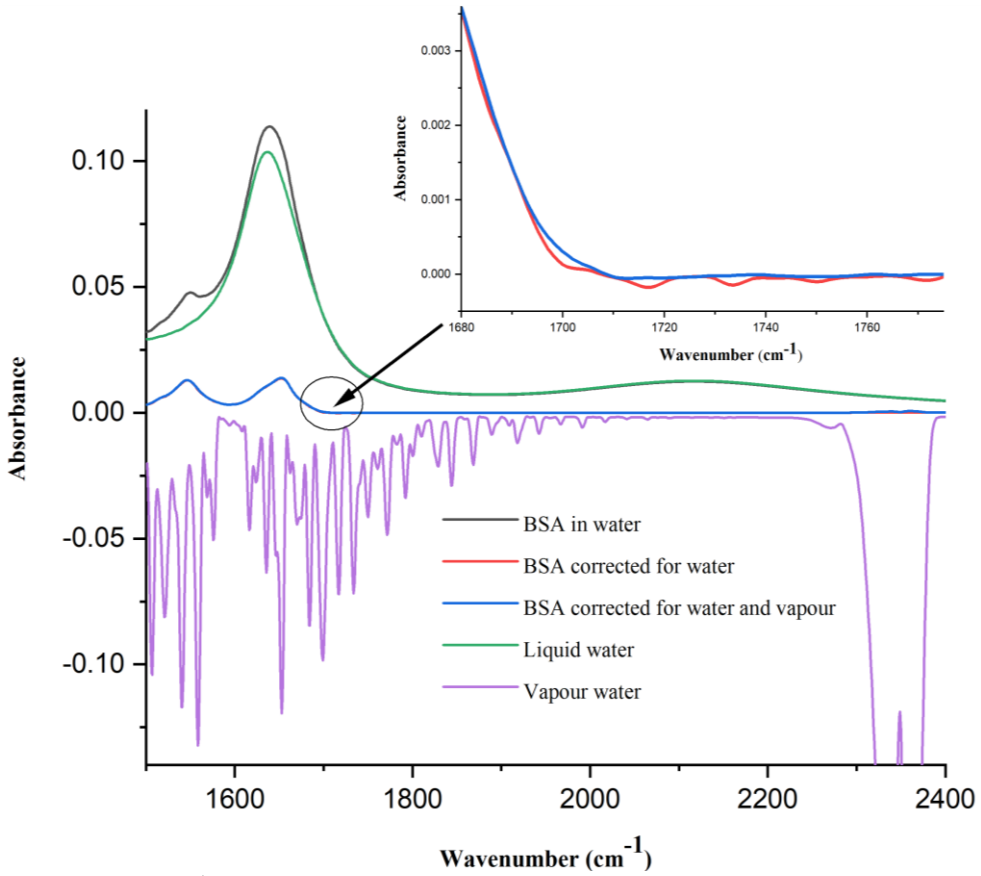

Figure 3.13. 50 mg.ml^{-1} BSA in water collected with a ZnSe 1-bounce ATR PIKE MIRAcle unit.

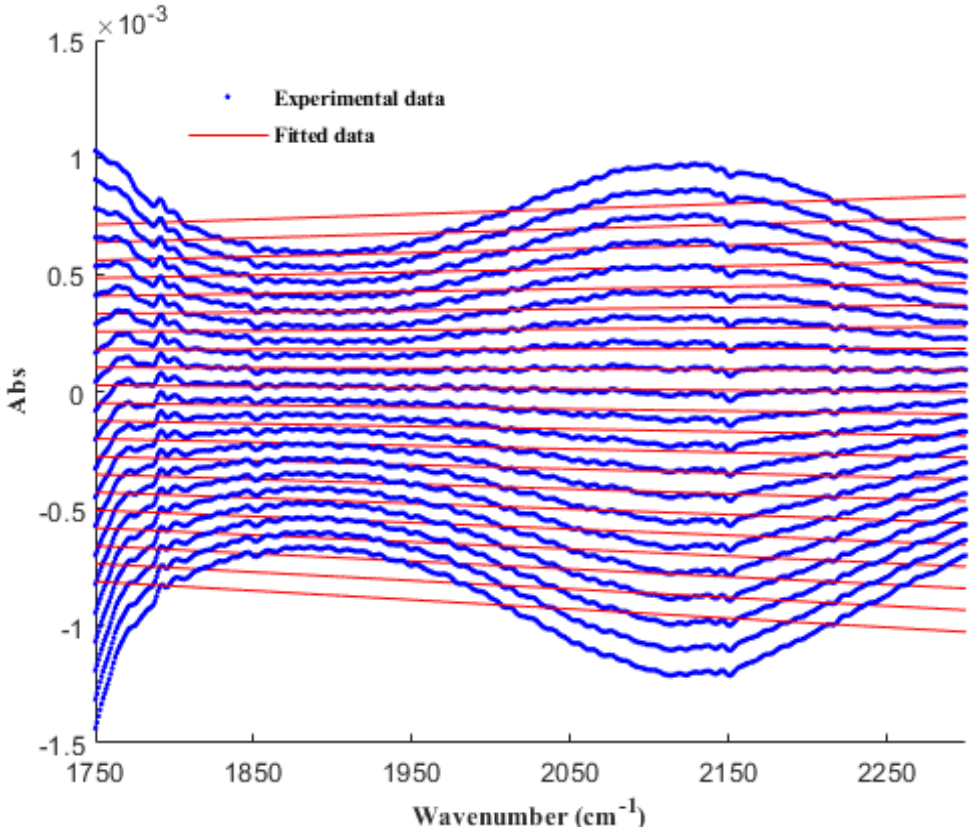

Figure 3.14. Subtraction of liquid water from a 2 mg.ml^{-1} BSA solution with different factors iteratively. The spectra were collected with a ZnSe 6-bounce ATR Specac unit and the subtraction performed through a MATLAB routine.

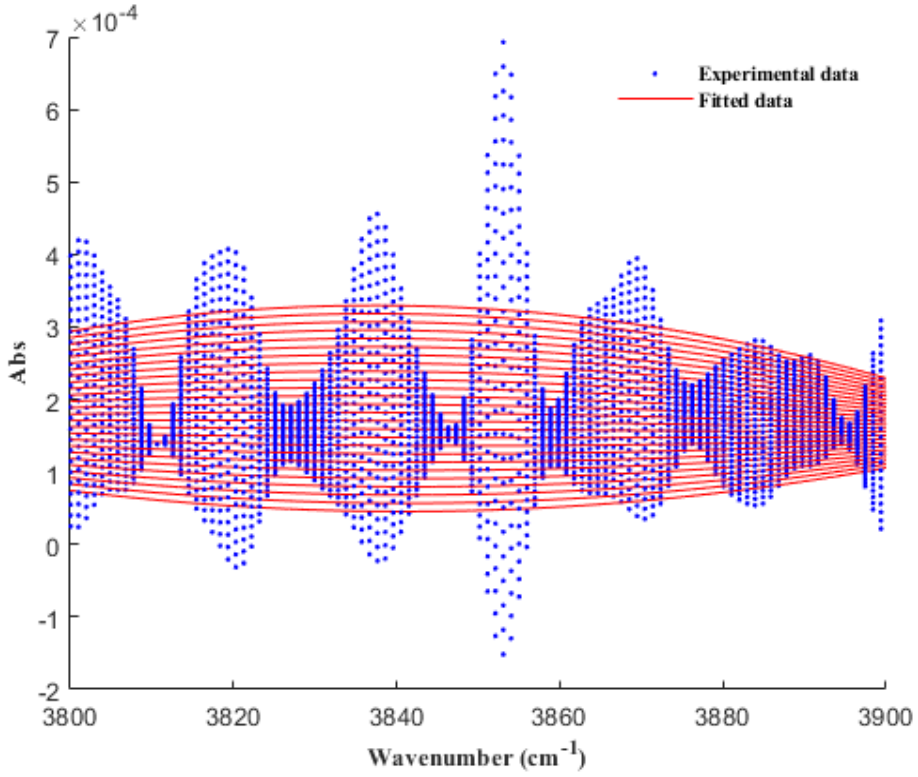

Figure 3.15. Iterative subtraction of water vapour from a 2 mg.ml^{-1} BSA protein solution. The spectrum was collected with a ZnSe 6-bounce ATR Specac unit and the subtraction performed through a MATLAB routine.

3.5 RESULTS

3.5.1 IR reference set in solid state

The SS analysis below focuses on the Amide I band which is accepted to represent the fractions of the different secondary structures present in the protein as stated previously. Although the presence of vapour was nearly neglectable, significant presence of liquid water was found in the solid samples (Figure 3.16). Water and vapour subtractions were performed on the 31 proteins in the set as described in methods. Because of variations in the water peak position, we decided to attempt the subtractions manually using the O-H stretch at 3400 cm^{-1} as a reference rather than the small combination band at 2120 cm^{-1}. The region used for the subtraction of vapour was 3800-3900 cm^{-1}. The subtractions were performed in spectra analysis (Jasco spectra manager version 2) subsequently, firstly the water and secondly the vapour.

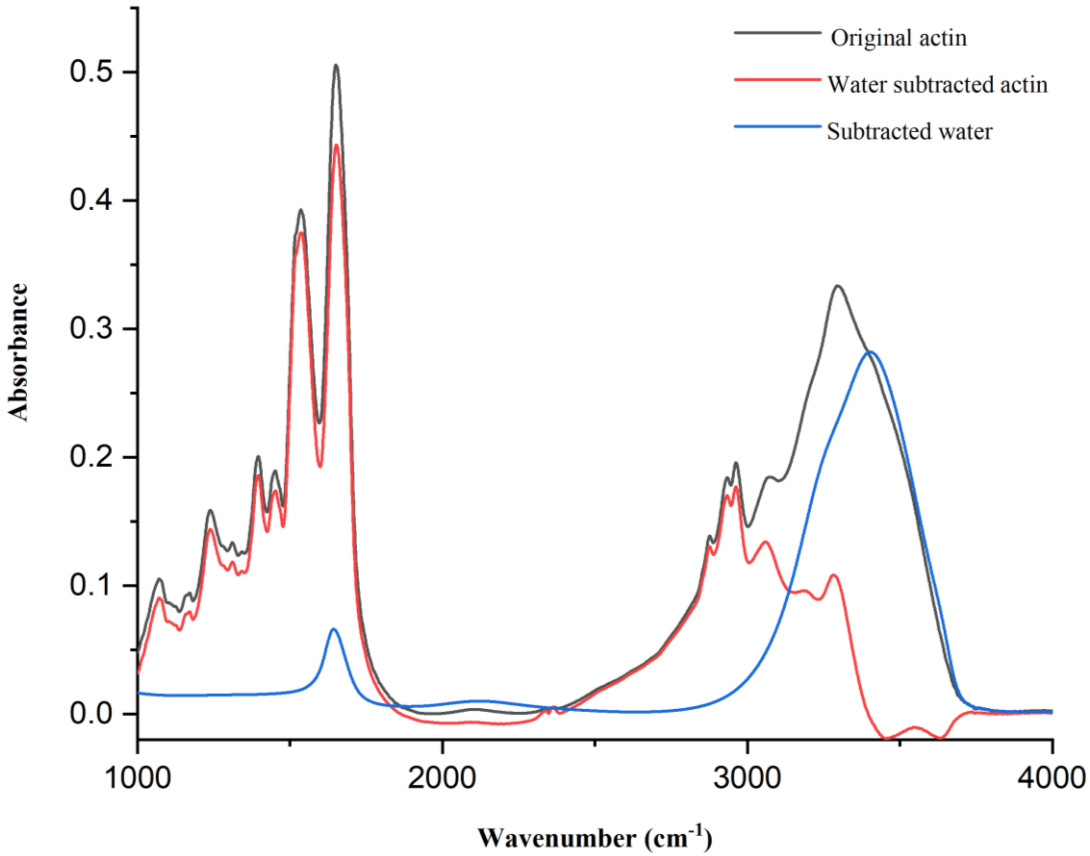

Figure 3.16. Solid state spectra of actin, water subtracted actin and subtracted water.

Because of the difference in thickness between pellets, the spectra were not comparable in magnitude which required scaling before SOM. The spectra were normalized in terms of intensity by the 'interval method (Equation 3.12)' (Figure 3.17) and put through SOM for a leave one out validation (not shown).

$$Normalized = \frac{A_i - A_{min}}{Amin_{max}} \tag{3.12}$$

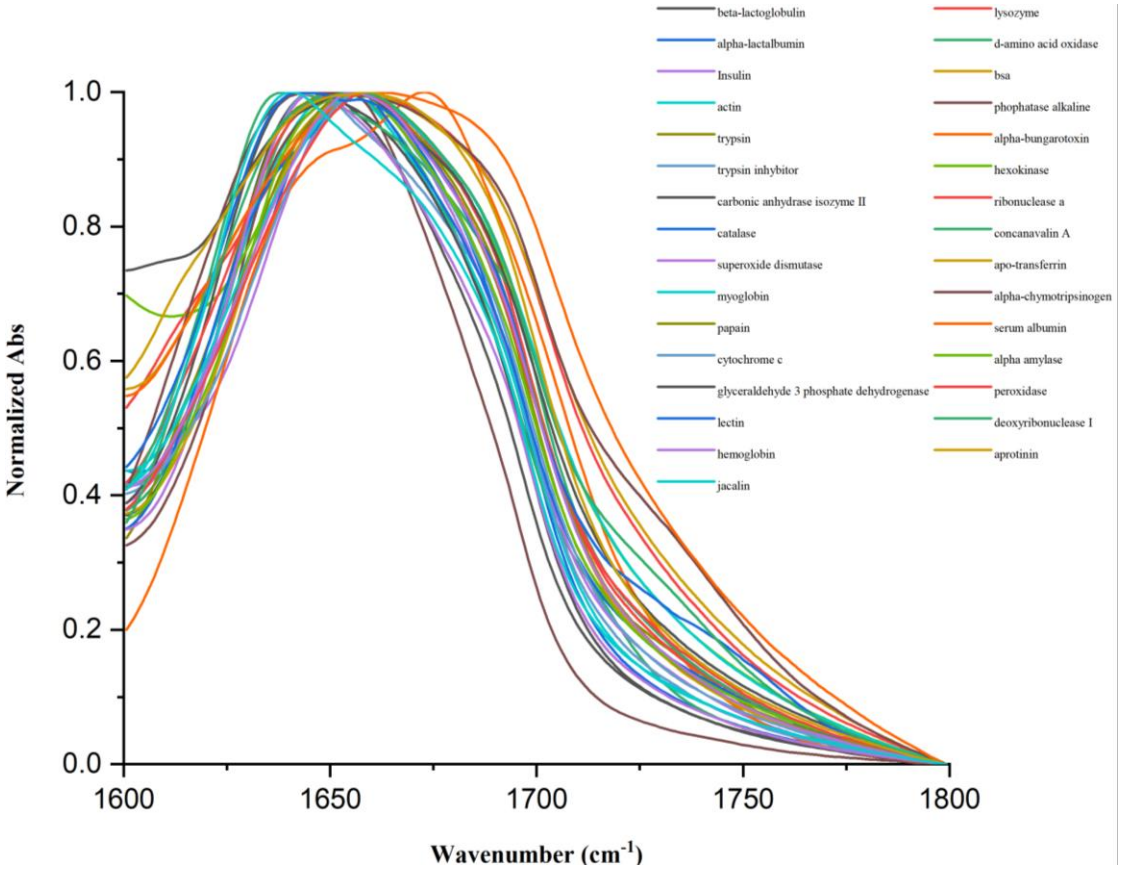

Figure 3.17. Transmission IR spectra of 31 proteins in solid state plotted in the region of the amide I band. The spectra were normalized by the interval method.

A preliminary exploratory analysis was performed on the reference set to determine correlations between secondary structure and ranges within the amide I band. For this purpose, we plotted the wavenumber corresponding to the maxima of the peaks against their helical and sheet contents (Figures 3.18 and 3.19). These figures show that there is a correlation between spectral range and secondary structure as claimed in the literature. Figure 3.18 shows how the max of the peak shifts down to 1640 cm^{-1} with increasing sheet structure and Figure 3.19 how the max shifts up to 1657 cm^{-1}

with increasing helical content consistent with the literature. Further analysis can be found in Appendices-B.1.

Furthermore, we also plotted helical *vs* sheet contents (Figure 3.20) to visualize if there is any obvious correlation between them. As it can be seen in Figure 3.20, there is a very strong inverse correlation between both structures which explains the trends about the edges in Figures 3.18 and 3.19. It can be noticed in them that when the content of helix is the smallest the sheet is the greatest and vice-verse.

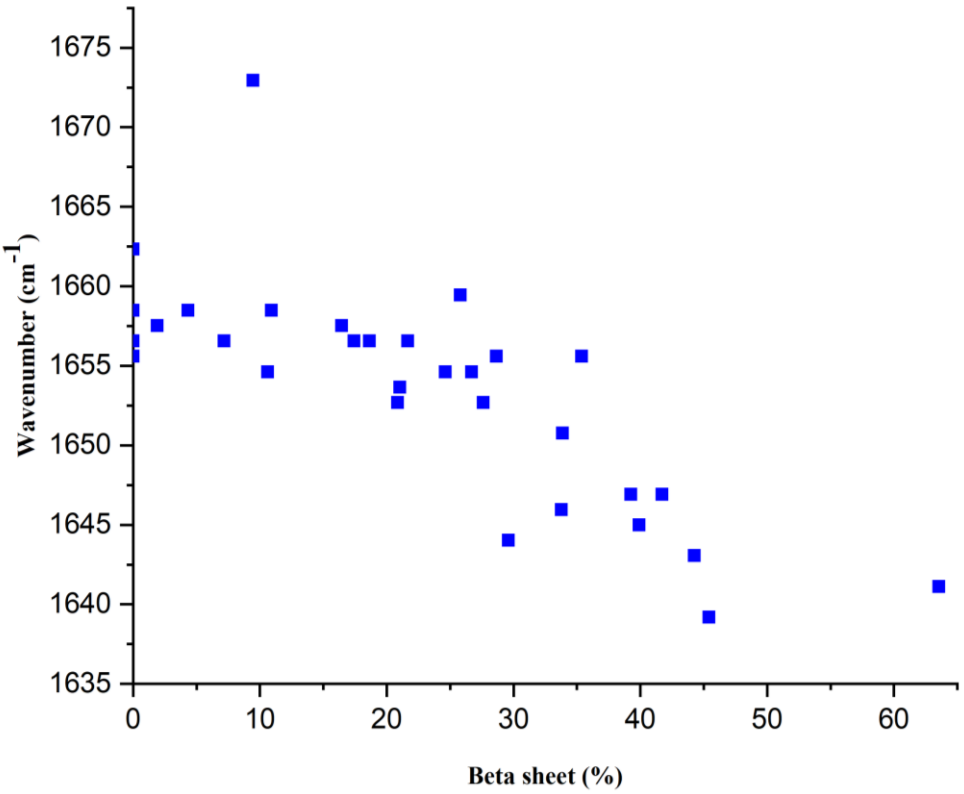

Figure 3.18. Wavenumbers corresponding to the peaks' max vs their beta sheet contents.

Although these results looked encouraging and in agreement with the literature, limitations exist in the treatment of the data that could partially explain the outcomes of the SOM validation (not included). These main limitations are the challenges of subtraction of water and the scaling of the spectra performed. The latter relies on the fact that normalizing works on the assumption that the extinction coefficients between proteins is practically the same and that only shape and position matter, something we found is not necessarily true for all proteins (next section). Also, the choice of normalizing in terms of intensity instead of area could be a potential source of error.

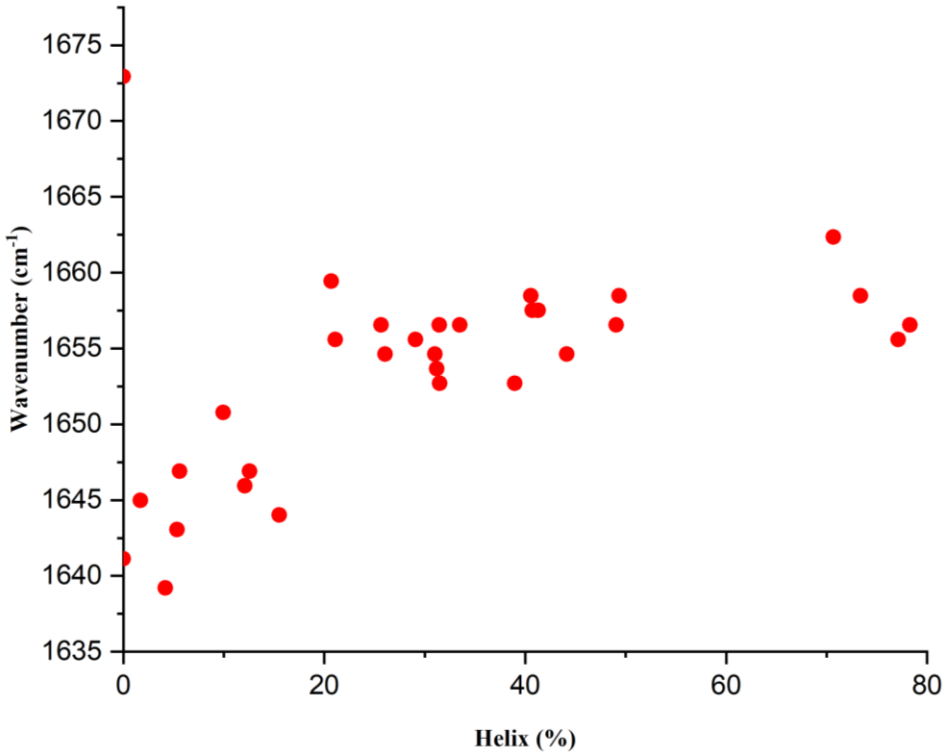

Figure 3.19. Wavenumbers corresponding to the peaks' max vs their helical contents.

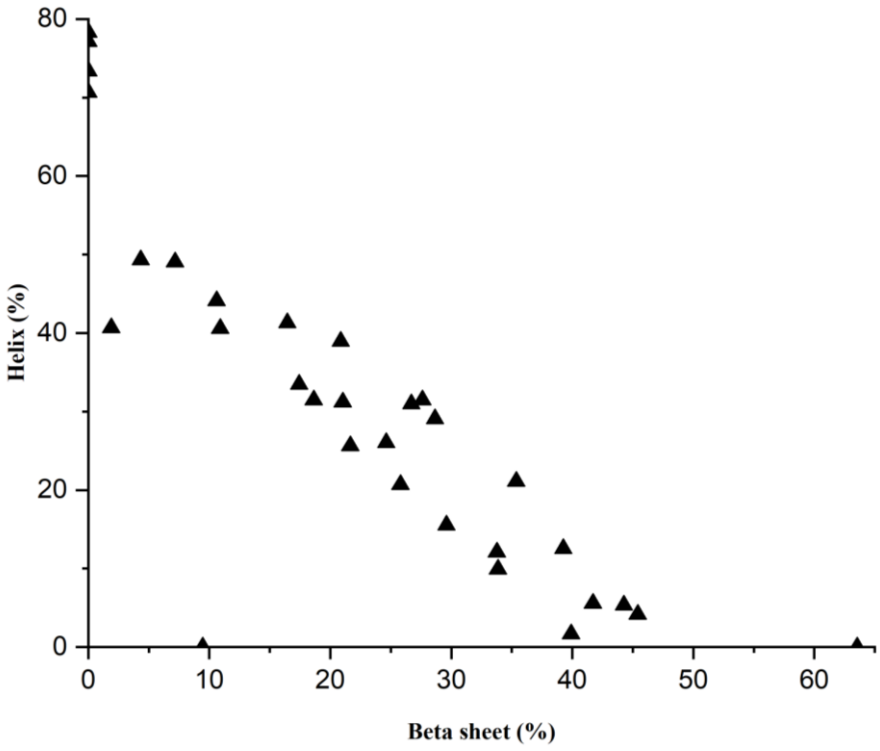

Figure 3.20. Helical *vs* sheet content of the transmission IR ref set in solid state (Figure 3.17) using the annotations from the server 2struc.cryst.bbk.ac.uk.

3.5.2 IR reference set in aqueous state

Although the main goal of this chapter was developing a method for characterization of proteins in solution, it was decided to firstly collect the reference set in solid state and so avoid solubility issues thinking it would be possible to predict aqueous state protein structure with it. However, only after collecting a fairly large number of them, it was noticed they do not compare to their corresponding aqueous spectra (Figure 3.21).

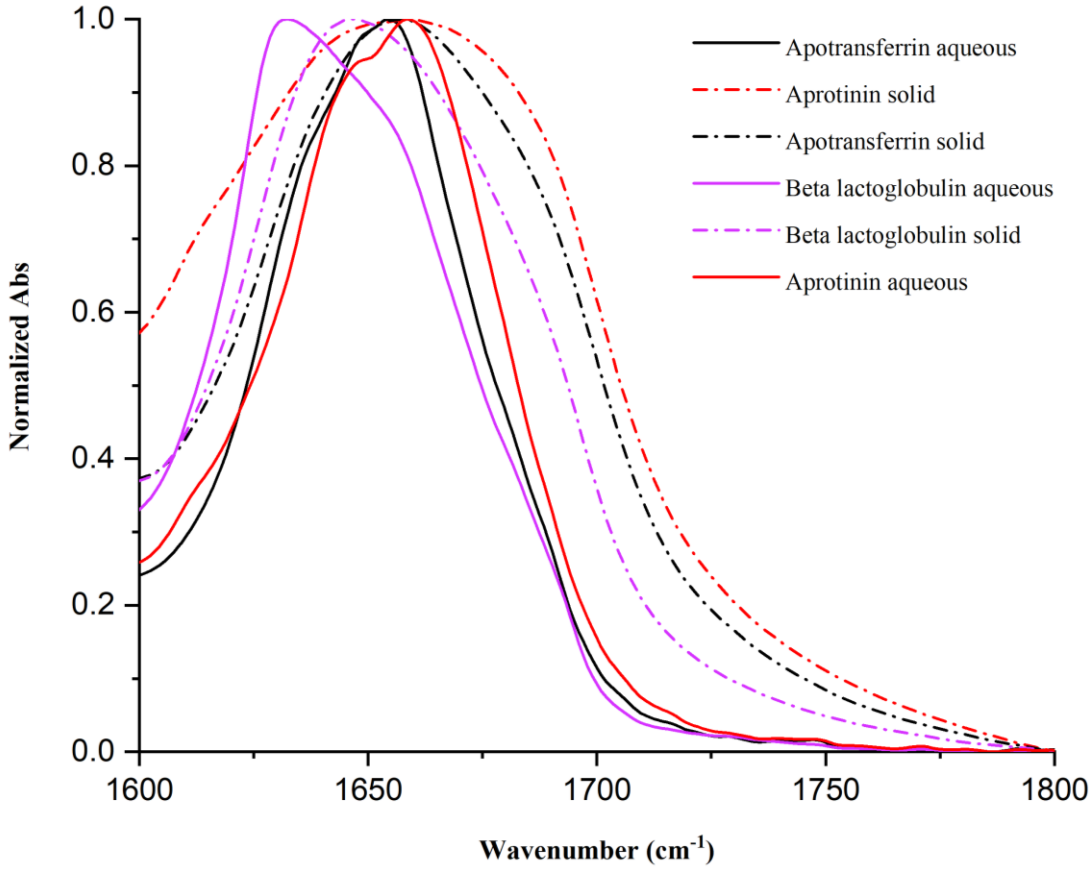

Figure 3.21. Comparison of aqueous and solid-state Apo-transferrin, Aprotinin and Beta lactoglobulin within the region of the amide I band. The spectra were collected with transmission and treated accordingly to what described in the methods section.

Although we had stocks of up to about 35 commercial proteins, we only succeeded in measuring 21 with a fairly good coverage of helix *vs* sheet structures (Figure 3.22).

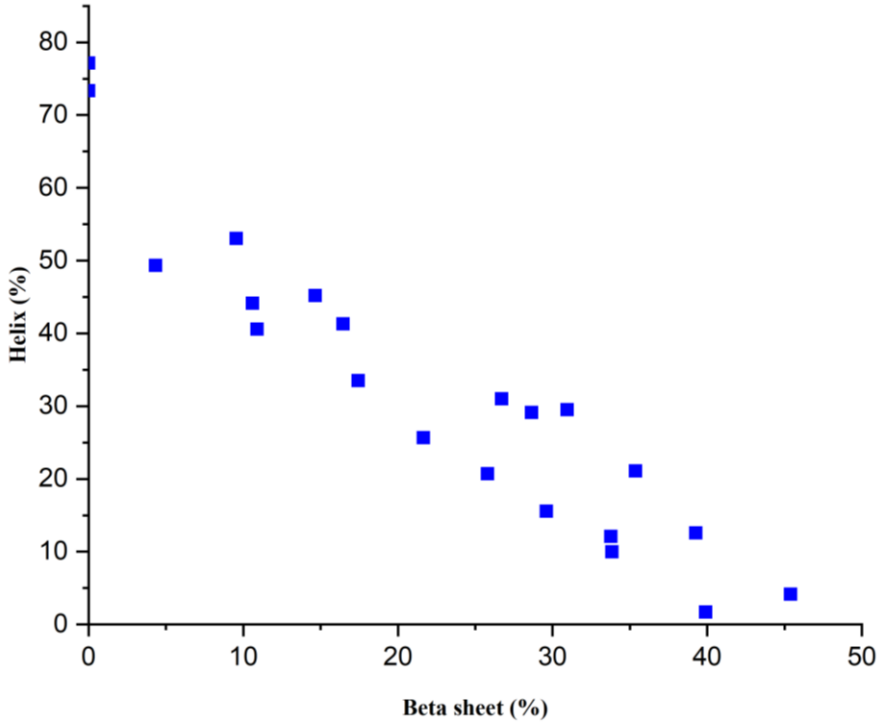

Figure 3.22. Coverage of SS by the transmission IR reference set in aqueous state (Figures 3.24, 3.25 and 3.27).

The baseline of the transmission spectra of some proteins required a baseline correction. The baselining was carried out by selecting points in the range from 1750 to 2700 cm^{-1} and interpolating them with linear splines. The edges of the spectra were defined by extrapolating the lines close to 1750 and 2700 cm^{-1} (Figure 3.23)

Besides the baseline correction, all the spectra were water and vapour subtracted as described in the methods section and converted to extinction coefficient in MRW (mean residue weight) (Figure 3.24) for SOM validation. Furthermore, another two validations of the ref set processed in a different manner were performed: normalized by the interval method (Figure 3.25) and deconvolved + normalized (Figure 3.27). The reason for the latter is that we found proteins (e.g., BSA, Aldolase) with similar overall shape when normalized (Figure 3.26) but significantly different SS annotations which affects their predictions. In order to solve this problem, the bands were subjected to Fourier-Self- Deconvolution (FSD) in Origin[80], hoping that resolving their intrinsic spectral features results in distinct overall shapes. The principles of FSD can be found in Appendices-B.2.

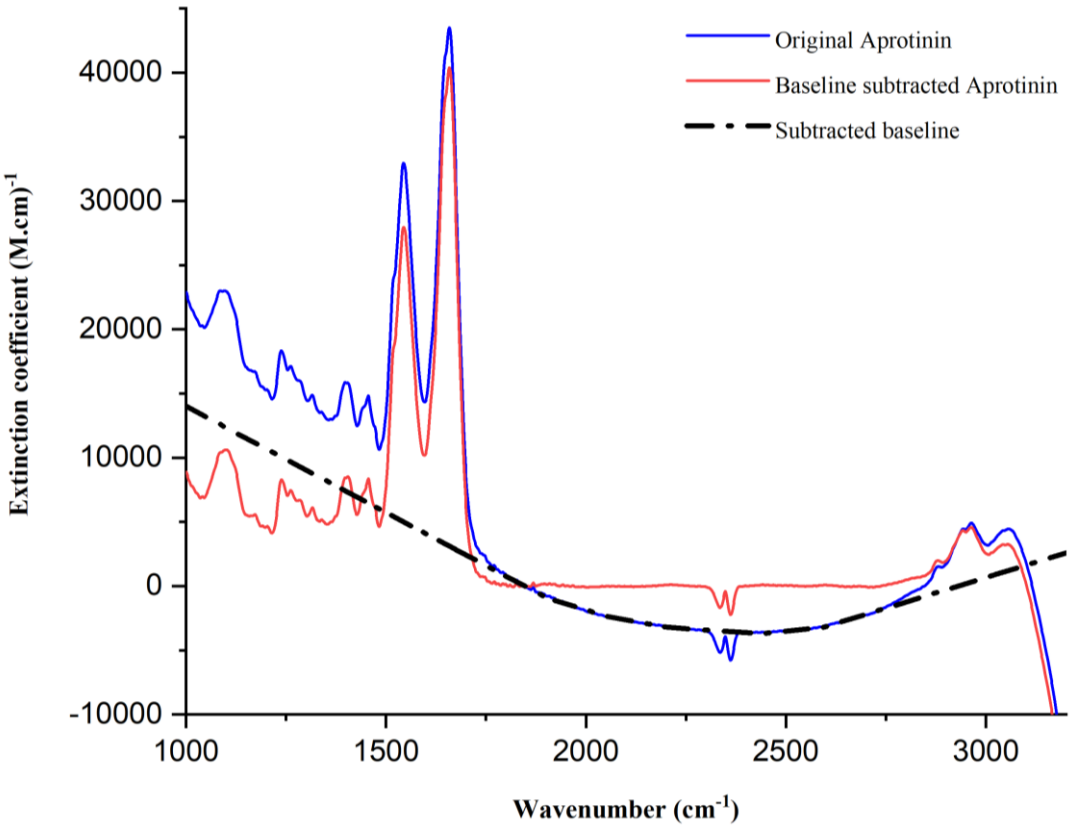

Figure 3.23. Baseline correction of transmission IR spectra of aprotinin in solution. Points between 1750 and 2700 cm^{-1} were interpolated by linear splines.

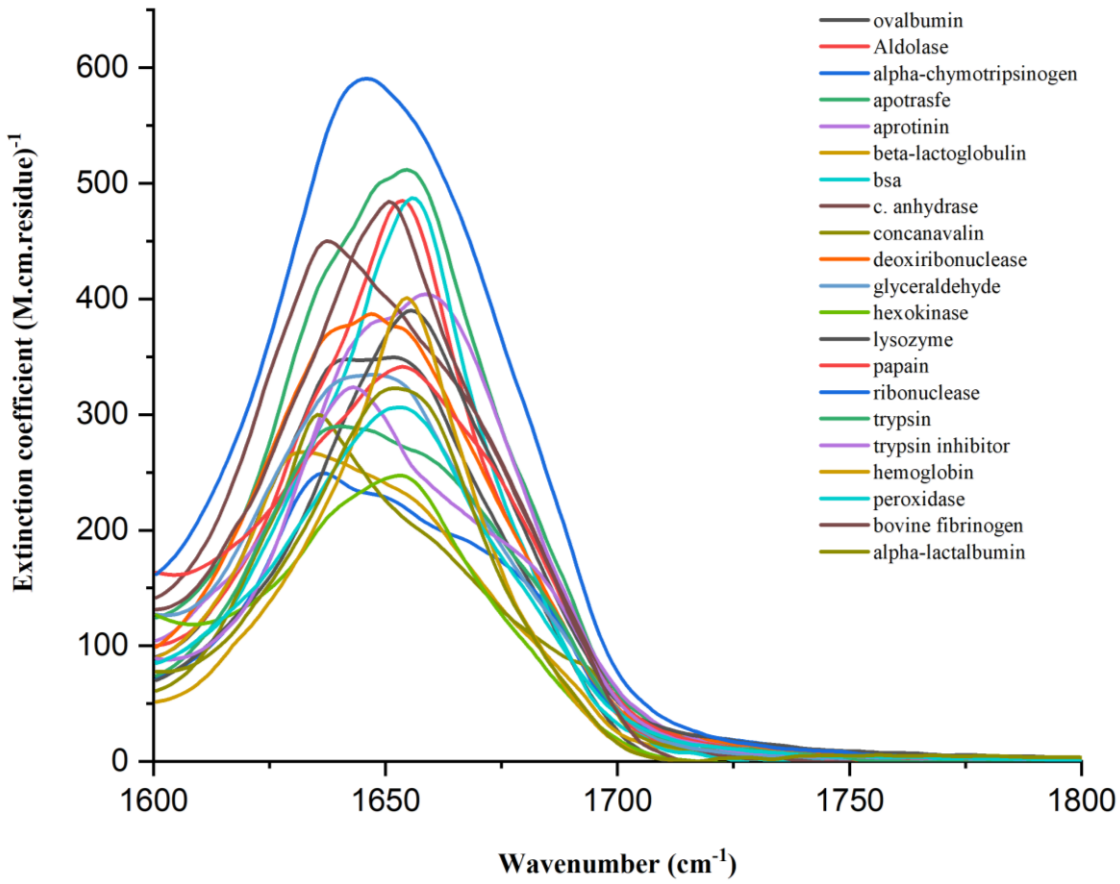

Figure 3.24. Transmission IR spectra of 21 proteins in aqueous state. The spectra were converted to extinction coefficient in MRW.

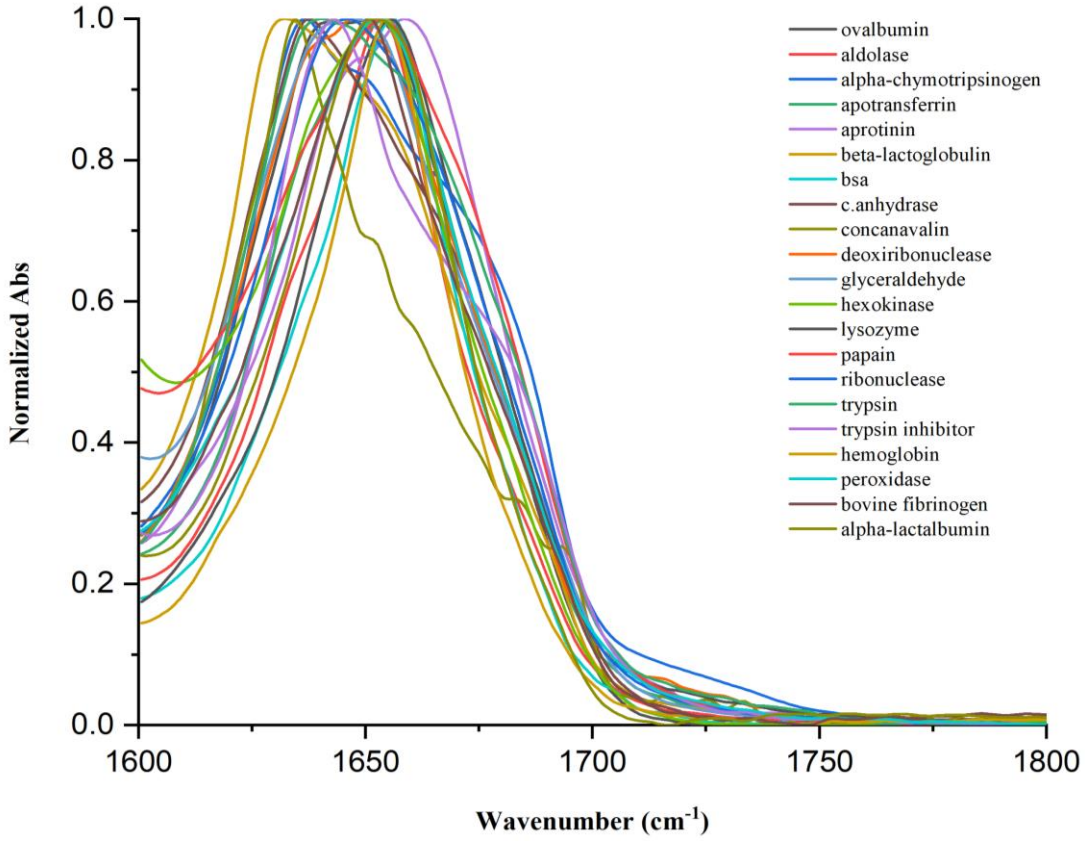

Figure 3.25. Transmission IR spectra of 21 proteins in aqueous state. The spectra were normalized by the interval method.

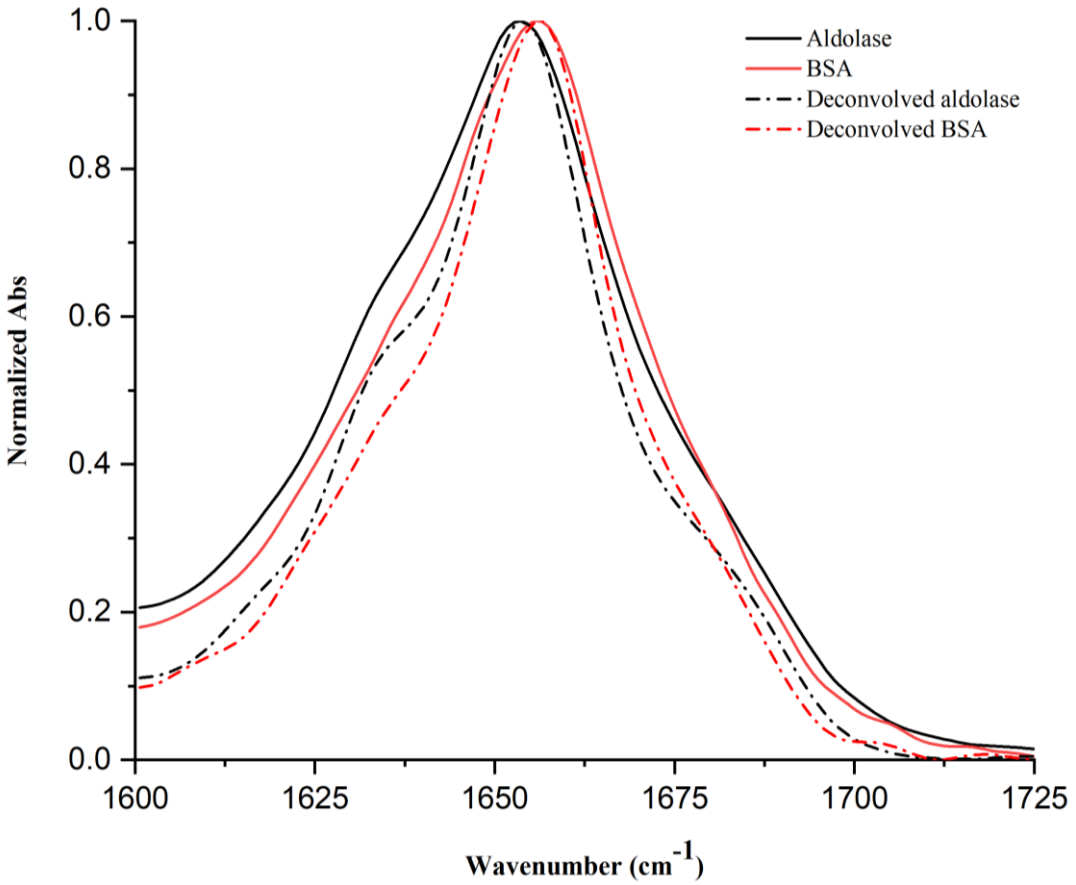

Figure 3.26. Comparison of transmission IR spectra of BSA and Aldolase in water, with and without deconvolution.

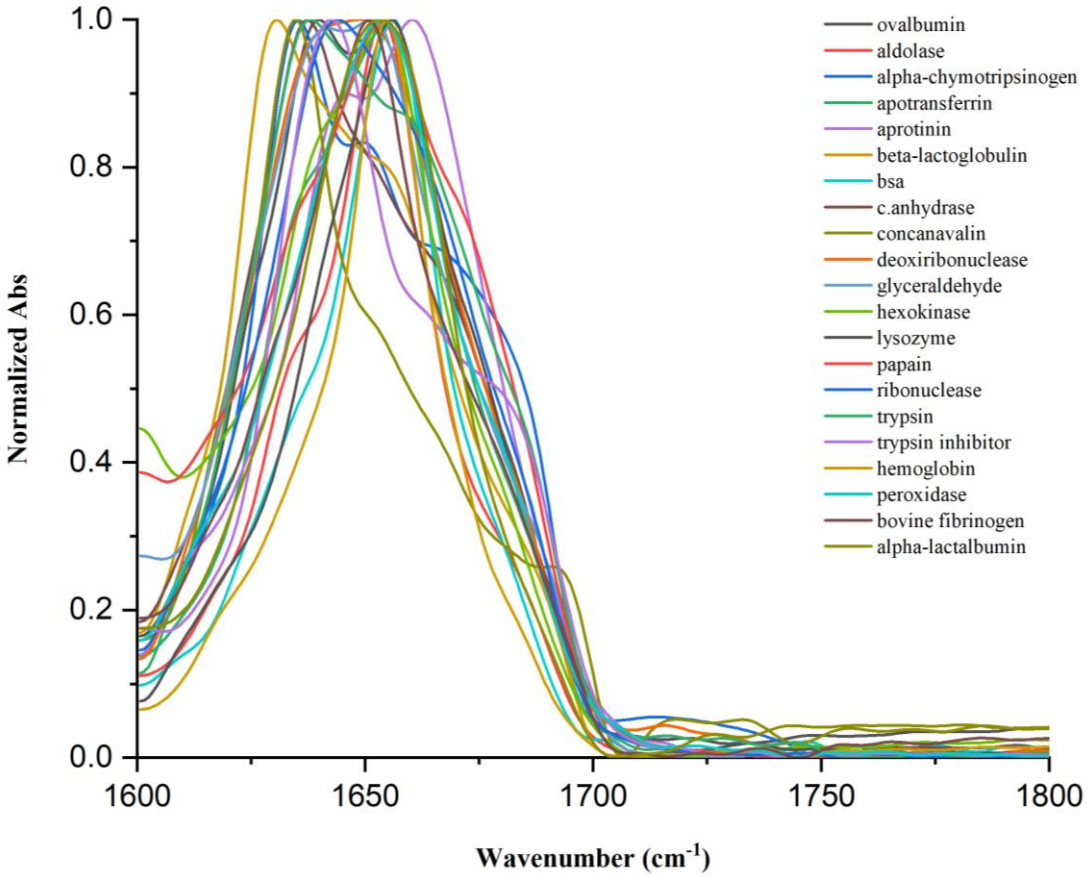

Figure 3.27. Normalized FSD transmission IR spectra of 21 proteins in solution. The deconvolution was performed in Origin with a gamma value of 10 and smoothing factor of 0.25.

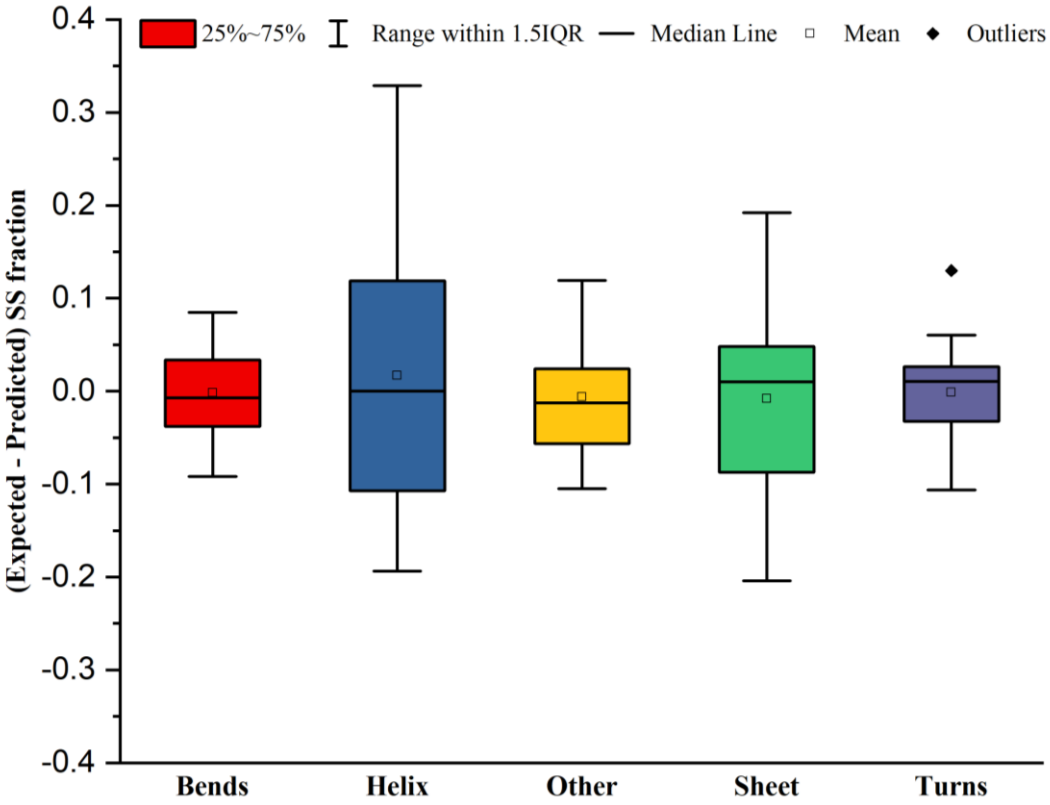

Figure 3.28. Results of the leave-one out validation of the aqueous IR 21-protein reference set expressed in MRW extinction coefficient (Figure 3.24). The validation was run with a 40x40 map and 40000 iterations.

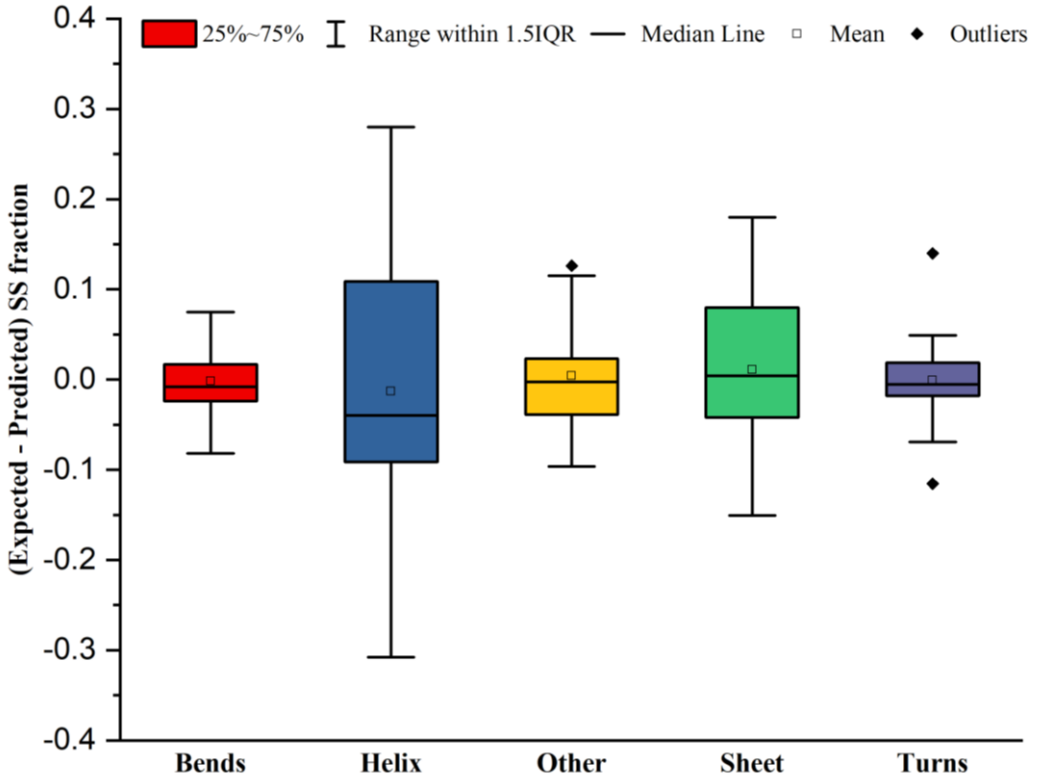

Figure 3.29. Results of the leave-one out validation of the normalized IR 21-protein reference set measured in water (Figure 3.25). The validation was run with a 40x40 map and 40000 iterations.

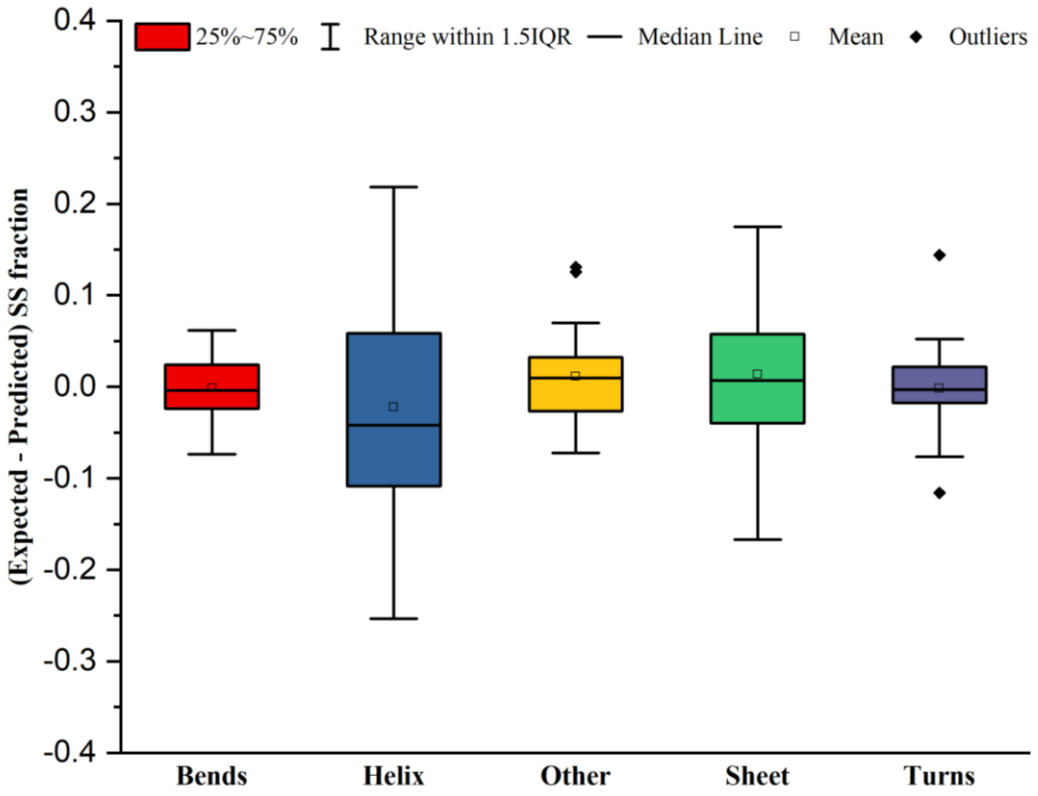

Figure 3.30. Results of the leave-one out validation of the deconvolved + normalized IR 21-protein reference set measured in water (Figure 3.27). The validation was run with a 40x40 map and 40000 iterations.

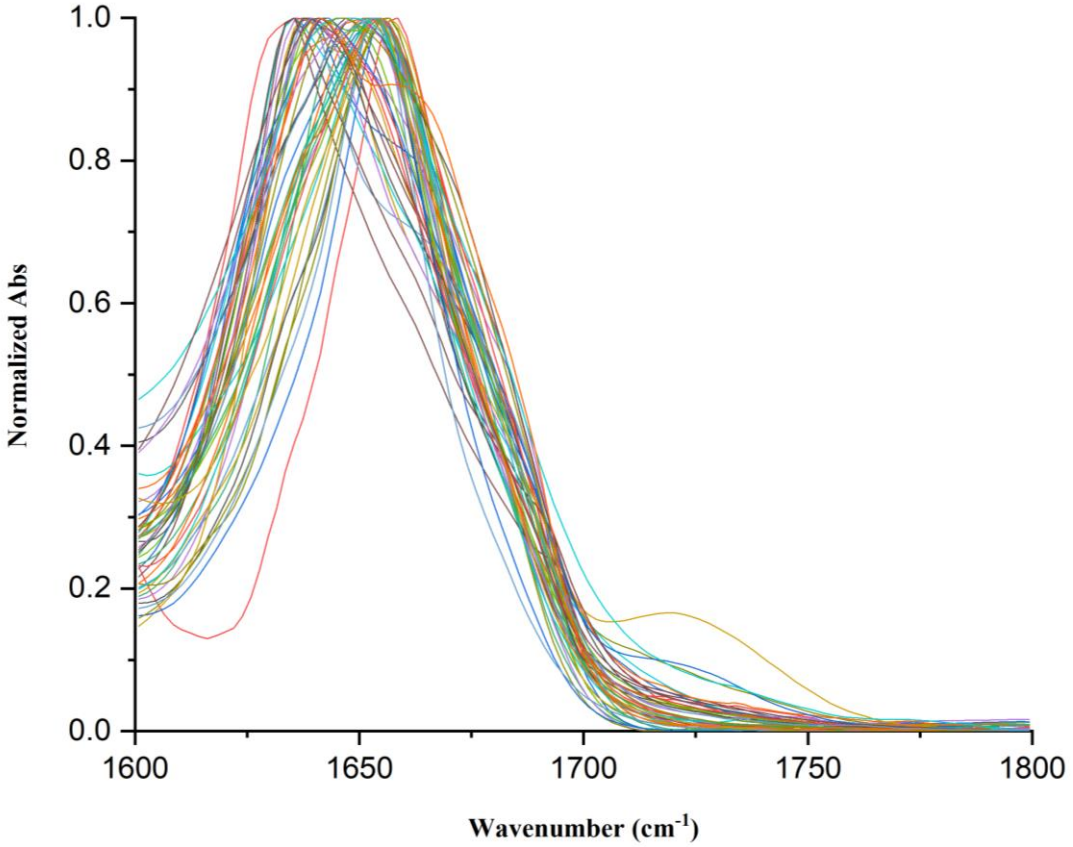

Figure 3.31. Transmission IR spectra of 47 proteins in aqueous state. The spectra were normalized by the interval method.

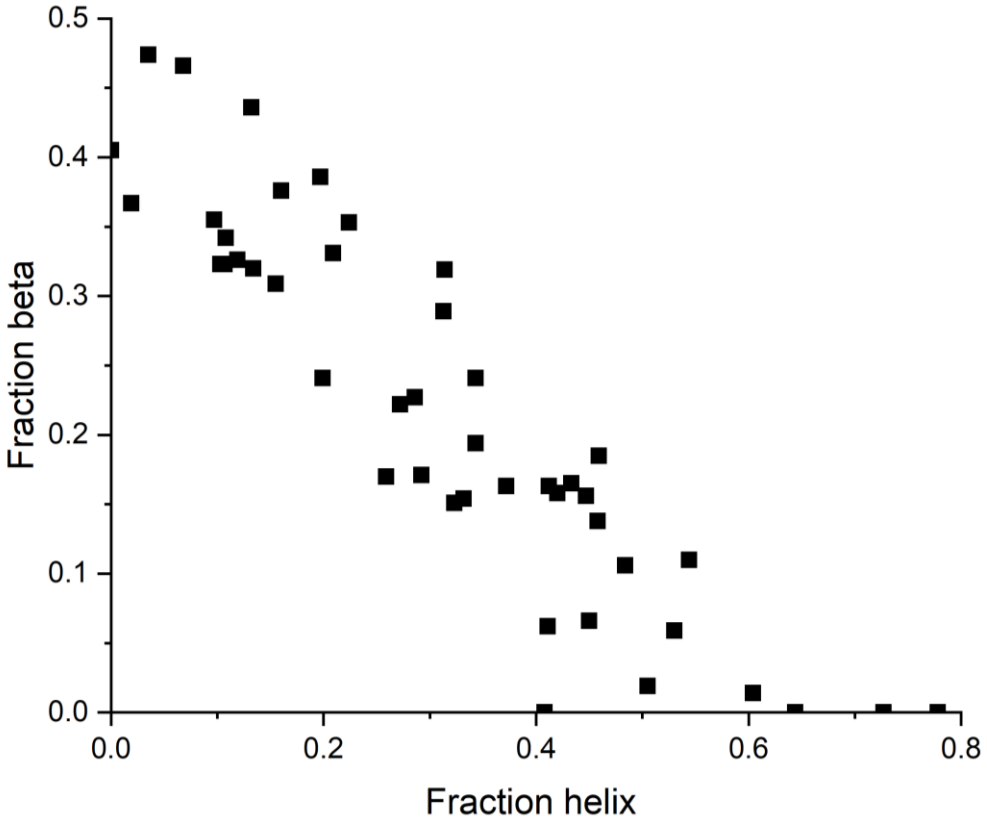

Figure 3.32. Coverage of SS corresponding to ref set in Figure 3.31.

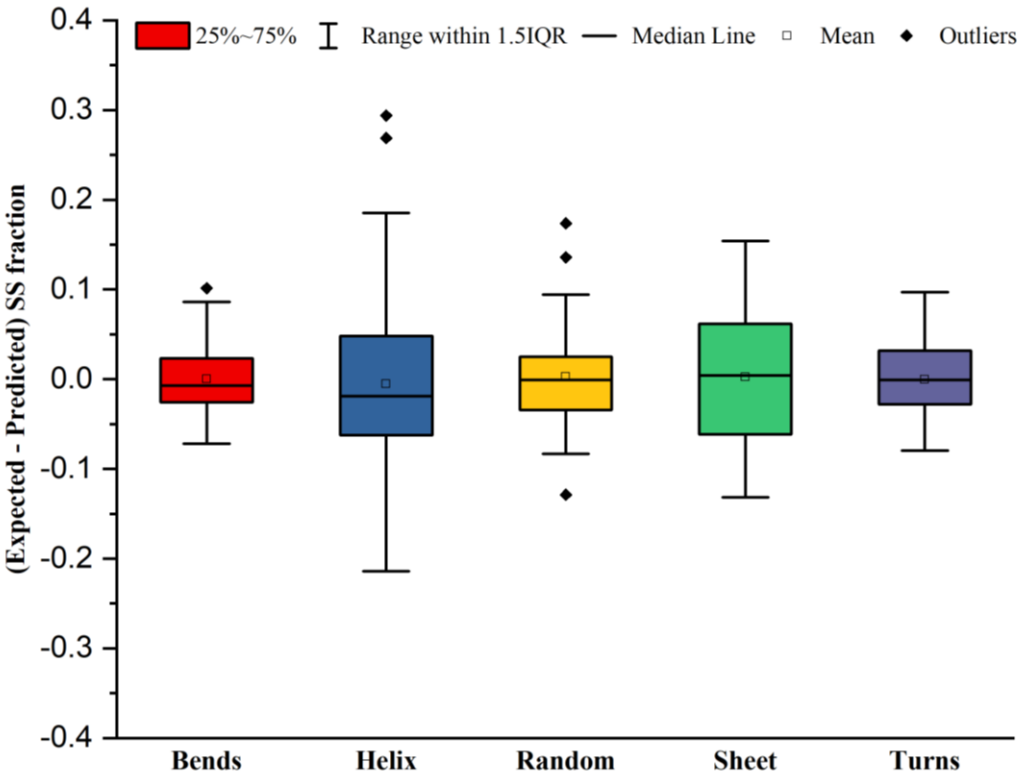

Figure 3.33. Results of the leave-one out validation of the normalized transmission IR 47-protein ref set measured in solution (Figure 3.31). The validation was run with a 40x40 map and 40000 iterations.

3.5.3 Circular dichroism

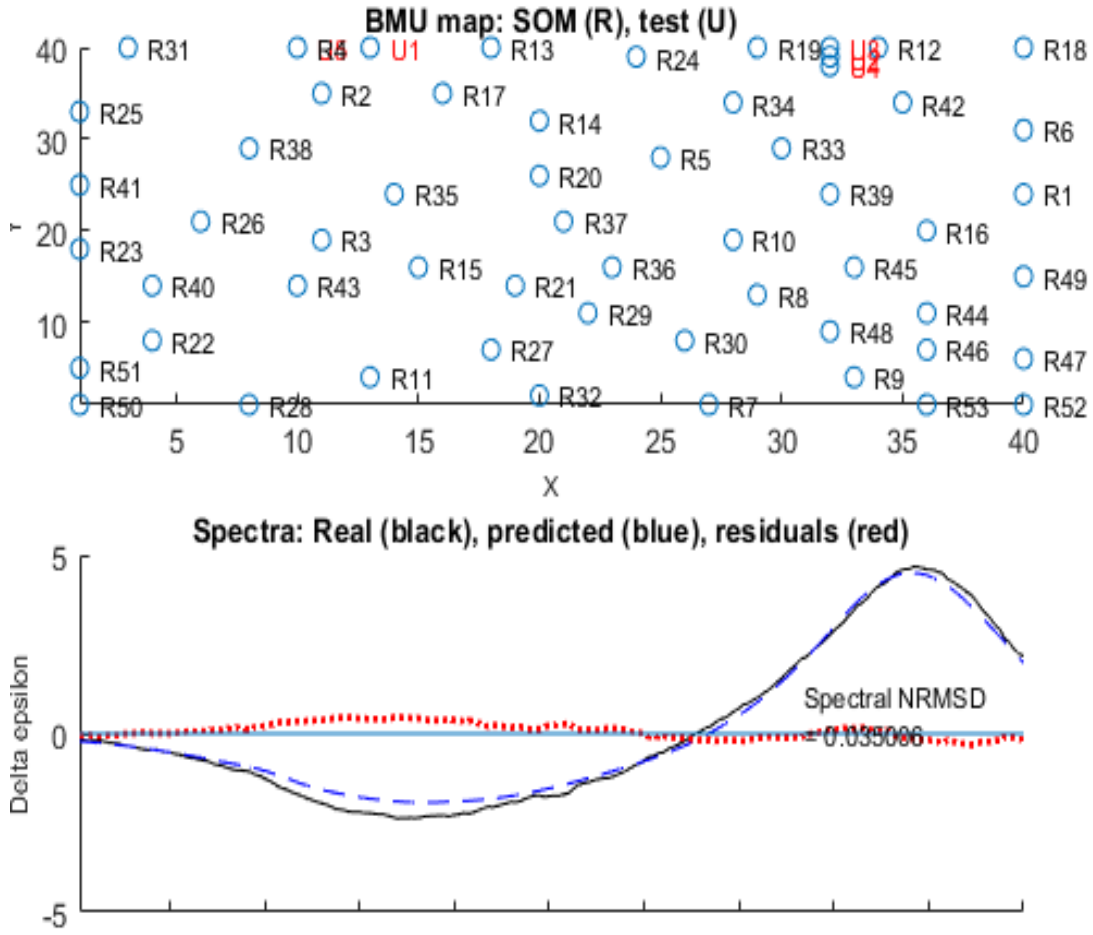

Figure 3.34. SOM-SS prediction of a Concanavalin CD spectrum.

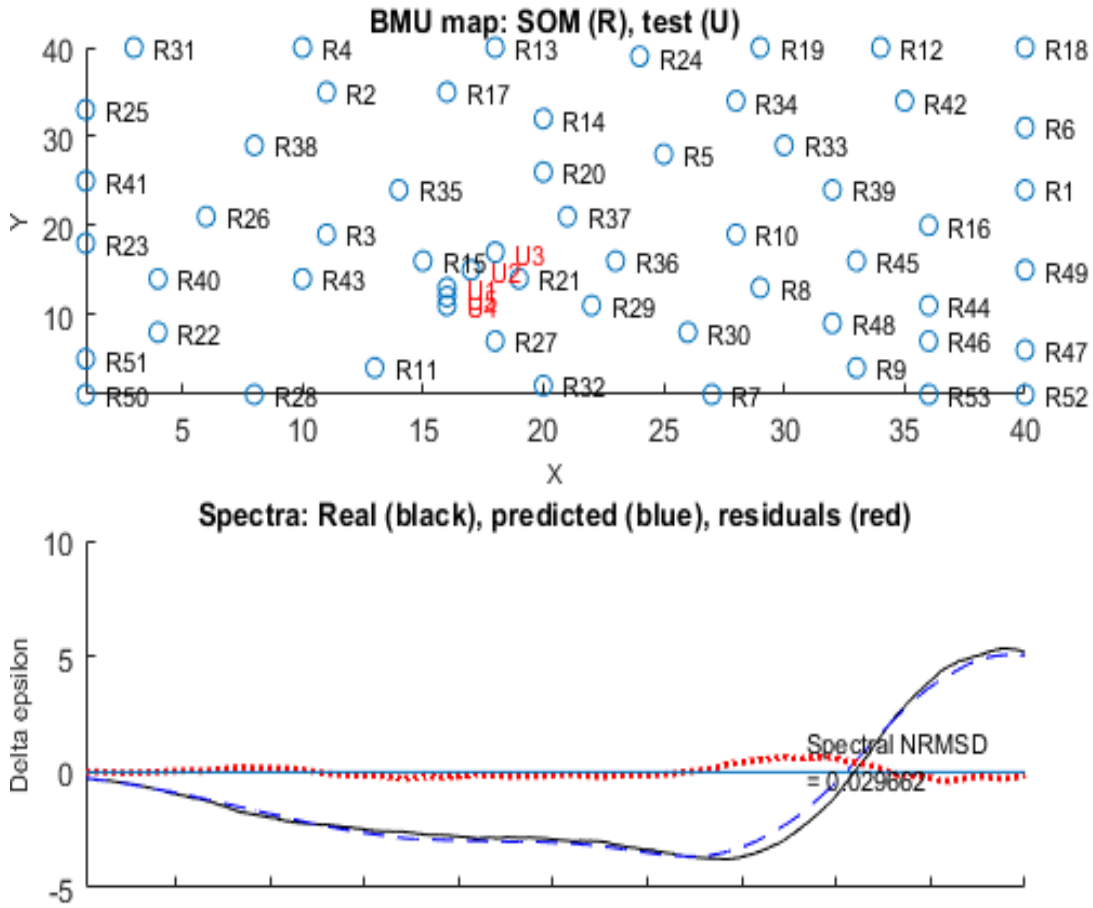

Figure 3.35. SOM-SS prediction of a Lysozyme CD spectrum.

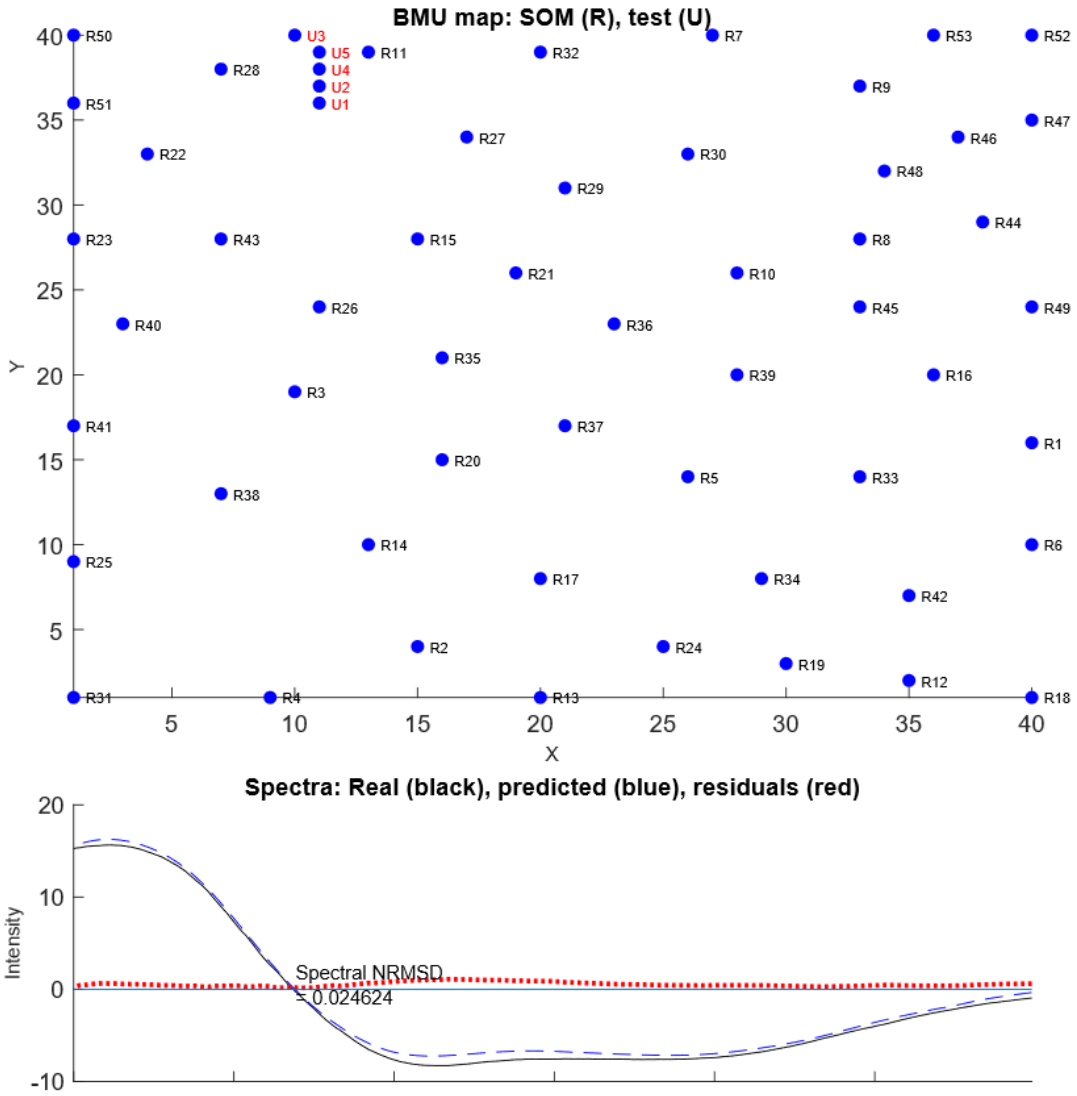

Figure 3.36. SOM-SS prediction of a BSA CD spectrum.

3.5.4 IR-ATR of proteins

In this section it is shown how the SOM algorithm works and the effect of the anomalous dispersion on the SOM predictions. 50 $mg.ml^{-1}$ solutions of 3 proteins representative of helical (Lysozyme), sheet (Concanavalin) and highly helical (BSA) structures (Figure 3.37) were collected in accordance with the protocol described in the methods section and tested through the chosen map from the previous section with the corresponding data processing (normalization).

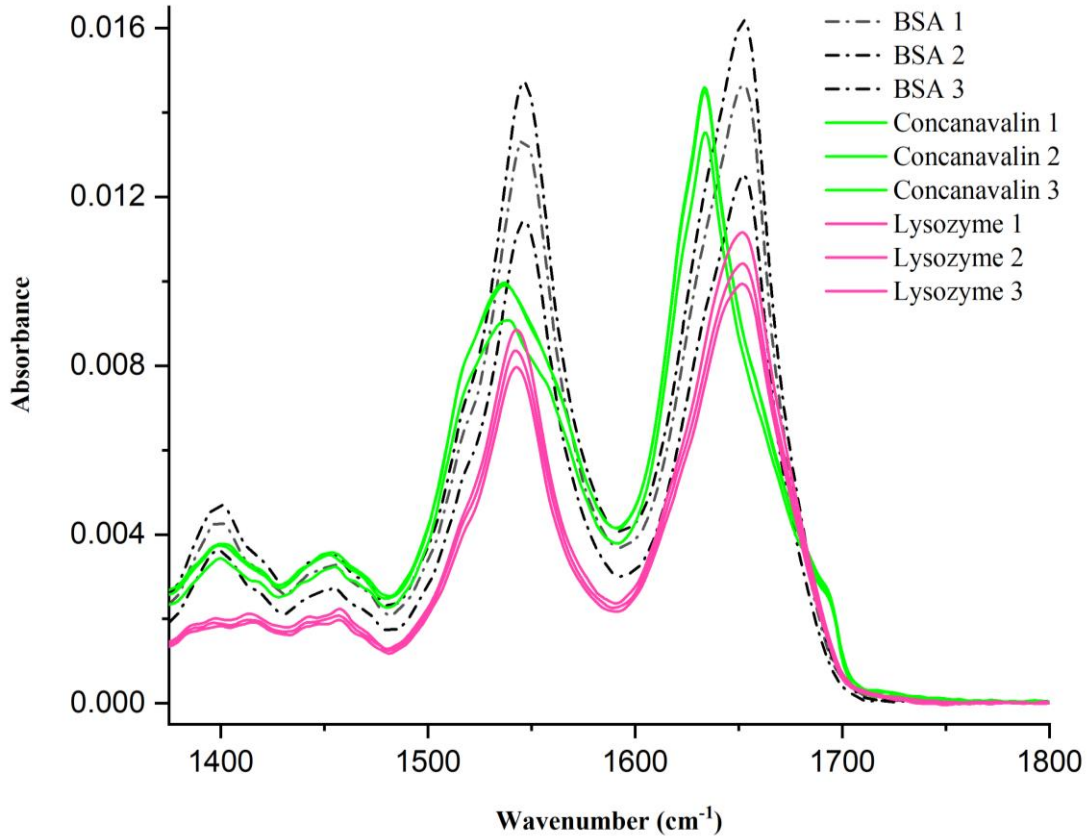

Figure 3.37. IR-ATR spectra of replicated 50 mg.ml^{-1} BSA, Concanavalin and Lysozyme solutions. The spectra were collected with a single bounce Specac ATR unit, accumulated over 250 scans and corrected as explained in the methods section.

The upper half of Figures 3.38, 3.39, 3.41, 3.42, 3.44, 3.45 is the trained map of the reference spectra with the position of the best matching units to the test spectrum indicated with the symbols U1-5. The NRMSD is a measure of the accuracy of the spectral fit which is illustrated at the bottom of each plot as an overlay of experimental (real) and predicted spectrum. The percentages of different secondary structures for the experimental ATR, experimental transmission, corrected ATR and CD spectra are given in Figures 3.40, 3.43 and 3.46 along with their X-Ray annotations.

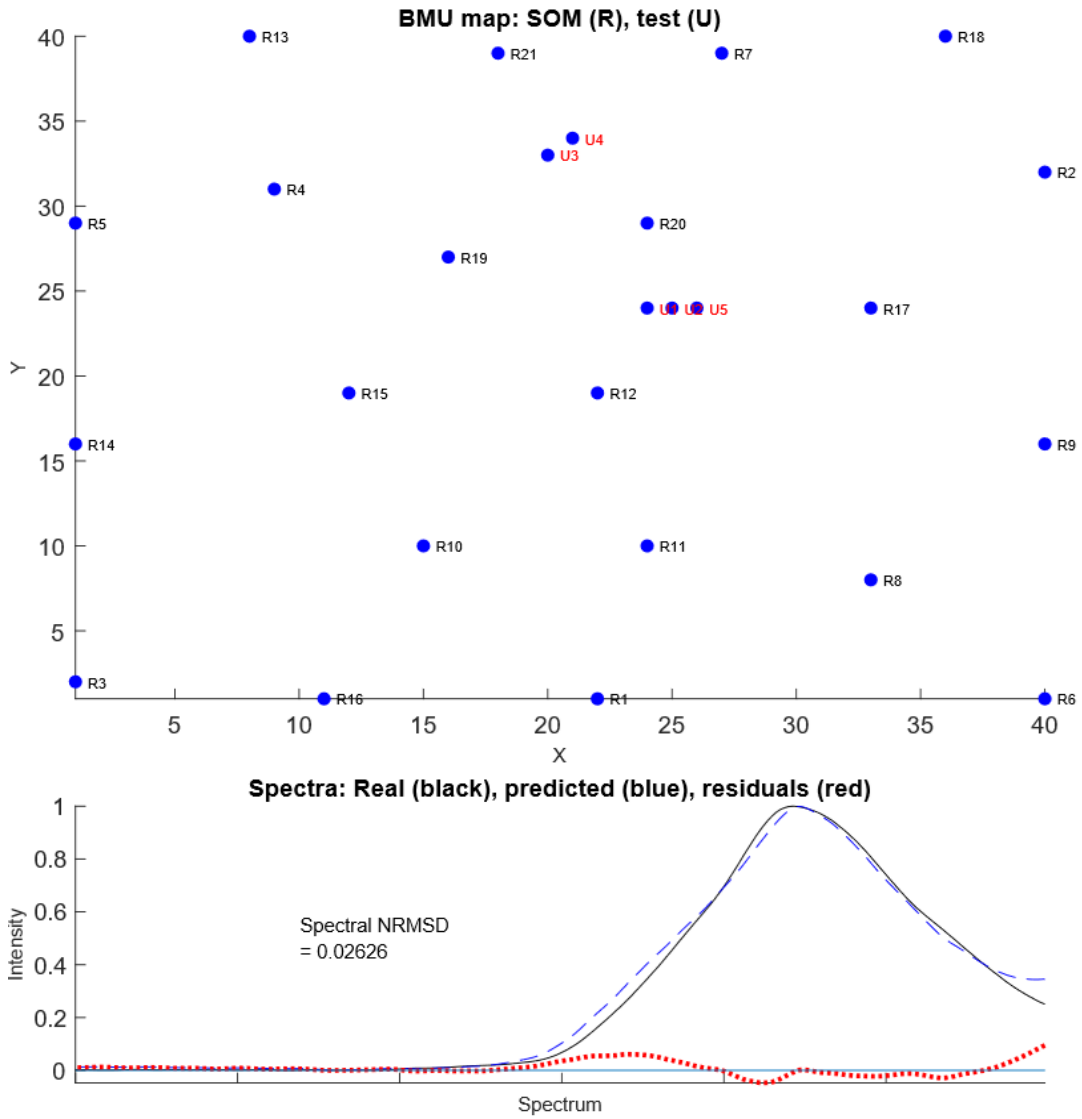

Figure 3.38. SOM prediction for the 50 mg.ml^{-1} experimental IR-ATR spectrum of Lysozyme by means of the 21 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm^{-1} and normalized).

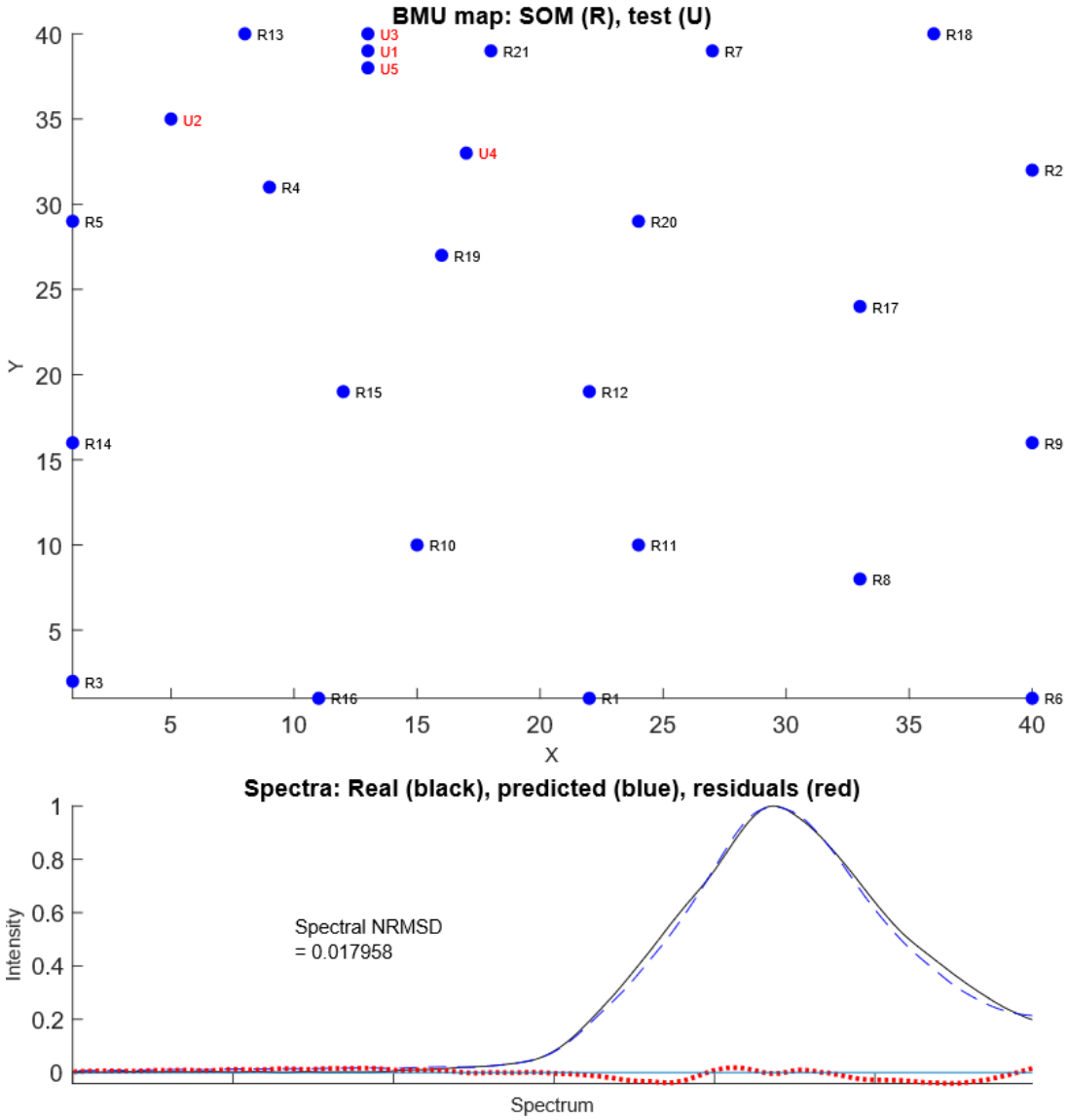

Figure 3.39. SOM prediction for the 50 mg.ml^{-1} corrected experimental IR-ATR spectrum of Lysozyme by means of the 21 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm^{-1} and normalized).

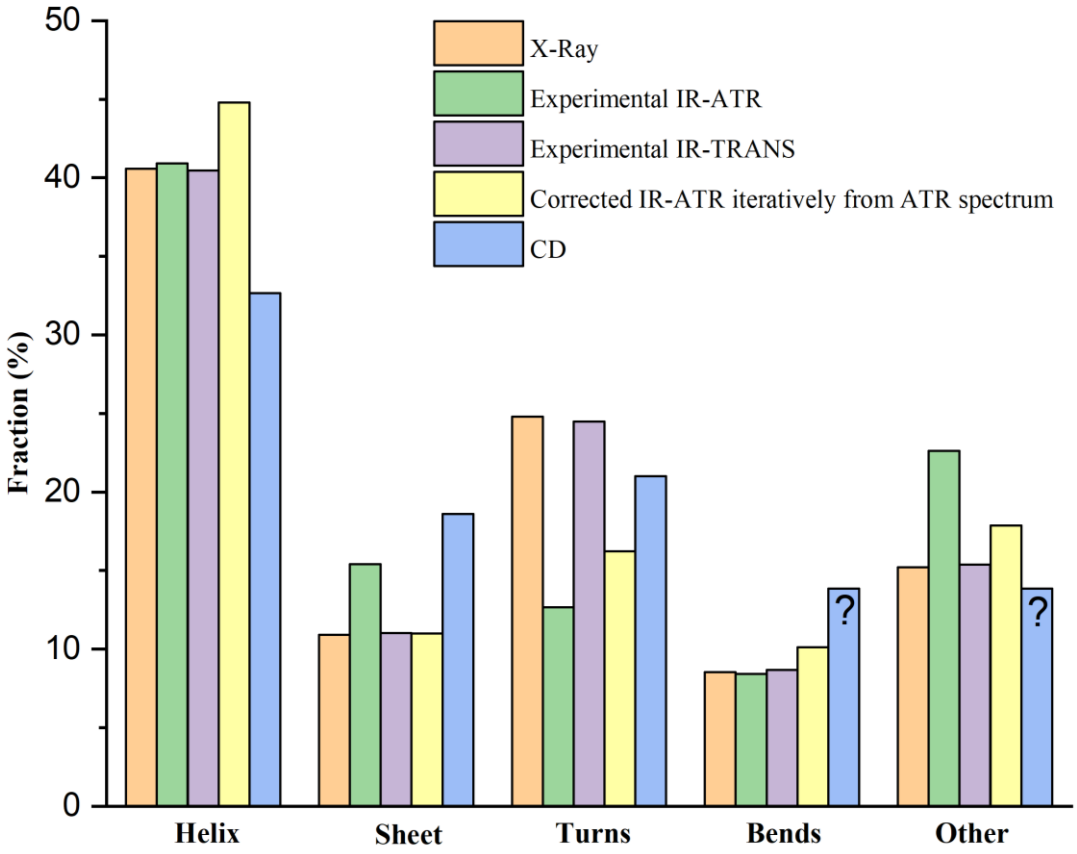

Figure 3.40. Comparison of X-Ray annotations of Lysozyme with the experimental IR-ATR, IR-Transmission, corrected IR-ATR iteratively from the ATR spectrum and Circular Dichroism-SOM predictions. The ATR was measured with a 50 $mg.ml^{-1}$ and the transmission with a 100 $mg.ml^{-1}$. Both were normalized by the interval method to test against SOM. The question marks mean we do not know what the relative amounts of bends and other would be from the CD set annotations. The training was performed with the 21-ref set and a 40x40 map with 40000 iterations.

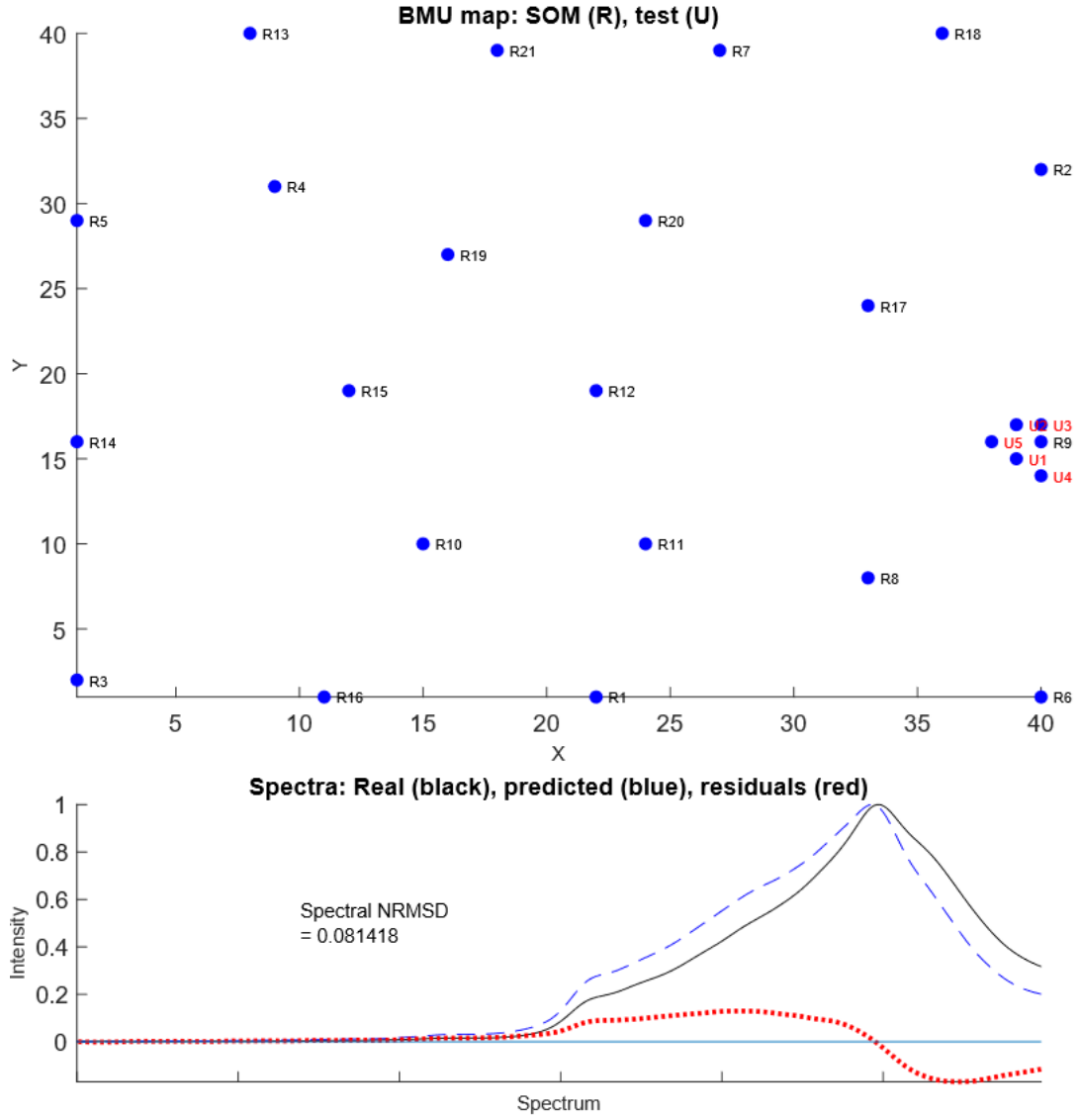

Figure 3.41. SOM prediction for the 50 mg.ml^{-1} experimental IR-ATR spectrum of Concanavalin by means of the 21 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm^{-1} and normalized).

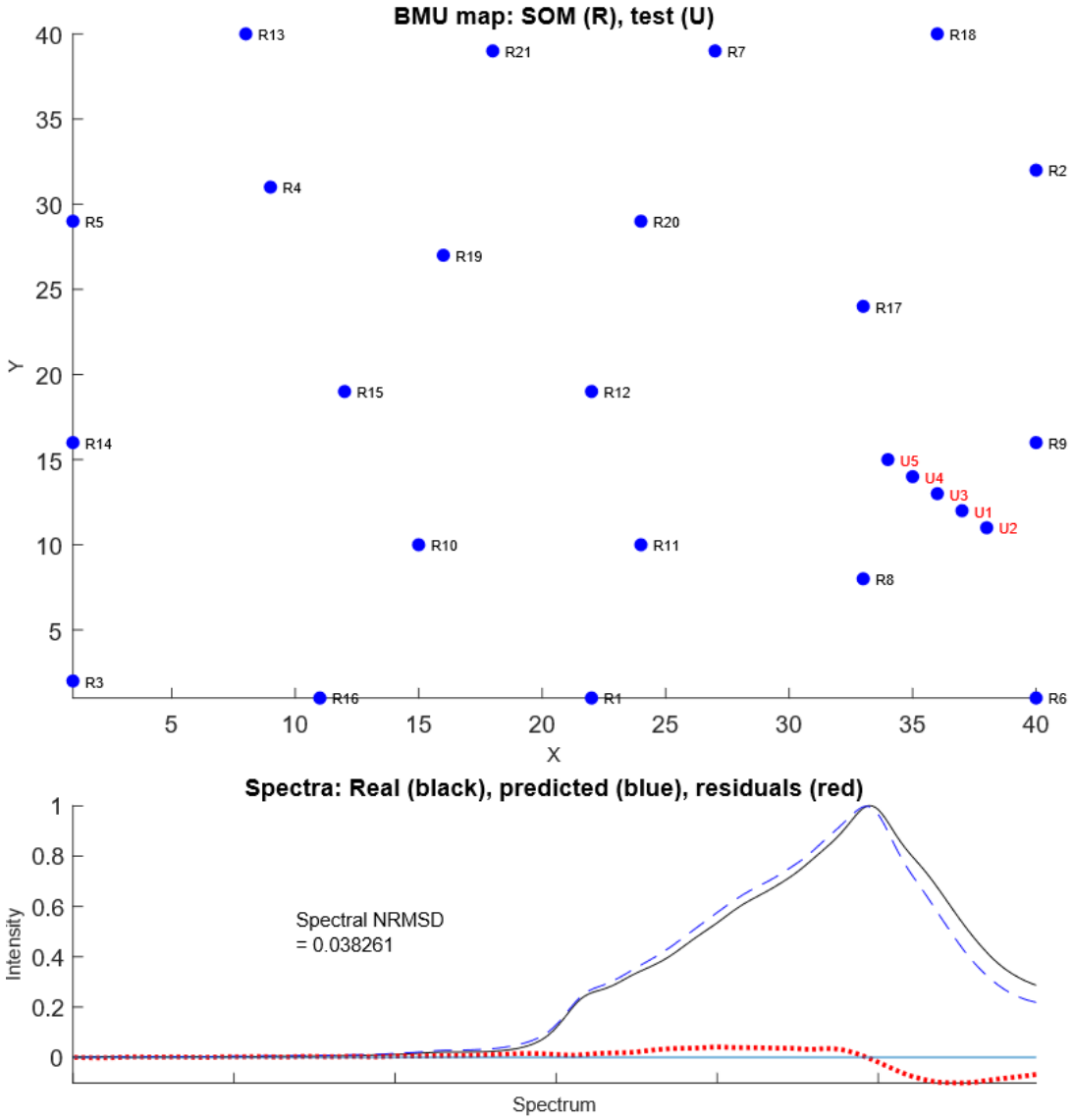

Figure 3.42. SOM prediction for the 50 mg.ml^{-1} corrected experimental IR-ATR spectrum of Concanavalin by means of the 21 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm^{-1} and normalized).

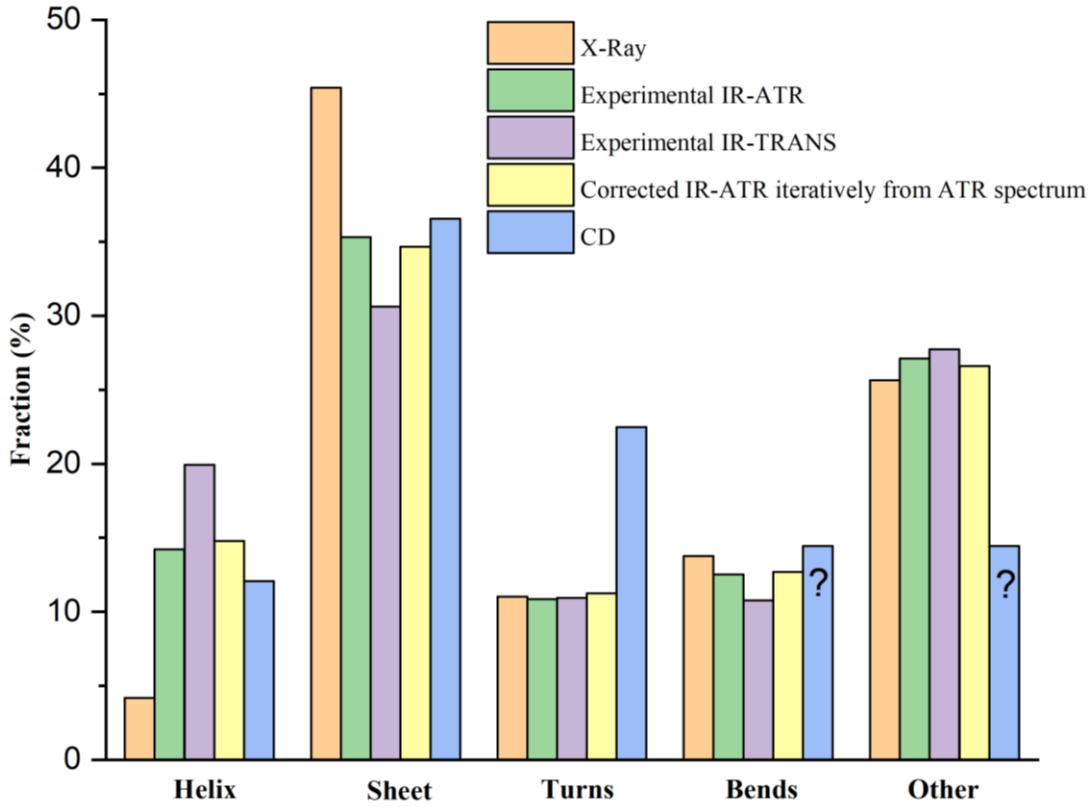

Figure 3.43. Comparison of X-Ray annotations of Concanavalin with the experimental IR-ATR, IR-Transmission, corrected IR-ATR iteratively from the ATR spectrum and Circular Dichroism-SOM predictions. The ATR was measured with a 50 $mg.ml^{-1}$ and the transmission with a 60.5 $mg.ml^{-1}$. Both were normalized by the interval method to test against SOM. The question marks mean we do not know what the relative amounts of bends and other would be from the CD set annotations. The training was performed with the 21-ref set and a 40x40 map with 40000 iterations.

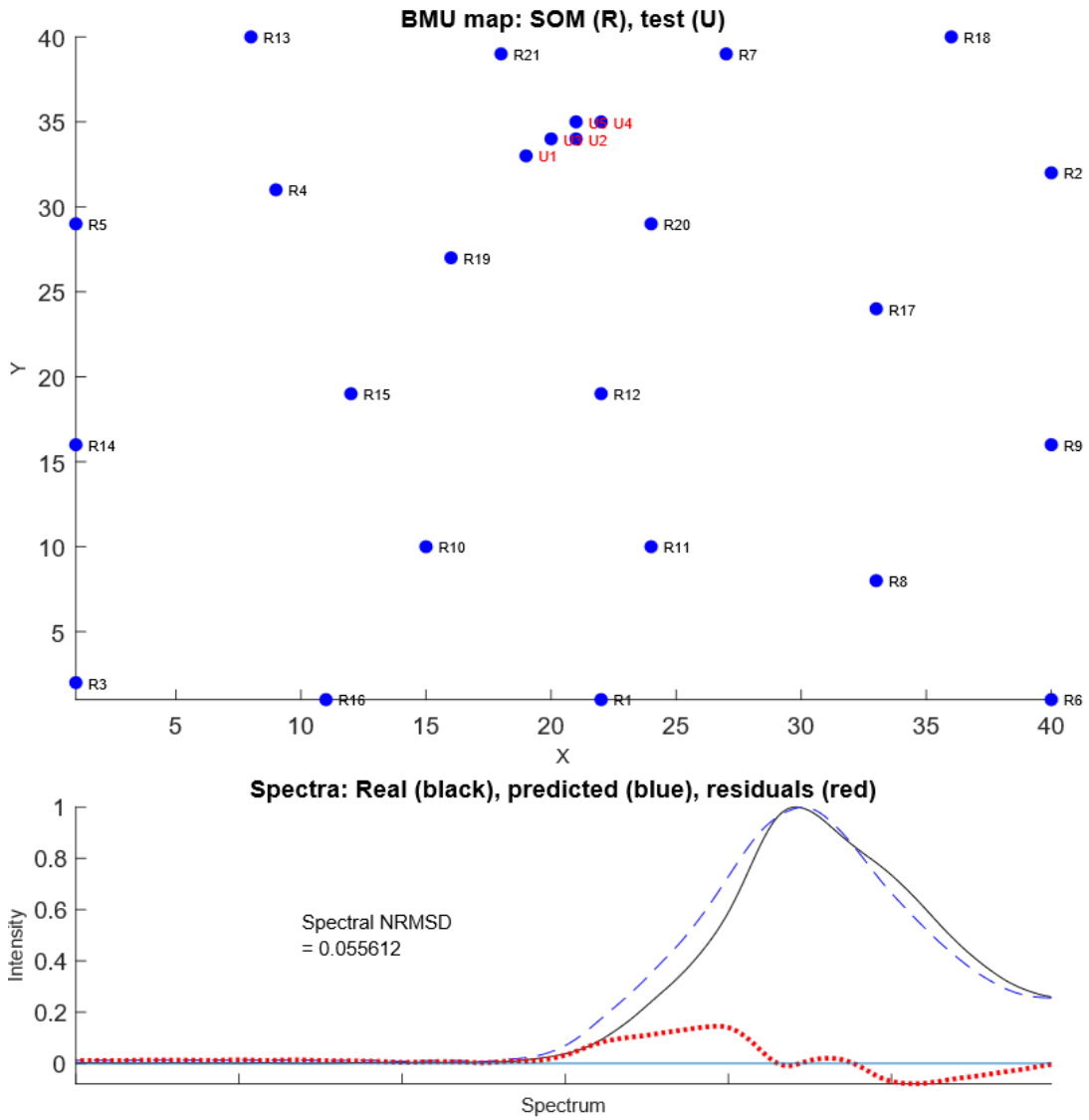

Figure 3.44. SOM prediction for the 50 mg.ml^{-1} experimental IR-ATR spectrum of BSA by means of the 21 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm^{-1} and normalized).

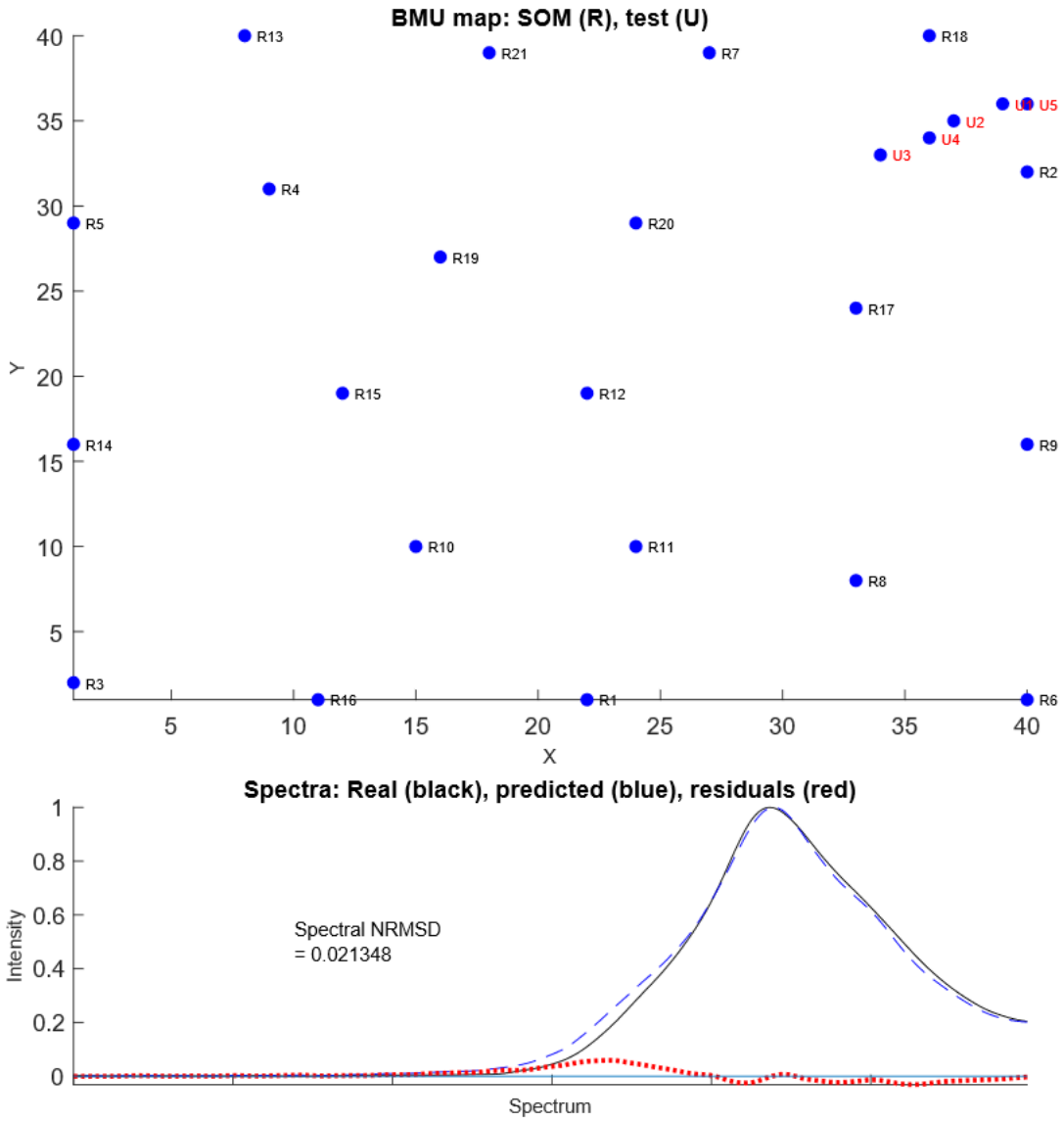

Figure 3.45. SOM prediction for the 50 mg.ml^{-1} corrected experimental IR-ATR spectrum of BSA by means of the 21 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm^{-1} and normalized).

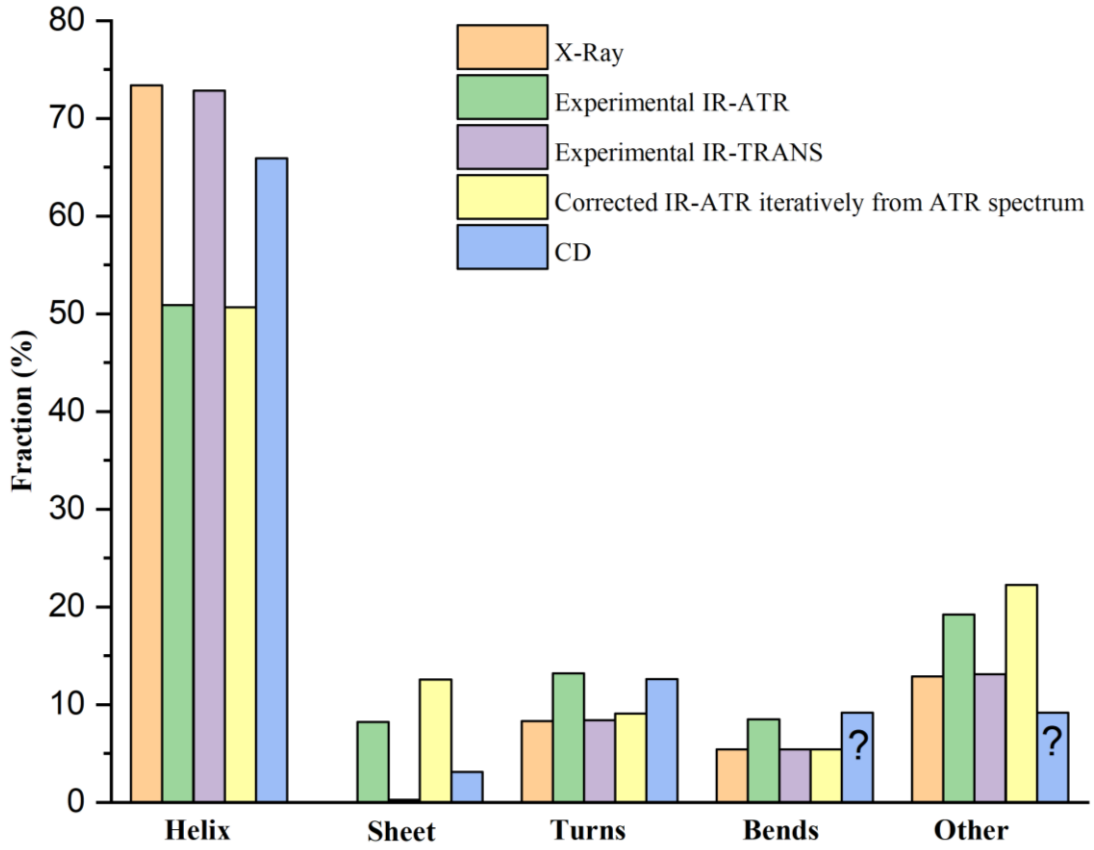

Figure 3.46. Comparison of X-Ray annotations of BSA with the experimental IR-ATR, IR-Transmission, corrected IR-ATR iteratively from the ATR spectrum and Circular Dichroism-SOM predictions. The ATR was measured with a 50 $mg.ml^{-1}$ and the transmission with a 80 $mg.ml^{-1}$. Both were normalized by the interval method to test against SOM. The question marks mean we do not know what the relative amounts of bends and other would be from the CD set annotations. The training was performed with the 21-ref set and a 40x40 map with 40000 iterations.

Although the predictions were fairly close to the X-Ray annotations and CD predictions, the effect of the anomalous dispersion correction was a bit contradictory. To judge by the improvement in the goodness of fit after applying the ATR correction, one would expect the predictions of the corrected ATR spectra to be closer to that of transmission but that was not the case. To check whether this was a consequence of the limited representation of mixed helix/sheet structures in the set, the predictions were repeated with a training set provided by BioTools Inc. with a larger number of proteins (47) collected by transmission. The results are shown in Figures 3.47-49. The SOM maps can be found in the Appendices-B.3.

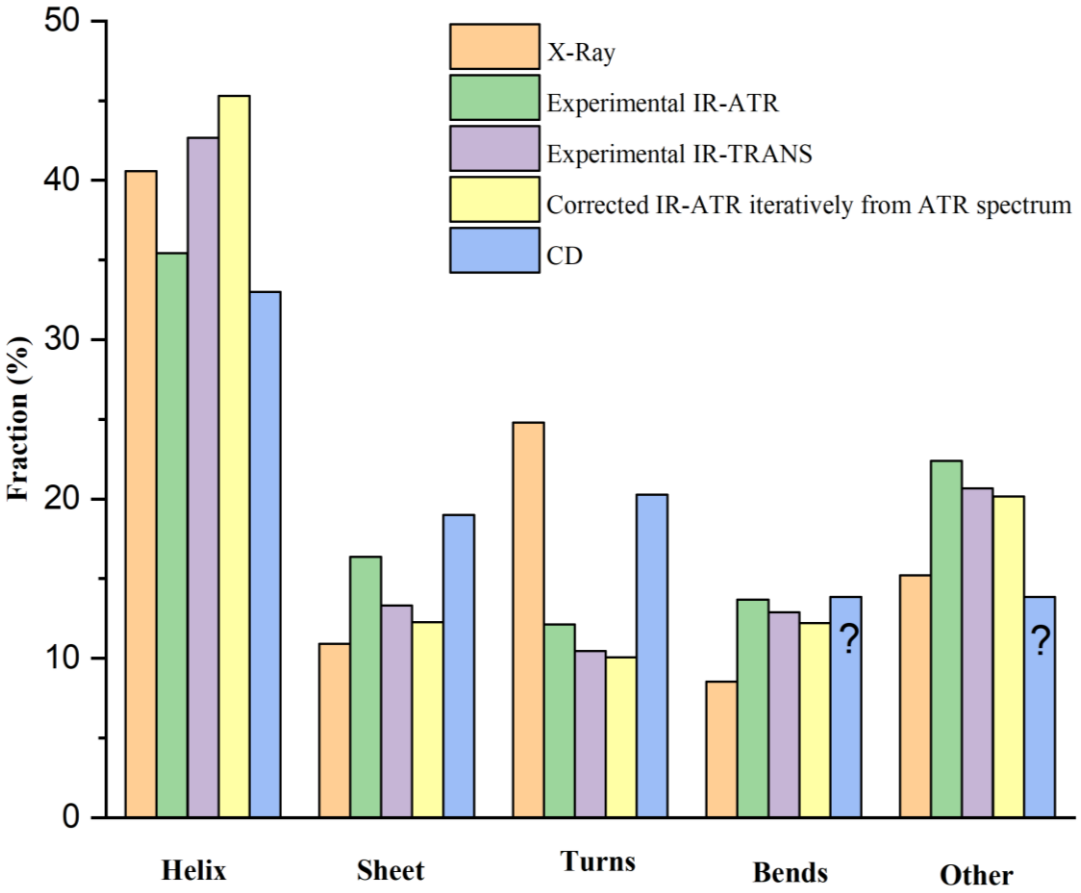

Figure 3.47. Comparison of X-Ray annotations of Lysozyme with the experimental IR-ATR, IR-Transmission, corrected IR-ATR iteratively from the ATR spectrum and Circular Dichroism-SOM predictions. The ATR was measured with a 50 $mg.ml^{-1}$ and the transmission with a 100 $mg.ml^{-1}$. Both were normalized by the interval method to test against SOM. The question marks mean we do not know what the relative amounts of bends and other would be from the CD set annotations. The training was performed with a 47-ref set provided by biopharma spec and a 40x40 map with 40000 iterations.

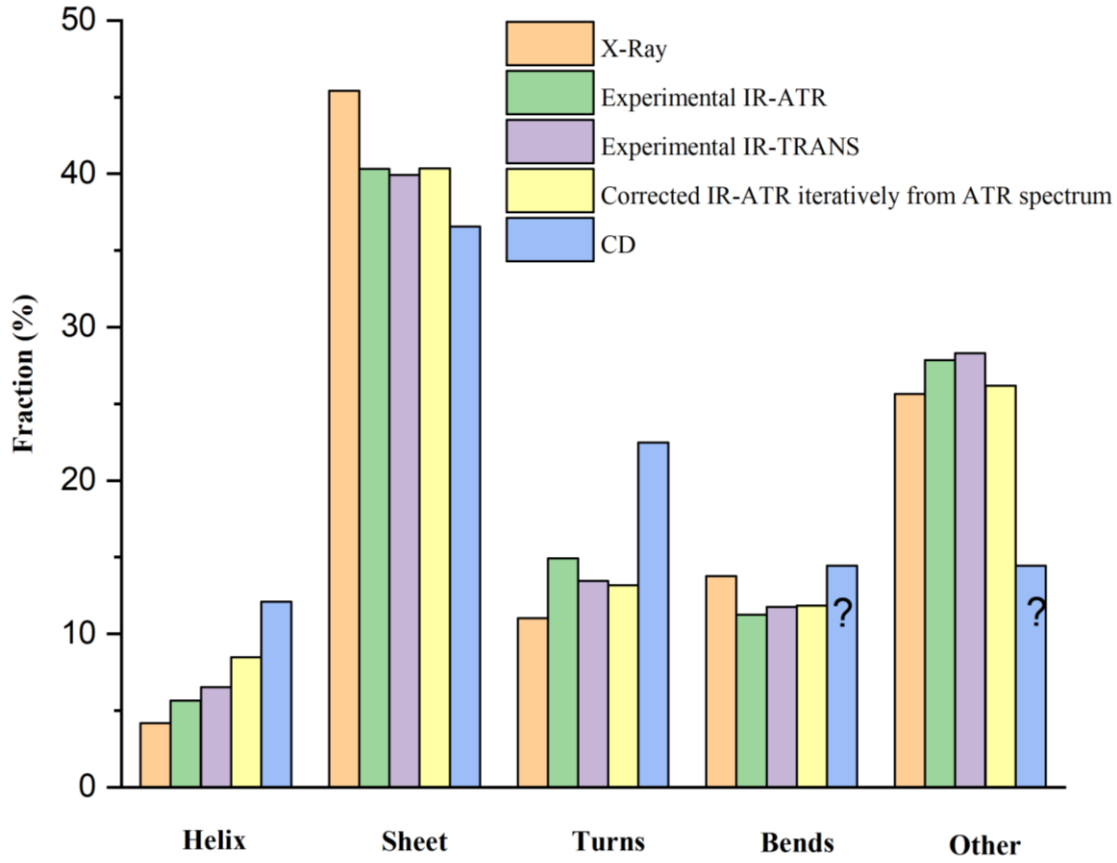

Figure 3.48. Comparison of X-Ray annotations of Concanavalin with the experimental IR-ATR, IR-Transmission, corrected IR-ATR iteratively from the ATR spectrum and Circular Dichroism-SOM predictions. The ATR was measured with a 50 $mg.ml^{-1}$ and the transmission with a 60.5 $mg.ml^{-1}$. Both were normalized by the interval method to test against SOM. The question marks mean we do not know what the relative amounts of bends and other would be from the CD set annotations. The training was performed with a 47-ref set provided by biopharma spec and a 40x40 map with 40000 iterations.

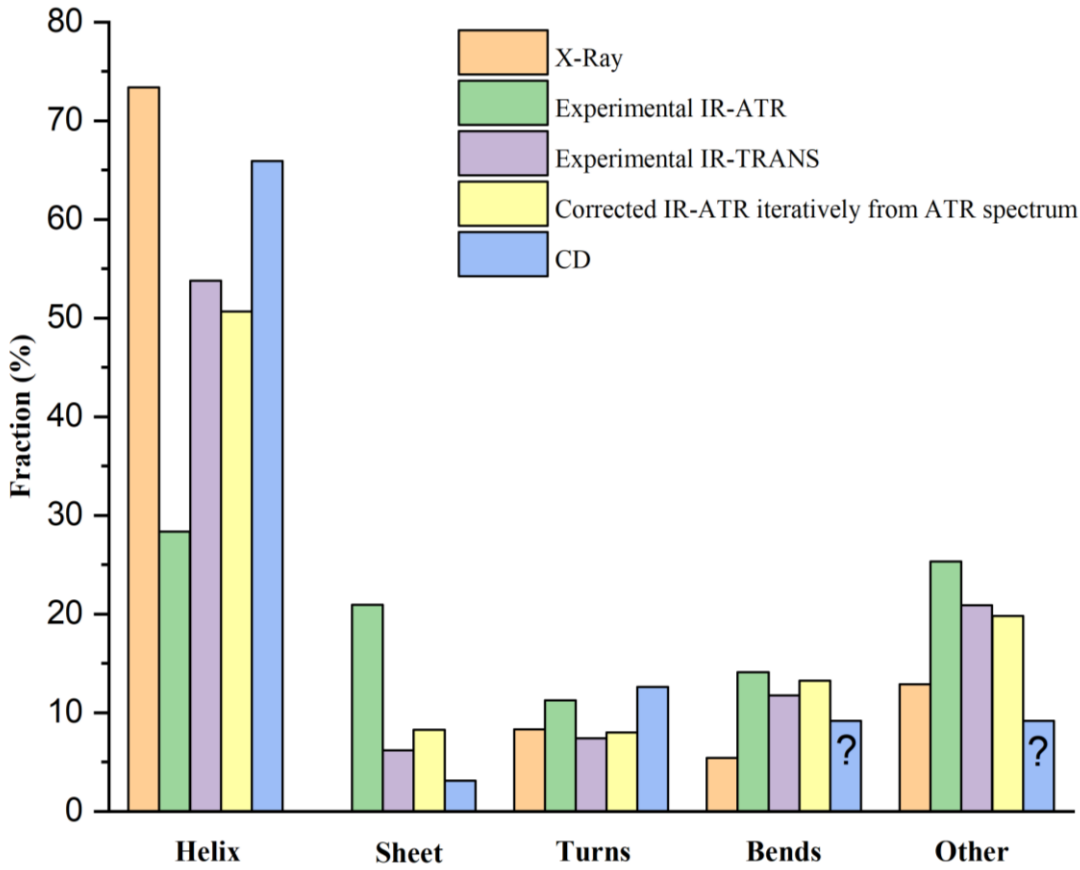

Figure 3.49. Comparison of X-Ray annotations of BSA with the experimental IR-ATR, IR-Transmission, corrected IR-ATR iteratively from the ATR spectrum and Circular Dichroism-SOM predictions. The ATR was measured with a 50 mg.ml^{-1} and the transmission with a 80 mg.ml^{-1}. Both were normalized by the interval method to test against SOM. The question marks mean we do not know what the relative amounts of bends and other would be from the CD set annotations. The training was performed with a 47-ref set provided by biopharma spec and a 40x40 map with 40000 iterations.

On a separate note, most literature data are collected on samples of at least 3 mg.ml^{-1} [46] but we were targeting data collection to overlap with CD concentrations in the near UV region which are typically 1 mg.ml^{-1}. However, we found significant challenge in getting down to 1 mg.ml^{-1} because of artefacts arising in the region of the amide I (Figure 3.51). We collected spectra of 2 mg.ml^{-1} solutions of BSA, Hemoglobin with a 6-bounce ATR unit and 1 mg.ml^{-1} of an antibody with a single bounce one (Figure 3.50) and tested them against SOM (Table 3-3) to compare their reproducibility with the 50 mg.ml^{-1} solutions of BSA, Concanavalin and Lysozyme used above (Table 3-4).

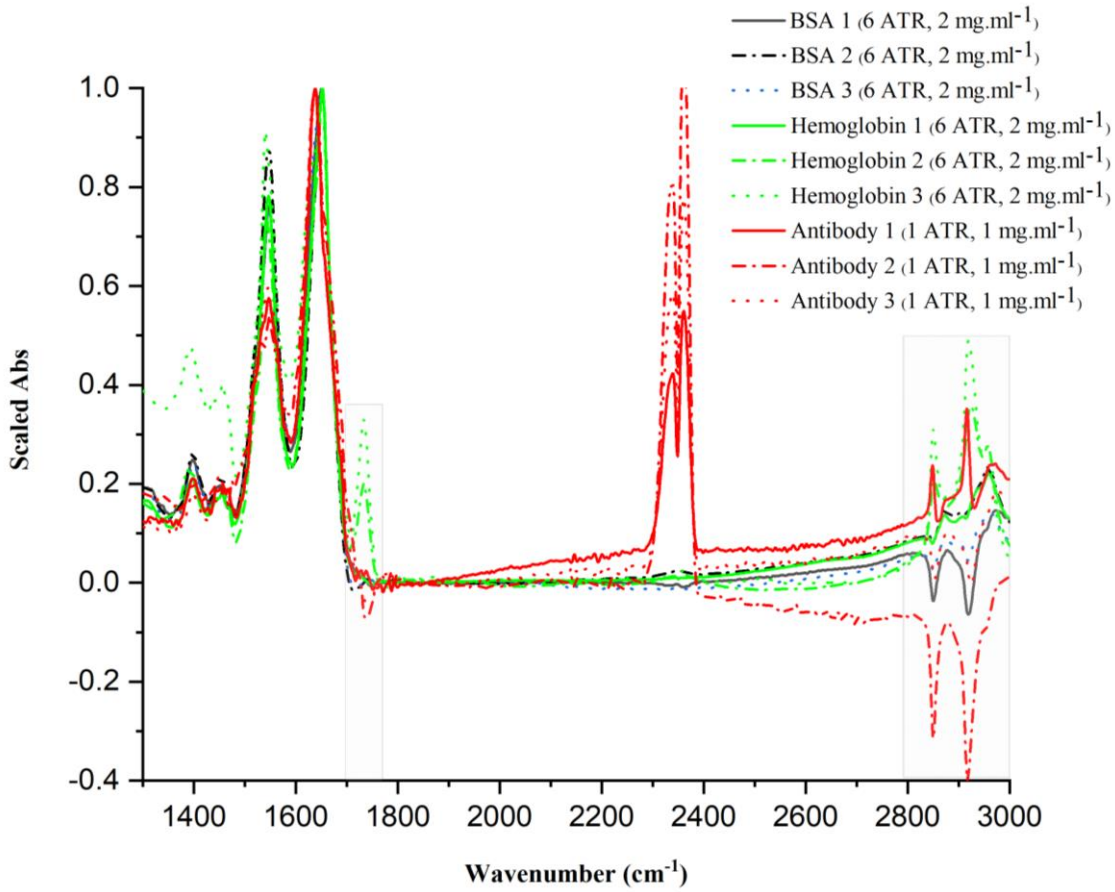

Figure 3.50. BSA, Hemoglobin and Antibody IR-ATR spectra of 2, 2 and 1 mg.ml^{-1} concentration respectively. The spectra were collected with a Specac 6-bounce ATR unit (BSA and Hemoglobin) and a PIKE MIRAcle single bounce ATR unit (Antibody) and scaled by dividing by the value of the amide I max.

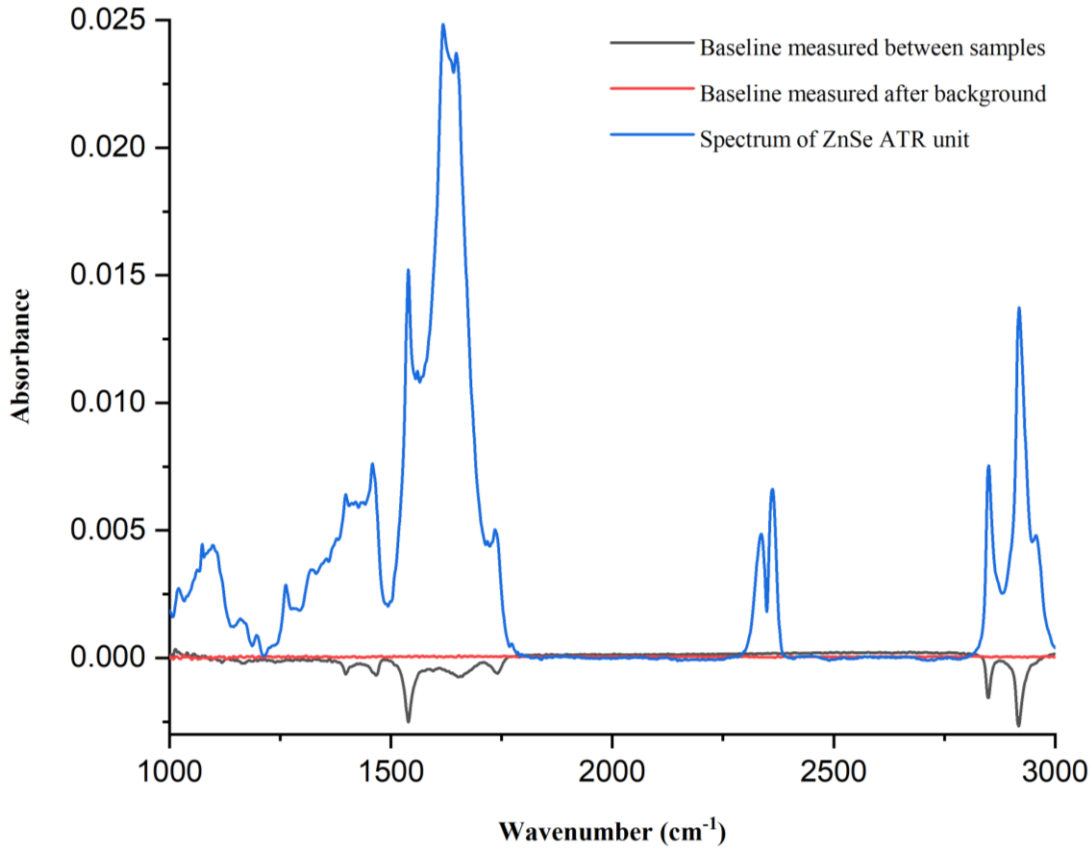

Figure 3.51. IR spectral baselines collected with a PIKE MIRAcle single bounce ATR unit between measurements.

Table 3-3. SOM predictions of ATR-IR replicates of BSA, Concanavalin and Lysozyme with a concentration of 50 mg.ml^{-1}. The predictions were performed with the 21 proteins map of 20x20, 5 BMU and 20000 iterations used before.

Proteins	α-helix	β-sheet	Other
BSA 1	50.900	8.235	40.775
BSA 2	50.900	8.235	40.775
BSA 3	50.900	8.235	40.775
Concanavalin 1	13.180	35.810	50.010
Concanavalin 2	14.210	35.320	50.470
Concanavalin 3	14.190	35.330	50.480
Lysozyme 1	36.660	18.560	44.780
Lysozyme 2	36.660	18.560	44.780
Lysozyme 3	41.020	15.390	43.590

Table 3-4. SOM predictions of ATR-IR replicates of Hemoglobin, BSA and Antibody measured with concentrations of 2, 2 and 1 mg.ml^{-1} respectively.

Proteins	α-helix	β-sheet	Other
BSA 1	29.540	25.040	45.420
BSA 2	50.900	8.235	40.865
BSA 3	28.610	26.560	44.83
Hemoglobin 1	45.670	12.550	41.780
Hemoglobin 2	42.590	14.770	42.640
Hemoglobin 3	17.650	34.720	47.630
Antibody 1	13.810	35.490	50.700
Antibody 2	18.010	35.520	46.470
Antibody 3	15.500	34.700	49.800

3.6 CONCLUSIONS

In this chapter we report the collection of a IR reference set spectra of proteins in aqueous state and train a neural network algorithm (SOM) with the spectra processed with different strategies (normalized, deconvolved + normalized and in MRW extinction coefficient) for prediction of SS. A leave-one-out validation was performed on each of the sets with 40x40 maps and 40000 iterations and represented as boxplots. “Deconvolved + normalized” was found to provide the most accurate results of the three, followed by “normalized” and “MRW extinction coefficient”. Furthermore, the normalized 21-protein set was also compared to the normalized 47-protein one with the latter showing a clear improvement over the first.

On a separate note, the experimental ATR spectra of three proteins with known SS and their respective ATR corrected versions were tested against a map trained with the normalized 21-proteins set and their predictions compared with those of the experimental transmission spectra and X-ray annotations. Moreover, the spectra were put through a larger normalized ref set (47 proteins) provided by biopharma spec and compared to the results by the 21-protein one. This showed that, as expected, larger number of proteins give more accurate predictions. As for the

anomalous dispersion corrections, no improvement was seen with the 21-protein ref set but much closer predictions to that of transmission were observed with the 47-protein one instead.

Regarding the limit of detection, it was clear that the reproducibility gets significantly affected by the concentration of the sample. This lack of reproducibility was blamed on the appearance of spectral artefacts in the background attributed to the ZnSe crystal whose magnitude and intensity across the spectrum seemed to change from measurement to measurement. These artefacts had orders of magnitude of $\sim 10^{-3}$ which is the reason why they were only significant at low levels of concentration (< 5 mg.ml^{-1}).

4 RAMAN AND RAMAN OPTICAL ACTIVITY SPECTROSCOPY OF PROTEINS AND SELF - ORGANIZING MAPS

4.1 INTRODUCTION

Raman spectroscopy is also a vibrational spectroscopic technique, although based on scattering as discussed in chapter 1, that gives information on different aspects of the structure of proteins of special interest in the pharmaceutical industry[105]. As in IR, the characteristic amide bands (A, B, I-VII) are also active in Raman with the amides I -between 1620 and 1720 cm^{-1}- and III -from 1230 to 1300 cm^{-1}- being the most prominent ones and of biggest interest for SS prediction[7,106]. The correlations of these bands with SS have been established back in the 1970s based on studies of induced alpha-beta transitions of poly-L-lysine and comparative analysis with X-Ray and CD spectroscopy of proteins[107,108]. Moreover, intensity and position markers of conformation, hydrogen bonding, protonation and hydrophobic interactions of tryptophan, tyrosine, cysteine and histidine have been identified and used as the basis for tertiary structure analysis[52,109]. Many methods exist in the literature for the extraction of information on the SS contents. They can be split into two classes: models based on reference sets of proteins of well-known SS[108,110,111] reviewed and collected by Bandekar[112] and some band fitting approaches in the more recent years[106,113].

Along with IR spectroscopy, Raman can measure in both aqueous and solid state and thus could potentially solve whether the disagreements between CD and X-Ray predictions acknowledged in the literature are due to limitations on the interpretation of CD or because of a different aggregation state[108]. However, compared to IR, the interference from the water vibrational mode in the region of the amide I is much smaller in Raman making water subtraction unnecessary in some cases[52,105]. Furthermore, Raman bands can be selectively enhanced by tuning the excitation wavelength to yield a resonance effect in what is commonly known as Raman Resonance Spectroscopy[52,105]. Although Raman spectroscopy offers many benefits compared to other spectroscopic techniques, impurity chromophores -even the smallest amounts and despite careful purification- can produce large amounts of fluorescence resulting in saturation of the Raman signal[105]. There is a wide range of

techniques for the suppression of background fluorescence but photobleaching, because of its simplicity, is the most popular approach and the one that was used most in this work[114,115].

Raman optical activity (ROA) is the chiral version of Raman spectroscopy and the scattering equivalent to CD, in which the scattered light acquires a small but measurable degree of circularity when interacting with chiroptical molecules[58]. The most prominent bands in ROA spectra of proteins are the amide I between 1630 and 1700 cm^{-1} and amide III from 1230 to 1340 cm^{-1} which, just like in Raman and IR, are related to the backbone of the protein and thus provide information on folding[116]. Although the correlation between SS and ROA spectra has been well established in the literature[60,117,118], we failed to find methods for the extraction of SS components. ROA can only be measured in aqueous state, so no solid reference set was measured. In this chapter, we focused our efforts in compiling a large set of protein Raman and ROA spectra to train SOM for protein SS prediction. Also, we reviewed the existing protocols and designed a SOP for protein data acquisition in both solid and aqueous state for Raman and aqueous only for ROA spectroscopy.

4.2 Secondary and tertiary structure markers for Raman and ROA

4.2.1 Overview

In Raman spectroscopy, molecular groups of different frequencies interact with an external electric field that disturbs the normal electronic distribution creating a dipole moment with three frequency components**.** Raman spectroscopy is based on the Stokes shift -excitation frequency minus the characteristic fundamental frequency of the oscillator-. The natural frequency of the oscillator, as discussed in chapter 3, depends on the reduced mass and constant forces of the bonds according to Equation 3.7 but unlike IR, Raman intensities depend on the variation of the polarizability with the normal coordinate of vibration. This is the reason why not all the amide vibrations mentioned in the introduction of chapter 3 are apparent in Raman spectra (e.g., absence of amide II and V)[112] as the polarizability change during a vibration is small. Instead, side chains -in particular those from aromatic groups- which are not visible in IR, show up in Raman spectra[105,119]. Figures 4.1 and 4.2 show the spectra of representative helical (Human albumin), sheet (Jacalin),

sheet + helical (Deoxyribonuclease) and irregular (Bungarotoxin) proteins by Raman and ROA respectively.

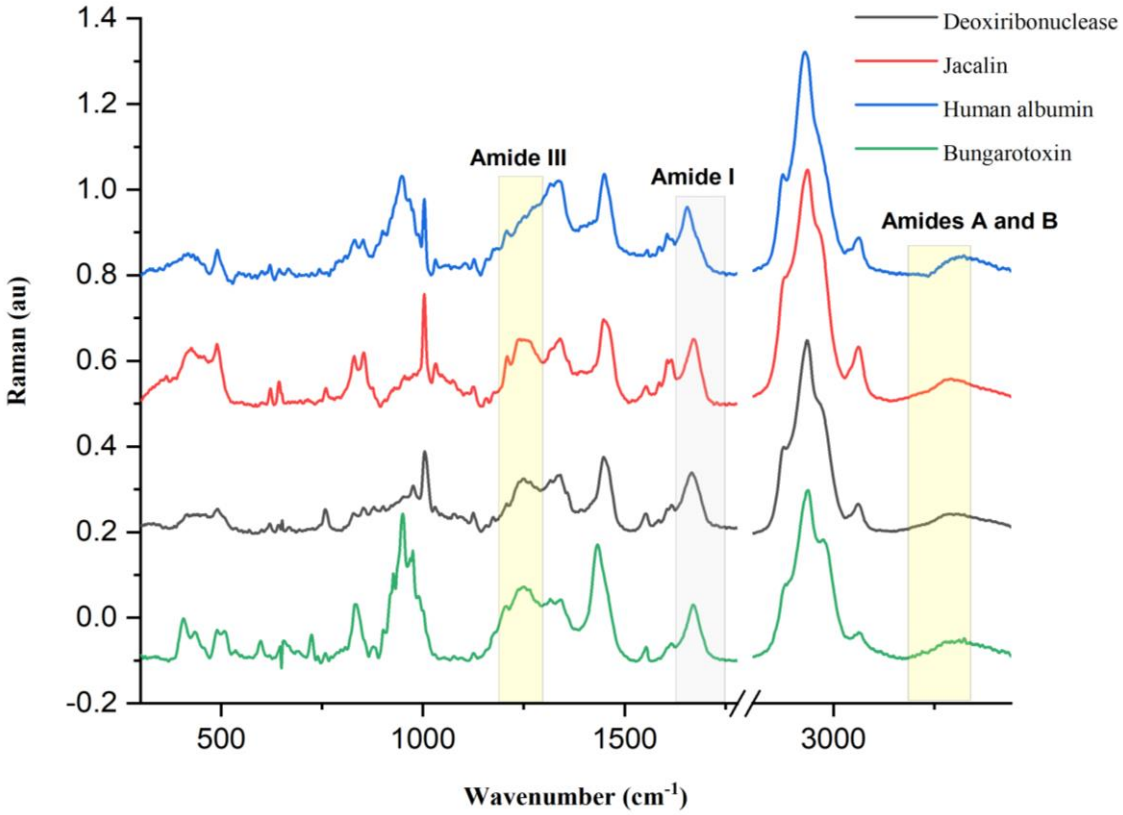

Figure 4.1. Raman spectra of Deoxyribonuclease, Jacalin, Human albumin and Bungarotoxin in solid state. The experimental procedure can be found below in the methods section.

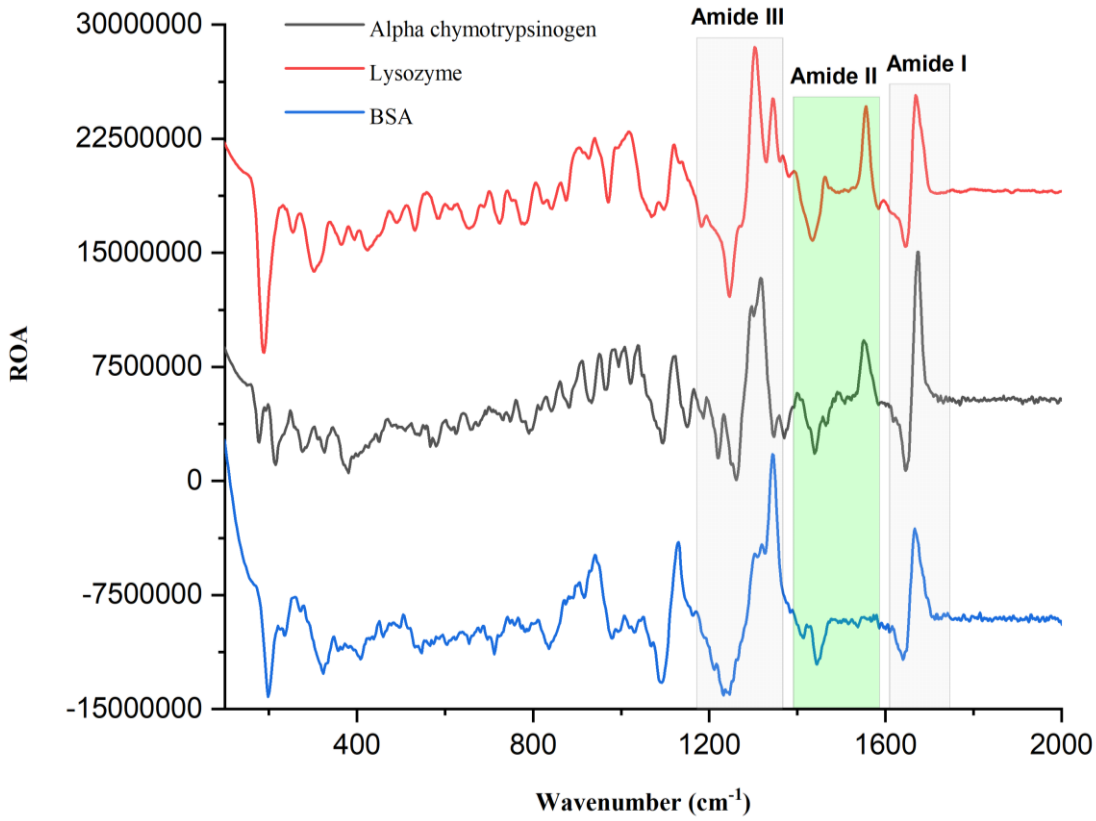

Figure 4.2. ROA spectra of α-chymotrypsinogen, Lysozyme and BSA in solution. The experimental procedure can be found below in the methods section.

4.2.2 Raman amide I and III bands

In 1972 Stynes and Ibers published a work on the Raman of glucagon in three forms (alpha, coil, beta) and found that the amides I and III maxima frequencies were consistent with the ones found earlier for alpha helical poly-L-alanine, irregular poly-L-glutamic acid and antiparallel beta-poly-L-glycine I[120]. More studies came after such as the one by Jen Yu and colleagues on the different conformations adopted by poly-L-lysine under different conditions of temperature, pH and ionic strength[107] used later on by Lippert and colleagues along with proteins with known SS (Lysozyme and Ribonuclease) to build up what was probably the first model to estimate SS based on Raman amides I and III[108]. More efforts were made over the years and collected in recent reviews[7,50,105] to establish more accurate assignments based on correlations between X-Ray and CD of proteins and homo polypeptides with their corresponding Raman amides I and III (Table 4-1). Theoretical studies based on Miyazawa's[98] *ab initio* calculations were conducted to explain the experimental characteristic frequencies for both IR and Raman and reviewed in 1986

by Krimm and Bandekart[50]. As in IR, the band position is sensitive to inter and intramolecular hydrogen bonding, through-bond coupling and also TDC[50].

Table 4-1. Combination of assignments from Rygula's et al[7] and Wen's[105] reviews.

Frequency (cm^{-1})	Secondary structure (SS)	Amide band
1650-1662	Helix	I
1670-1680	Sheet	I
1680	Turns	I
1640	Loose sheet/disordered	I
1272-1340	Helix	III
1227-1250	Sheet	III
1260	Disordered	III

Kitagawa shows in his review a table with characteristic positions for a set of proteins with predominant helix, sheet and non-regular structures[52]. For the amide I: 1652-1662, 1672-1674 and 1665-1668, respectively. The ranges for the amide III: 1264-1272, 1227-1242 and 1239-1254, respectively. These values seem to agree with the data presented in table 1 except for the non-regular under the assumption it means the same as disordered. Sane and colleagues published a table that gathered theoretical and experimental assignments from the literature[121]. The values for the helix (1652-1659), the sheet (1663-1672), turns (1689-1691) are in fair agreement with the values above but the irregular, referred to in this paper as 'unordered' and ranging from 1649 to 1656 cm^{-1}, clearly overlaps with the helix region. Other values reported for purely irregular proteins and peptides are the ones of poly-L-glutamic acid and glucagon: 1665 cm^{-1} (amide I), 1248 cm^{-1} (amide III) and 1248 cm^{-1} amide III respectively[120]. Another review by Miura et al generalizes intervals for the different structural components: helical structures for amide I and III, 1645-1655 cm^{-1} and 1260-1310 cm^{-1} respectively; sheet structures for amide I and III, 1665-1680

and 1230-1245 cm^{-1} respectively; and random coil ('irregular or disorder'), 1655-1665 and 1245-1270 cm^{-1} for amide I and III respectively[122] which like most of literature suggests unordered structures range between helix (~1655 cm^{-1}) and sheet (~1972 cm^{-1}). In this chapter, we measured the Raman spectrum of alpha bungarotoxin in solid state (random coil) and whose max was found at ~1670 cm^{-1} which is fairly close to the lower limit of the sheet interval. This probably reflects a reality that supposedly irregular structures actually adopt dynamic hydrogen-bonded structures.

4.2.3 Raman environmental markers

Many protein sidechains have Raman signals that depends on their environment. Some of the most useful to this work are discussed below.

4.2.3.1 Cysteine

Cysteine shows a band between 2500 and 2600 cm^{-1} corresponding to the S-H stretch whose position depends on the hydrogen bonding between S lone pair electrons and electron acceptors from other side chains or peptide bonds in the vicinity[119] (Figure 4.3 a). Furthermore, there is a correlation between the positions of the C-S and S-S stretch modes and the conformation of the Cys-Cys linkage in disulfide bridges[109] (Figure 4.3 b).

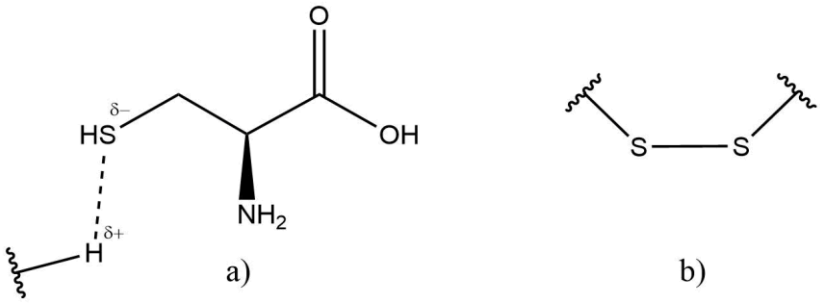

Figure 4.3. Hydrogen bonding between sulfydryl and electron acceptors (a). Disulfide bridge between two cysteines (b).

4.2.3.2 Methionine

Oxidized methionine shows two bands at ~1010 and ~704 cm^{-1} corresponding to the S=O stretch and C-S stretch respectively of the methyl sulfoxide group (Figure 4.4) that can be used as a marker for protein degradation[119].

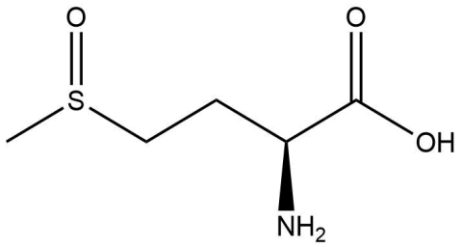

Figure 4.4. Sulfoxide methionine.

4.2.3.3 Histidine

Histidine shows a ring mode about 1275 cm^{-1} and a C=C stretch mode about 1570 cm^{-1} that shift upwards in wavenumber upon binding of imidazole ring to a metal[119].

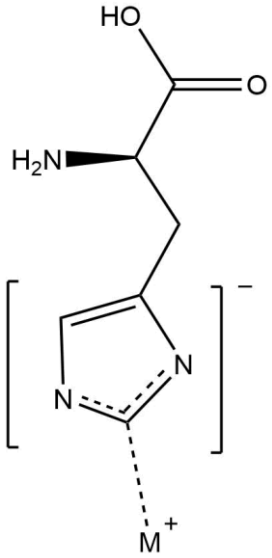

Figure 4.5. Metal binding histidine.

4.2.3.4 Tyrosine

Tyrosine shows a Fermi doublet at ~830 and ~850 cm^{-1} corresponding to the coupling of the breathing mode of the aromatic ring with a C-C-O deformation overtone[119]. The ratio between the intensities of those two peaks is known to provide

information on the phenoxyl group hydrogen bonding. Value of $I_{850}/I_{830} > 6$ are considered indicative of lack of hydrogen bonding and thus hydrophobicity[119].

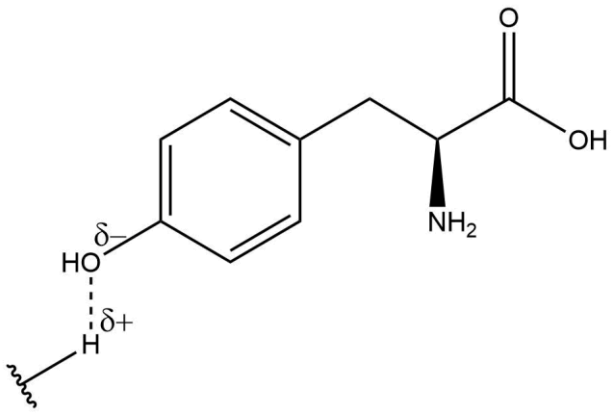

Figure 4.6. Hydrogen bonded tyrosine.

4.2.3.5 Tryptophan

Tryptophan has a characteristic band about 1550 cm^{-1} as a result of an indole ring vibration that is related to the torsion angle defined by χ between the plane of the aromatic ring and the peptide bond[109]. Equation 4.1 shows an empirical correlation between the torsion angle and the position (wavenumber) of this Tryptophan mode which is thought to be applicable within the range from 60° to 120° [109].

$$\tilde{\nu} = 1542 + 6.7(cos(3\chi) + 1) \tag{4.1}$$

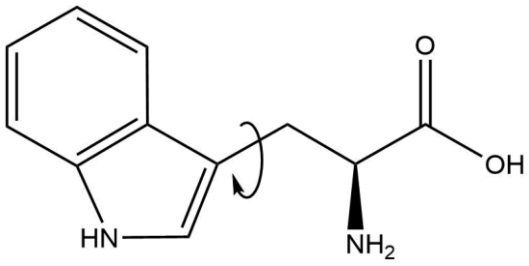

Figure 4.7. Tryptophan with an arrow pointing the bond about which the aromatic group rotates.

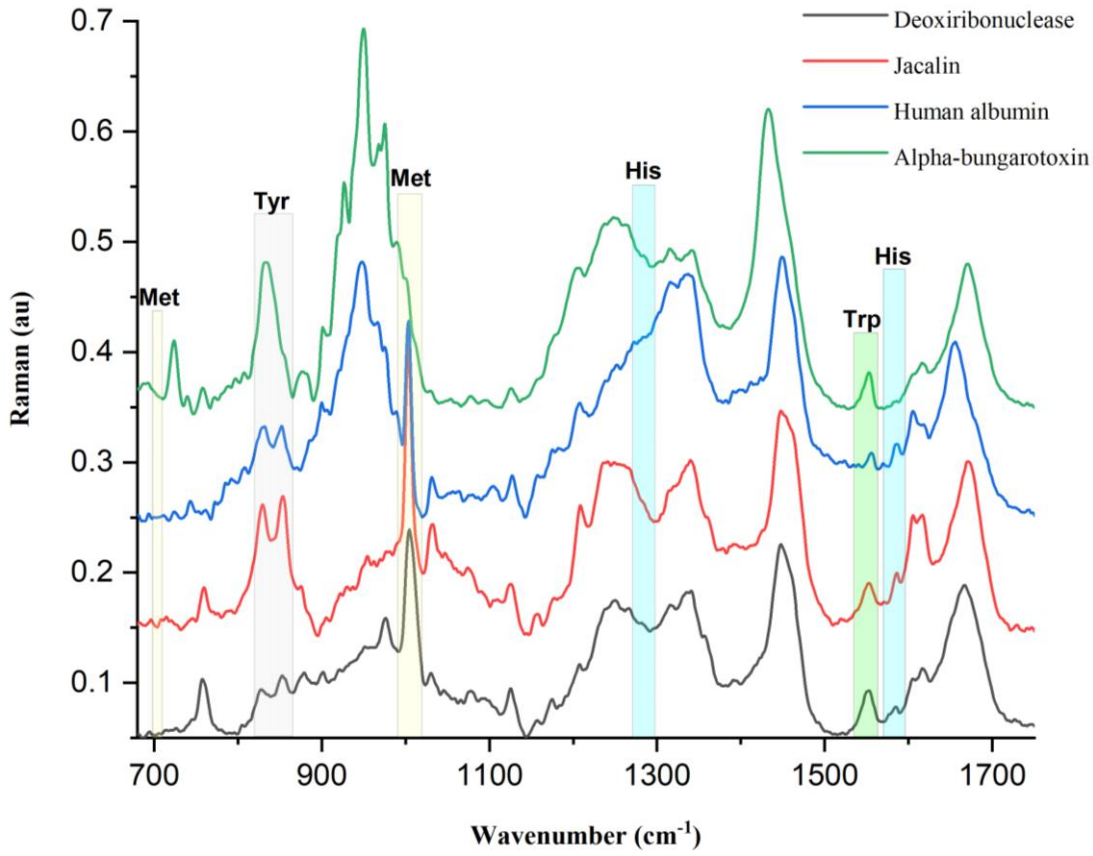

Figure 4.8. Tyr, Met, His and Trp most common environmental markers.

4.2.4 ROA amides I, II and III

The vibrations that give place to the amides I, II and III are optically active or equivalently, scatter right and left circularly polarized light with different strengths. Amides I and III are visible in both Raman and ROA but unlike in Raman where it can be seen only in resonance Raman, the amide II is perfectly visible in ROA within the region from 1510 to 1570 cm^{-1} [116] (Figure 4.2). The amide I is a couplet that consists of a negative and a positive counterpart[118]. The positive peak in helical proteins is prominent compared to that in sheet proteins[118]. The couplet has the minima at ~1638 and maxima at ~ 1662 in helical proteins (Human serum albumin) and is shifted to higher wavenumbers by ~10 cm^{-1} in beta-sheet proteins, e.g. ~1648 and 1677 cm^{-1} (Human IgG)[116]. Another example is that of equine Lysozyme, chicken Lysozyme and Alpha-lactalbumin with minima at ~1643 and maxima at ~1666 cm^{-1} [123] with a shift upwards in wavenumber compared to that of serum albumin that could be explained by the small content of beta structure in them. Although there are distinctive features for helical and beta proteins in the region of

the amide I, most of the efforts in the literature to assign SS to the ROA spectra of proteins focus on amide III because of its individual and well resolved peaks for the different SS contents. We collected from different references in the literature the most well-established assignments for the amide III (Table 4-2). These assignments were deduced from correlations between ROA of proteins and their corresponding X-Ray annotations (α, β, α+β, α/β). The peaks at ~1300 and ~1340 cm^{-1} in helical proteins were assigned to hydrophobic and hydrated helix respectively and their relative intensity to be informative of the relative amounts of helix hydrated and buried in non-polar media. However, recent publications pointed out contradictions in those assignments and suggested it might actually be the opposite, ~1300 cm^{-1} for hydrated and ~1340 cm^{-1} for hydrophobic[59].

Table 4-2. Band assignments for amide III.

Position (cm^{-1})	**Sign and band**	**Assignment**	**Proteins**
~1297-1305	+ Amide III	α-helix	(Lysozyme Chicken, Lysozyme Equine, Alpha-lactalbumin)[123]
~1340-1345	+ Amide III	α-helix	(Human serum albumin, Subtilisin Carlsberg)[116], (Lysozyme Chicken, Lysozyme Equine, Alpha-lactalbumin)[123]
~1264-1276	- Amide III	α-helix	(Lysozyme Chicken, Lysozyme Equine, Alpha-lactalbumin)[118,123]
~1244-1247	- Amide III	β-sheet	(Lysozyme Chicken, Lysozyme Equine)[123], (Ribonuclease, Subtilisin and

			IgG)[116]
~1296	+ Amide III	β-turn	IgG[116]
~1347	- Amide III	β-turn	IgG[116]
~1376	- Amide III	β-turn	IgG[116]
~1315-1320	+ Amide III	PPII	Beta-casein[116], poly-L-Ala[124],[123],[125]

4.2.5 ROA side chains

The most distinctive side chains are Tryptophan at ~1545-1560 cm^{-1}, which as in Raman can be related to the torsion angle, phenylalanine and its breathing ring mode at ~1000 cm^{-1} ,Tyrosine and other aromatic groups within ~1600 and 1630 cm^{-1}, CH_2 and CH_3 deformations of side chains resulting in a -/+ couplet (low/high wavenumber respectively) about ~ 1400-1480 cm^{-1} and a weak couplet from the N-H deformation of the indole ring in Tryptophan but with the signs inverted, +/- for low and high wavenumber respectively[116,123,126].

4.3 MATERIALS AND METHODS

4.3.1 Samples and reagents

All the proteins used were purchased from Sigma Aldrich and any dilution and blank measurement carried out with Milli-Q (ultra-pure) water. The number of proteins measured by Raman in aqueous state, Raman in solid-state and ROA in aqueous state respectively were 17, 32 and 14 (Appendices-D).

4.3.2 Instrumentation

Centrifuge sigma D-37520 14k, Millex-GV 0.22 µm syringe filter disks, syringe, Ika vortex genius 3, marble mortar, and a Jasco 660 UV-Vis, Jasco 1500 CD, Thermo-fisher DXR2 Smart Raman, Jasco FP 6500 and BioTools ChiralRAMAN-2X™ ROA spectrometer.

4.3.2.1 Thermo-fisher DXR2 Smart Raman instrument

The instrument used for the Raman experiments was a Thermo-fisher DXR2 Smart Raman Spectrometer (Figure 4.3.2.1 b) equipped with a He laser of up to 8 mW output power, 633 nm filter, a diffractive monochromator (grating) and a charge coupled device (CCD) detector. The sample holder sits on an adaptable unit that can be moved up, down, left, right, for and backward (Figure 4.9 a). Before hitting the sample, the laser goes through a lens that focus the beam on the sample, which is the reason why it is critical to adjust the relative distance between the holder and the objective in order to maximize the amount of light hitting the sample and the amount of scattered light collected by the objective. The light scattered back into the objective is passed through an edge filter to remove the excitation wavelength and then directed into a grating unit to extract the different wavelength components before going into the CCD for its detection (Figure 4.10).

Figure 4.9. Picture of the DXR Smart Raman Spectrometer: sample holder (a) and outside (b)

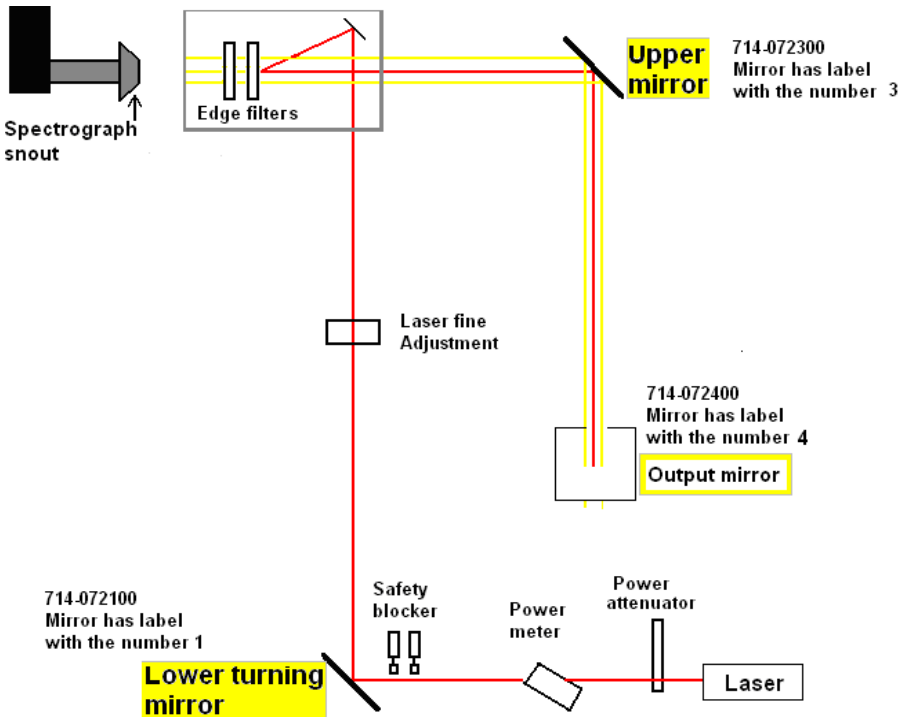

Figure 4.10. Diagram of the Raman instrument used in the experiments here reported.

4.3.2.2 BioTools ChiralRAMAN-2XTM ROA

The instrument used for the collection of Raman and ROA spectra of proteins in aqueous state was a BioTools ChiralRAMAN-2X^{TM} ROA spectrometer equipped with a 532 nm Nd-YAG laser of up to 2000 mW and a CCD detector. The light from the laser is passed through lenses that reduce the cross section and collimate the beam and then through a linear polarizer. Next, the linearly polarized beam is passed through linear rotators (half wave plates) that scrambles light to produce non-polarized light that passes through another lens that further collimates and focuses the beam on the sample. The scattered light is passed through another linear rotator and then through an edge filter that removes the excitation wavelength before passing through a liquid crystal retarder (quarter wave plate) that creates circular polarized light and whose *s* and *p* components will be split by a dichroic mirror into the two fibre optics for separate detection in the CCD (Figure 4.11).

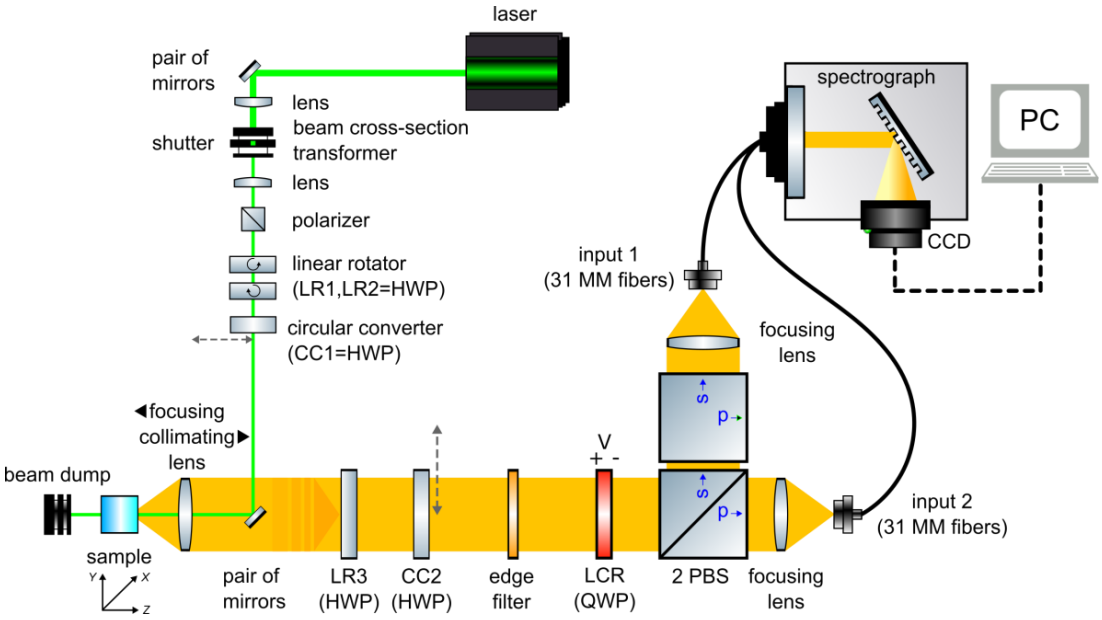

Figure 4.11. Scheme of BioTools ChiralRAMAN-2X™ ROA spectrometer.

4.3.3 Experimental procedure

4.3.3.1 Quenching of background fluorescence by photobleaching

Many studies in the field of molecular analytical science are carried out with biological samples treated to extract a variety of components e.g. DNA, RNA, lipids and proteins[55,66] Although these molecules are put through different techniques, e.g., ion exchange chromatography, affinity chromatography, gel filtration, gel electrophoreses, to separate them from one another and other cellular components, trace concentrations of these impurities often remain in the sample with the molecule of interest, proteins in our case[66,127]. Because Raman spectrometers measure all the light at the Stokes shifted frequency that reaches the detector whether it is from Raman or other sources, and because fluorescence has a yield several orders of magnitude larger than that of scattering even the smallest number of chromophores in the sample cause significant background fluorescence[55]. Among the main effects that the interference of fluorescence causes in a Raman measurement are the degradation of the signal-to-noise ratio because of the shot noise in the detector, the rise of a broad Gauss-like shaped baseline that complicates the spectra interpretation and the overload of the CCD camera resulting in loss of linearity in the instrumental response and complete masking of Raman peaks when the cut off is reached[56,114,128]. Fluorescence can be quenched by different molecular interactions such as energy transfer, ground-state complex formation and collisional quenching[129,130]. They are

based on different physical-chemical interactions between a quenching agent and the fluorophore that results in a lower fluorescence quantum yield and so fluorescence intensity[129]. These interactions can be reversed by removing the quenching agent, generally a chemical with certain properties, e.g., I^-, acrylamide in the case of collisional quenching[129].

Another way to supress fluorescence is by means of excited state reactions and molecular rearrangements, that result in irreversible degradation of the fluorophore and occurs when exposed to large amounts of light in a phenom referred to as photobleaching[115]. It is because of its irreversible character that it is not considered among the quenching mechanisms aforementioned but separately. Photobleaching is the most extensively used fluorescence supressing method in Raman spectroscopy because of its simplicity. The extent of the bleaching or in other words the number of bleached molecules, depend on the time exposed to the laser, the laser power and the excitation wavelength used[115,131]. Although the molecular pathways by means of which photobleaching occurs are not well known, some authors have suggested a two-photon excitation caused by a multiplicity exchange between molecular oxygen and the chromophore with the latter undergoing an intersystem crossing might be the main reason behind. Figure 4.12 shows the decrease of background fluorescence of a Beta-lactoglobulin sample in solid state over time due to the exposure to the instrument laser (633 nm and 8 mW output power).

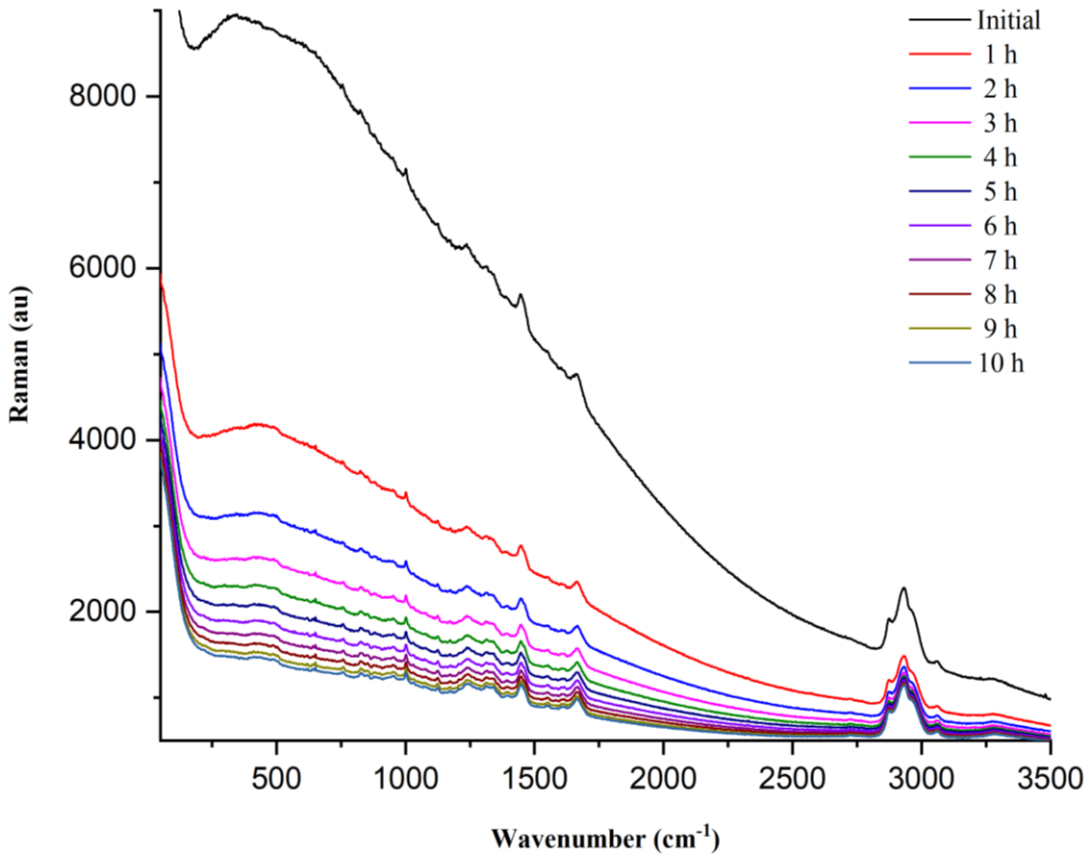

Figure 4.12. Raman spectra of Beta-lactoglobulin in solid state photobleached over 10 h. The Raman spectra were collected every hour with 8 mW, 10 s exposure and 10 accumulations. More details of the experimental procedure can be seen in the data acquisition section below.

The experiments on the proteins in solid state were carried out with the DXR spectrometer which consists of a fairly large power for Raman measurements but too low for photobleaching (~8 mW). A macro was written to perform cycles of photobleaching and subsequent measurement. The number of cycles was decided in accordance to the initial levels of fluorescence and how the sample responds to few minutes of exposure. With the same power, some samples decreased in fluorescence much faster than others which suggested the wavelength did not always match the maximum of absorption of the chromophore or some impurities were more stable to light-induced degradation and thus the chemical nature of the impurities changes from protein sample to protein sample. It was found that for some of the proteins it was a matter of minutes for the background fluorescence to significantly decrease, whereas some others required longer exposures ranging from 1 to 12 hours and several days in some extreme cases. It is because of those troublesome proteins that we home-made a photobleaching unit out of industrial diode lasers with an output power between 150 and 180 mW and 635 and 532 nm wavelength (Figure 4.13).

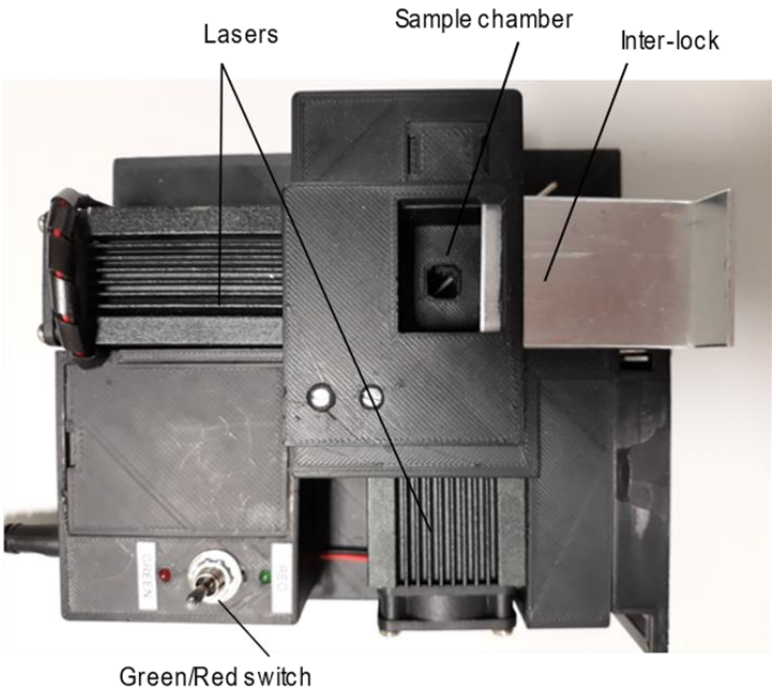

Figure 4.13. Home-made photobleaching unit equipped with a green and a red diode laser.

4.3.3.2 Data acquisition

The procedure used for the spectra acquisition of both reference set and test samples were the same so no distinction will be made in the description of the operational procedure.

4.3.3.2.1 Raman data collection in solid state

About 10 mg of the protein powder were grounded in a mortar and then introduced into a 4x4x15 m^3 (width + thickness + length) Starna quartz cuvette.

The measurements in solid state were done with the DXR Smart Raman spectrometer described in the instrumental section. Since the maximum photobleaching time in the instrument set up was only 1 hour, it was necessary to create a macro consisting of several cycles of photobleaching and subsequent measurement. Usually one short cycle was performed to check how fast the fluorescence decreased and in consequence the total number of cycles necessary to minimize the fluorescence was estimated. Although the instrumental photobleaching proved enough with some of the samples it was not for many others with higher levels of fluorescence. In such cases, we photobleached externally with the home-made photobleaching unit described above over periods ranging from 1 to 12 h.

Although more effective due to the choice of wavelength and higher power, photobleaching externally means moving the cuvette around which might result in mixing of the powder and 'loss' of photobleached sample.

The number of accumulations used for the measurements ranged between 1500 and 9999 (the maximum the instrument allows), chosen in accordance to the levels of fluorescence and noise and the exposure time used was 10 s. The power was always set to the maximum possible output which dropped from ~8 in early experiments to ~6 in the late ones due to the lifespan of the He laser. Alignments of the laser beam and calibration of the grating and CCD were done weekly since the first affects the actual power at sample position and the second the peak positions. Pinene was used as a standard to assess the goodness of peak position and need for calibration. The background subtraction option selected was the 'smart background correction' which averages measurements taken in regular intervals of time throughout a pre-set period of time (about a week in our case). Another critical parameter in the set up was the position of the sample holder which can be moved along the 3 spatial coordinates (side to side, up and down, for and backward). The side to side and up and down positions were optimized to centre the sample in the beam which we usually did by looking directly at the position of the light spot on the cuvette. The forward and backward adjustments relate to the focal length of the convergent lens which we optimized by first lifting the holder to make the beam pass through the bottom of the cuvette and then either manually or automatically moved it forward and backward until the maximum scattering from the cuvette quartz was reached (Figure 4.14). Once the optimal position was found, the holder was put back down for the measurement.

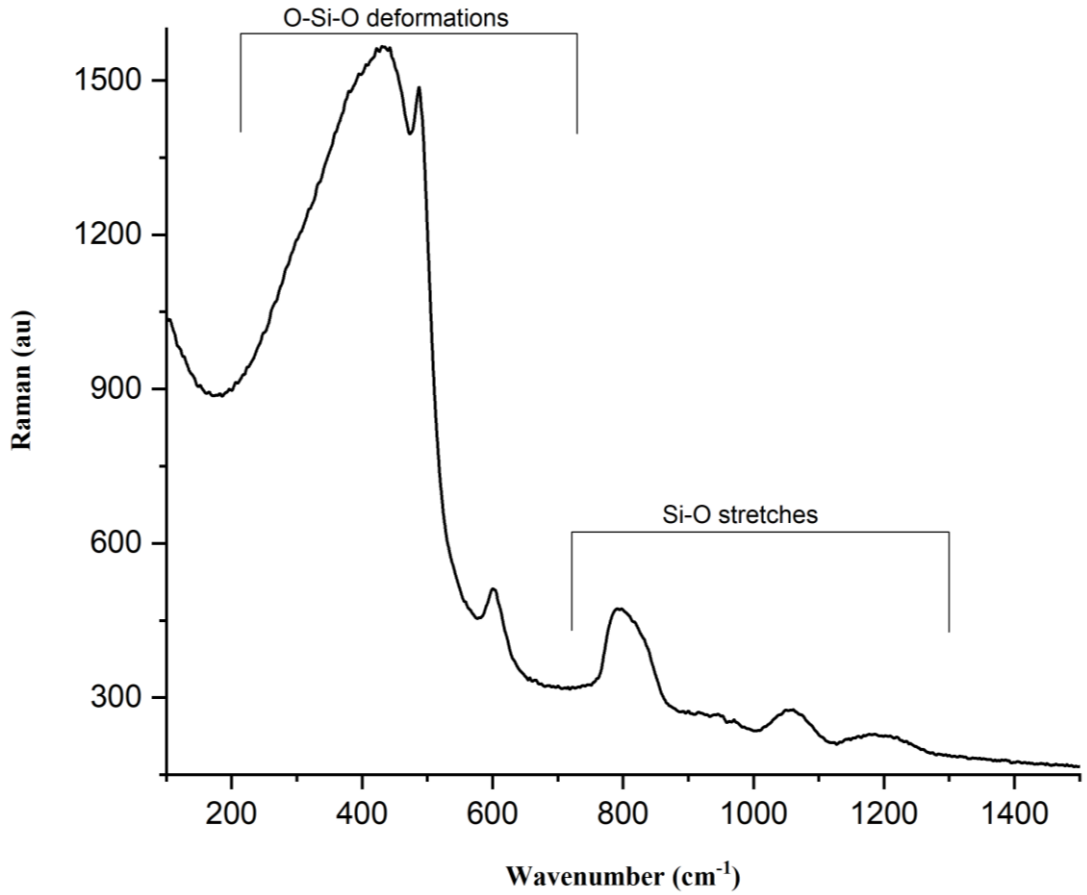

Figure 4.14. Raman spectrum of cuvette quartz.

4.3.3.2.2 Data collection of Raman and ROA in aqueous state

The proteins were dissolved in water in concentrations ranging from 40 to 80 mg.ml^{-1} which depended mostly on the amount of protein available and its solubility. After, they were centrifuged for 5 min at about 10k rpm to remove any undissolved protein. The eluent was recovered and filtered with a Millex-GV disk filter of 0.22 µm diameter to further remove any solid particles in suspension. A small volume of the stocks was diluted down to 1 mg.ml^{-1} for concentration determination by UV-Vis measurement. About 40 µl of the stock solutions were used for the Raman/ROA measurements. The instrument used for all the Raman and ROA experiments in aqueous state was the BioTools ROA spectrometer described above. A few samples did not require photobleaching, but the ones that did were exposed to the laser for periods that depended on the amount of fluorescence displayed. Although it is true that the rate at which fluorescence is depleted depends on the incident power, we soon learnt that increasing the power with no mechanism to refrigerate the sample will often leads to its calcination. Because of this, we decided to photo-bleach mild

fluorescent samples with up to 1000 mW for about 3-5 hours whereas other samples that fluoresced intensely with powers between 300 and 500 mW for up to 24 hours.

There are three halt criteria in the BioTools ROA: halt after a pre-set number of scans, halt after a pre-set amount of time and halt on convenience by pressing 'halt'. Although very high-quality Raman spectra were obtained within the first minutes of a measurement, many accumulations were necessary to improve the signal-to-noise ratio (S/N) up to 'acceptable' standards for ROA. Because of this, we decided not to set any number of accumulations nor amount of time but just stop the measurement when we considered it to be good enough in terms of S/N. The power used for the measurements ranged from 800 to 1000 mW and the exposure time between 0.36 and 1.02 s, always seeking to maximize the signal without saturating the detector.

The proteins exposed the longest to the laser (photobleaching + data acquisition) were measured before and after the Raman/ROA measurement by CD and Fluorescence to ensure no unfolding took place during the experiment.

The instrument alignment and calibration are not automated. +/- pinene need to be measured on regular basis to assess the performance of the instrument and decide when the alignment and calibration are due. The alignment involves the optics that define the path of the beam up to the sample, the position of the sample holder, the position of the fibre optics and the position of the camera; and seeks to maximize both the power that incises on the sample and the collection of the corresponding back-scattered light. The intensity and position calibration are done using the fluorescence of a reference material and the lines emitted by a Ne lamp as a reference. The detailed superuser procedure for alignment and calibration of the instrument exceeds the pretentions of this chapter and thus will not be discussed here.

4.3.4 Data processing

The solid state spectra were baseline-corrected in Origin[80] by subtracting a fit to the spectra in points assumed to be only due to fluorescence that were interpolated with cubic polynomials in the range between 1508 and 3500 cm^{-1} (Figure 4.15). The water vibrational modes discussed in chapter 3 are also Raman active but the O-H bending mode is far less intense than in IR (Figure 4.16) so there was no need for

water subtraction from the proteins in solid state although we did have to subtract it from the ones in solution (see below). As discussed in chapter 1, the Raman intensity depends on the incident power, the Raman cross-section and the concentration[132]. Moreover, the signal recorded by the instrument is proportional to the total exposure time. In that case, the total Raman signal of an aqueous protein measured by the instrument can be expressed as

$$I^{pw} \propto I_o^{pw} \cdot t^{pw}\left(\sigma^p \cdot C^p + \sigma^w \cdot C^w\right) \tag{4.2}$$

where σ is the Raman cross-section, C means concentration, t stands for the total exposure time, I_o is the incident power, pw stands for protein in water, p means protein and w stands for water. For just the water the absorption is

$$I^w \propto I_o^w \cdot t^w \cdot \sigma^w \cdot C^w \tag{4.3}$$

and so, the Raman signal of the protein under consideration is

$$I^p = \frac{I^{pw}}{I_o^{pw} \cdot t^{pw}} - \frac{I^w}{I_o^w \cdot t^w} \propto \left(\sigma^p \cdot C^p + \sigma^w \cdot C^w\right) - \sigma^w \cdot C^w \tag{4.4}$$

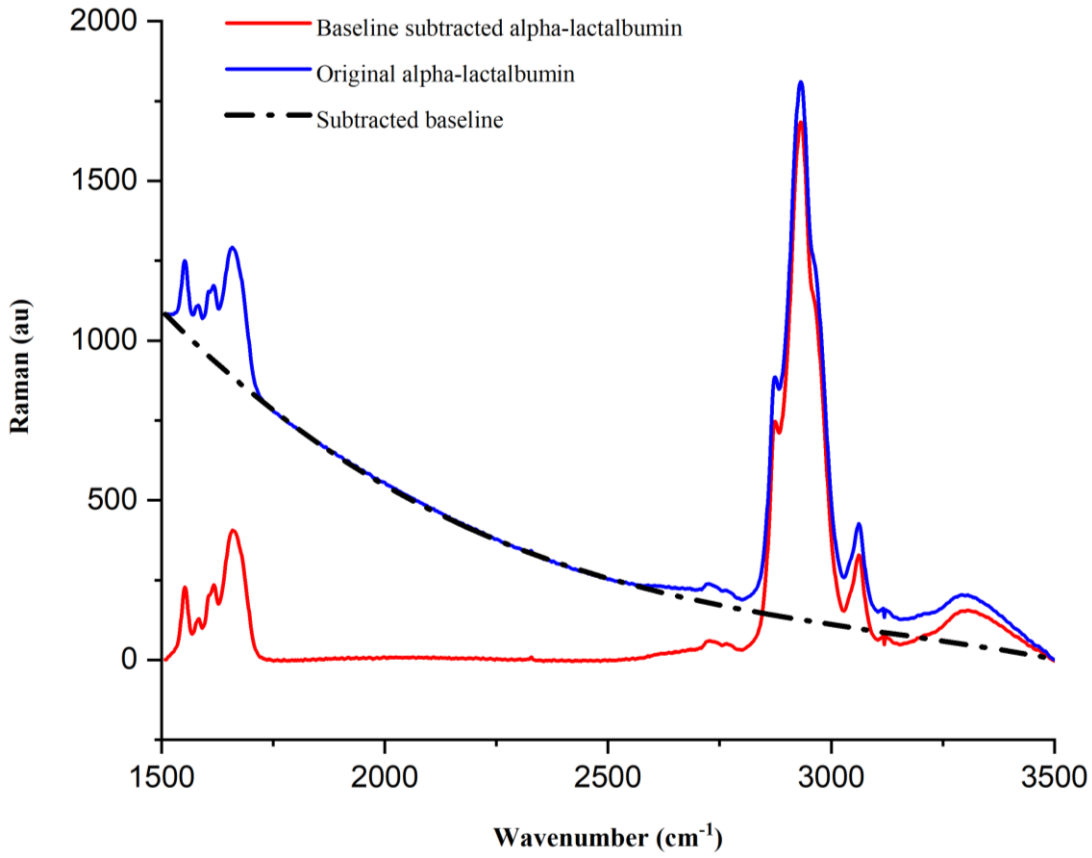

Figure 4.15. Baseline correction of alpha-lactalbumin in solid state. A cubic polynomial was fitted to points between 1508 and 3500 cm^{-1} and subtracted.

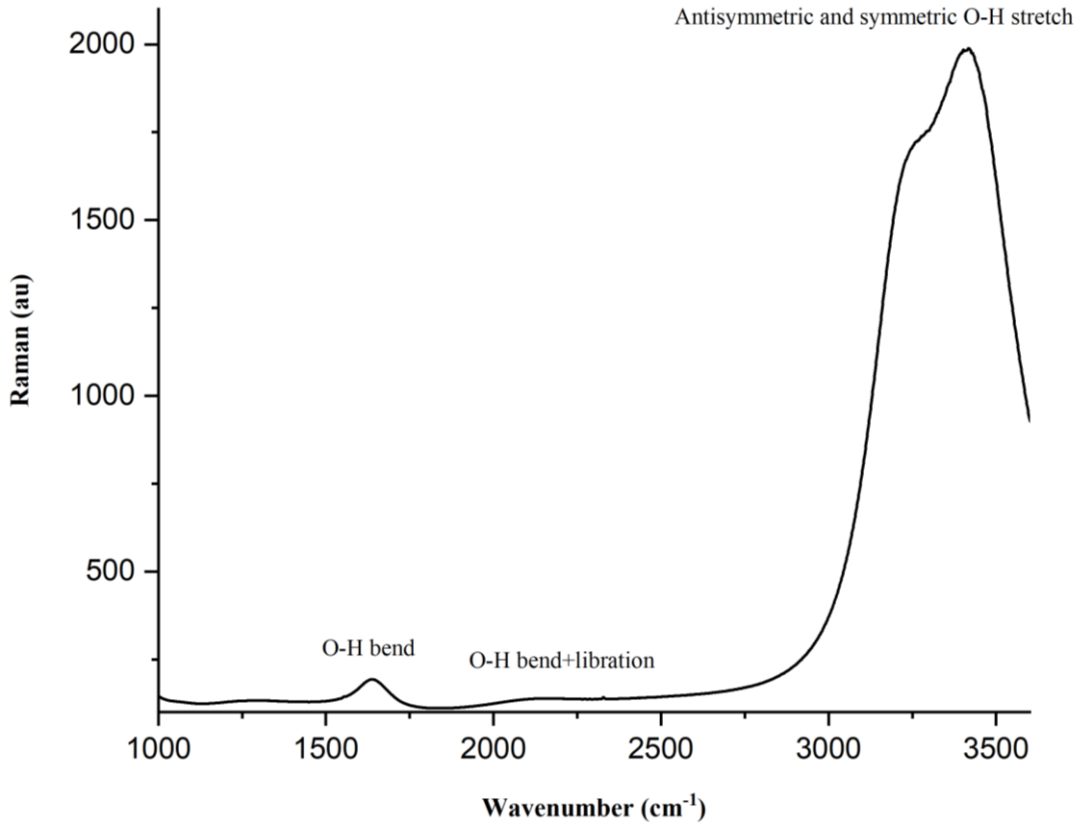

Figure 4.16. Raman spectrum of water collected with the DXR smart Raman spectrometer with 100 scans and 10 s exposure.

Although the instrument had an option to export accumulated data, we chose to export the data files every 5 min to have independent through-time spectra and do the average ourselves as a part of the post-processing. Since the exposure times and powers used differed from measurement to measurement, the spectra had to be further scaled by dividing the Raman and ROA signal by the power and the exposure time -as expressed in Equations 4.2-4.4- in order to make them comparable.

Due to fluorescence, the baselines of the different measured proteins were not comparable and thus a baseline correction was necessary. The baseline correction is probably the most challenging part in any Raman spectra post-processing. Most of the existing methods are based on interpolation and subtraction of the baseline but they defer in the approach used to select the points that define the baseline (manual or automatic) and the method to interpolate between them (splines or fitting)[56]. The procedure used here consisted of selecting certain points across the spectrum that we believed to be only due to fluorescence, interpolate them with cubic splines for subsequent subtraction. For the proteins in aqueous state, a MATLAB[79] code was written to ease some of the operations involved which include fitting and subtraction

of the baseline, subtraction of the blank and scaling and trimming of the data (Appendices-A.4).

The ROA spectra were smoothed by Savitsky-Golay with a 3 points window and a grade 1 polynomial. After they were zeroed by subtracting the average of the values in the range from 1750 and 1850 cm^{-1}. Finally, the spectra were scaled to account for the different incident powers, exposure times and concentrations, and expressed in MRW (Equation 4.5). All the processing was carried out manually in Origin[80].

$$ROA^{scaled} = \frac{ROA^{original} \cdot MRW}{t \cdot I_o \cdot C} \tag{4.5}$$

where *MRW* means mean residue weight $MW/(n-1)$, *t* is the total exposure time in min, I_o stands for incident power and *C* means concentration. *MW* means molecular weight.

4.4 RESULTS

4.4.1 Photobleaching experiments

As explained above, it was necessary to expose the sample to the laser before the measurement to reduce the background fluorescence as much as possible. Several time course measurements were carried out with different powers to assess how significant the differences in the rate decays with powers between 500 and 900 mW were (Figure 4.17).

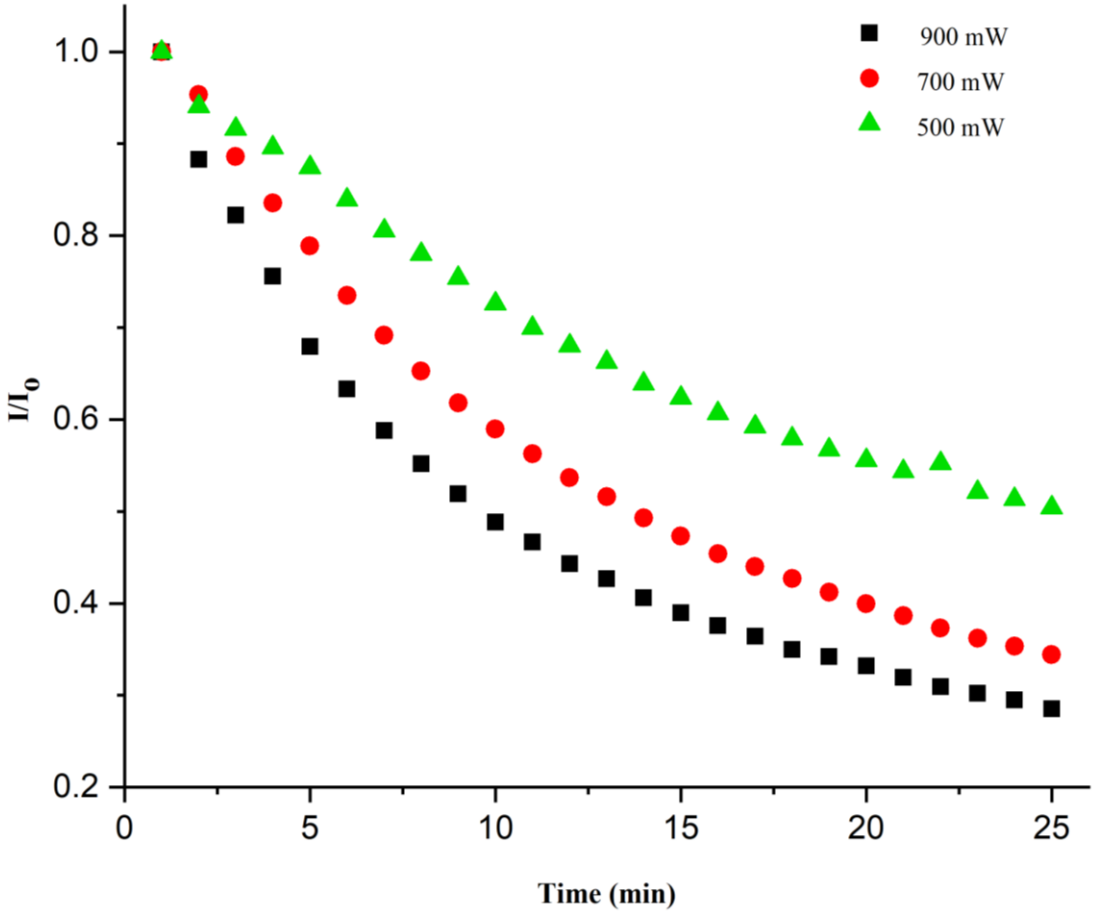

Figure 4.17. Relative Raman intensity at 838 cm^{-1} (attributed only to fluorescence) of a 65 $mg.ml^{-1}$ BSA solution with different laser powers over exposure time with 532 nm excitation. High powers seem to speed up the photolysis of the fluorophore. The measurements were accumulated and exported every 5 min.

The time course measurements in Figure 4.17 were carried out with no stirring. Although the trends seem to be well defined in the time interval showed in the figure, deviations from the exponential-like behaviour were observed when wider time windows were recorded. Those deviations disappeared when stirring the sample all the way through the experiment (Figure 4.18).

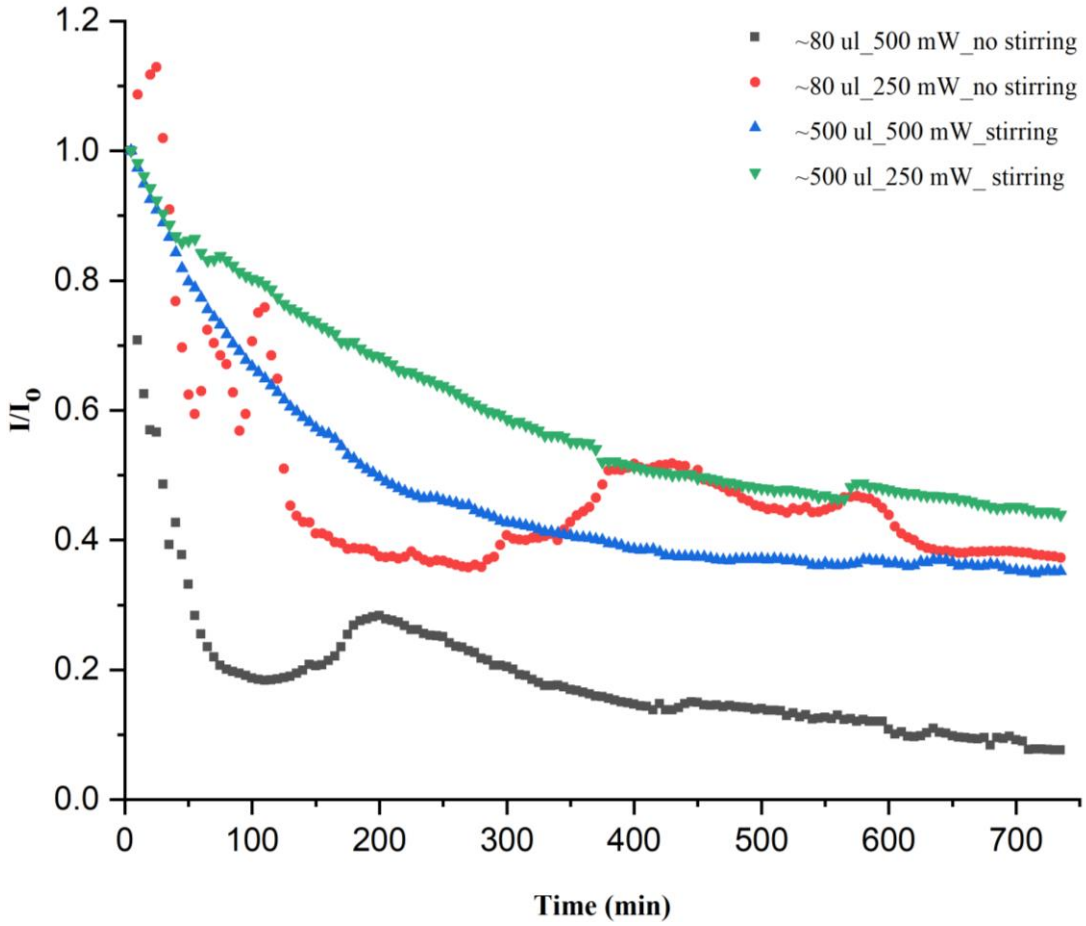

Figure 4.18. Photobleaching of ~80 ul at 500 and 250 mW and ~500 ul at 500 and 250 mW with and without stirring. The stirring was achieved by recirculating the sample through a polymer home-made mask by means of a HPLC pump that required 0.5 ml to fully fill the circuit.

Another thing to consider was whether there is any thermal degradation taking place on top of photolysis. The fluorescence decays in Figure 4.17 can be expressed in terms of exponentials with different decay rate constants. If only photodegradation occurred and assuming the background was only due to fluorescence, the data should be linearized by taking the logarithm. The contrary could be indicative of both photo and thermal degradation[133]. In Figure 4.19, it can be seen 500 mW is the closest to a linear curve, followed by 700 and 900 mW which is consistent with presence of thermal degradation as it increases with power.

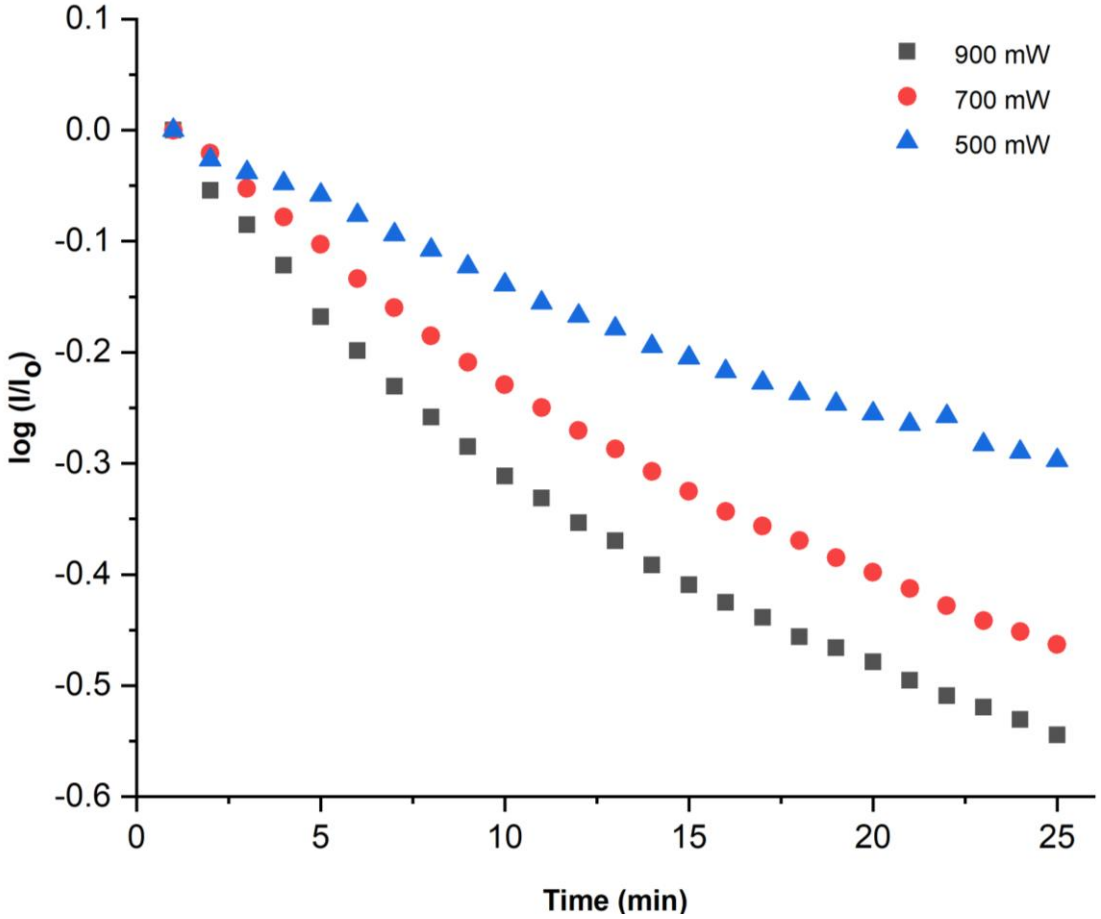

Figure 4.19. I/I_o of 900, 700 and 500 mW photobleaching curves in log scale.

Furthermore, the 900 mW curve from Figure 4.17 was fitted firstly with a one exponential decay model (Equation 4.6) and secondly with a two exponential one (Equation 4.7), shown in Figure 4.20 (the fitting of the one exponential model decay can be found in Appendices-C). The R^2 value and the distribution of the residuals (Figure 4.21) indicates that the data is better fitted with two exponentials than by only one which could be a confirmation of thermal degradation taking place along with photodegradation. Alternatively, it could also be indicative of the presence of more than one fluorescent impurity. Similar analyses were performed on the 700 and 500 mW (not shown) resulting in better fittings with one single exponential model.

$$y = y_o + A_1 \cdot e^{-x/t} \tag{4.6}$$

$$y = y_o + A_1 \cdot e^{-x/t_1} + A_2 \cdot e^{-x/t_2} \tag{4.7}$$

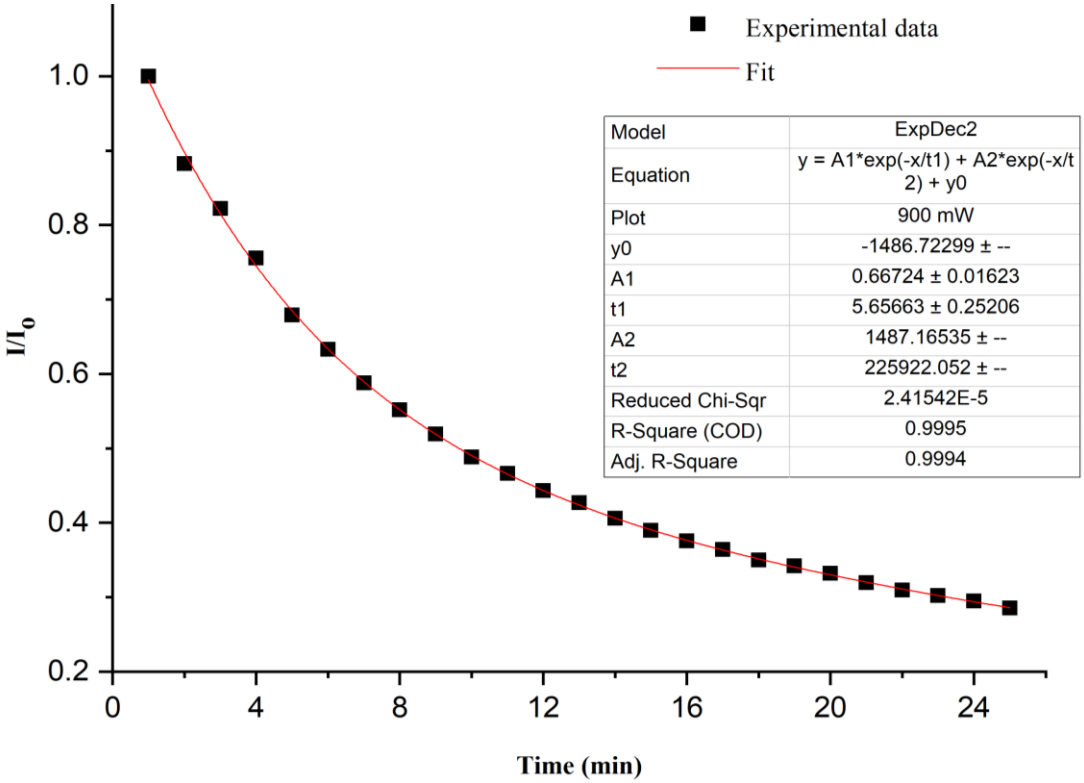

Figure 4.20. Fitting curve plots of 900 mW photobleaching curve with a two-exponential model. The algorithm used for the fitting was Levenberg Marquardt.

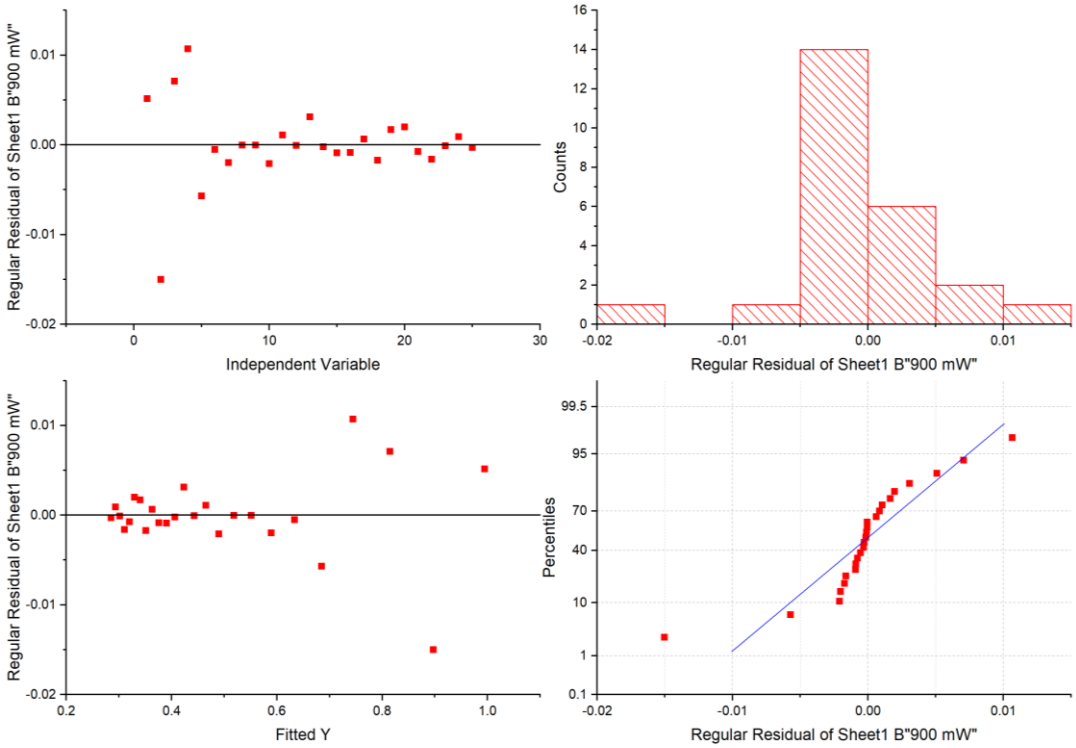

Figure 4.21. Residuals plots corresponding to Figure 4.20.

On a separate note, photobleaching is based on irreversible chemical reactions as stated above. However, after waiting for some time, it was found the intensity of the peaks often recover giving a false impression of being reversible. This could be explained by the fact that the small cross section of the laser only illuminates a small portion of the bulk solution during the experiments and that impurities in the surroundings diffuse into the illuminated area long after the end of the exposure resulting in apparent recover of fluorescence.

To test this hypothesis, an experiment was performed in which the intrinsic fluorophores of BSA were exposed to the lamp for long periods and the fluorescence spectrum measured in regular periods of time first with no mixing (Figure 4.22) and then in a separate experiment with only a mixing between measurements (Figure 4.23).

Although the fluorescence increased again after mixing up the sample, it did not go back to the initial value but a fraction of it.

Another question that arose was what wavelength is the most suitable for photo-bleaching. The first approach was to shine with the same wavelength that was to be used in the Raman experiments in order to destroy all the fluorophores that could potentially be excited during the experiments. However, the efficiency of the process must be dependent on the maxima of the extinction coefficient of the fluorophore. Some proteins showed massive fluorescence that faded away after short periods of laser exposure whereas some others, even despite not displaying that much fluorescence, required very long exposures to reduce it to a similar extent.

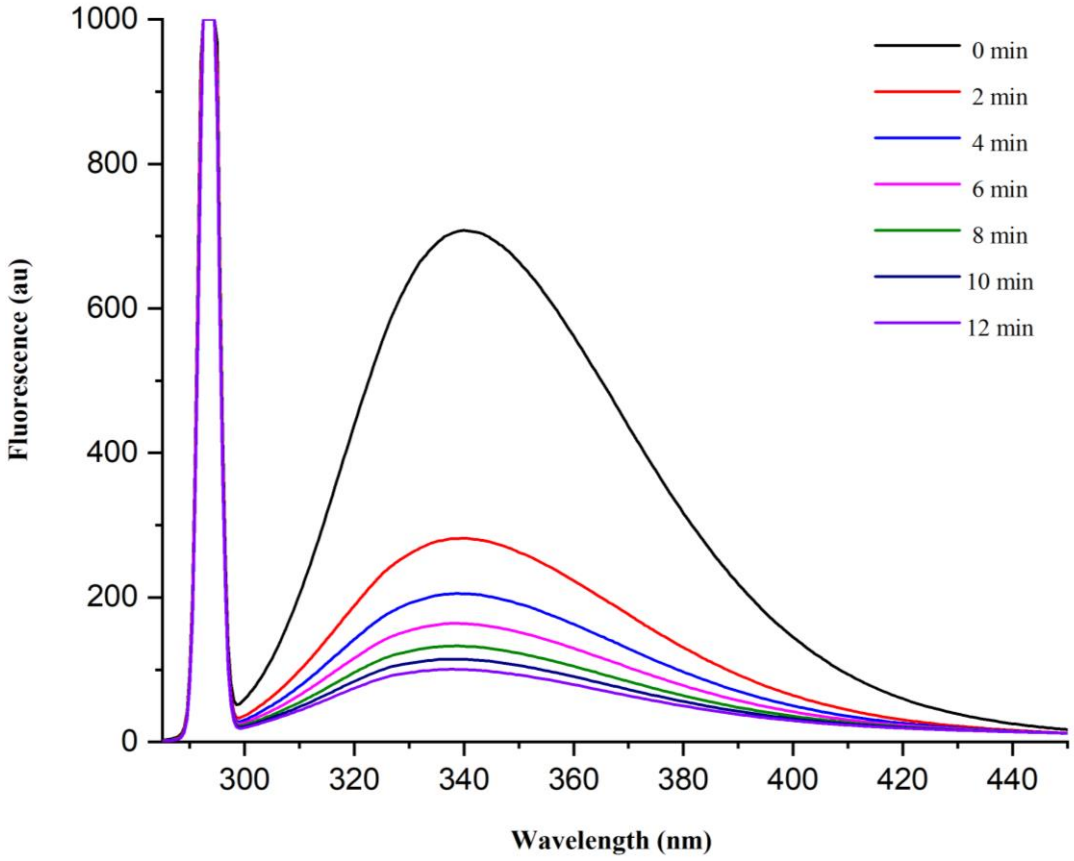

Figure 4.22. Fluorescence spectra of a 1mg.mL^{-1} BSA solution recorded at regular time intervals throughout continuous light exposure. The sample was exposed to the lamp by leaving the incident shutter open between measurements (~2 min). The measurements were carried out with a 3 nm resolution aperture to minimize the exposure during the measurements whereas the photobleaching was done with an aperture equivalent to a 10 nm resolution to increase the amount of irradiation. The excitation wavelength used was 295 nm (tryptophan) and the pathlength used for the experiment 4 mm.

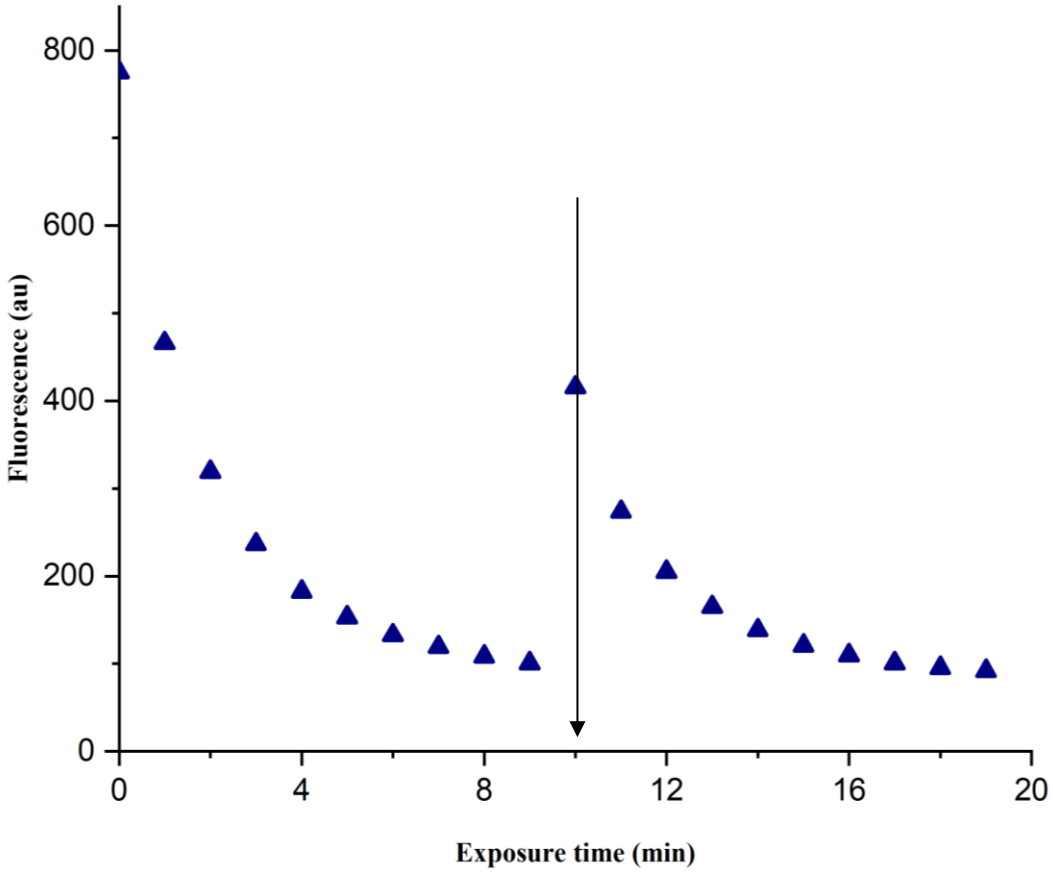

Figure 4.23. Fluorescence Recovery After Photobleaching (FRAP) of a 1mg.ml^{-1} solution of BSA in water. The measurements were carried out with a 4 mm pathlength and 3 nm of resolution while the photobleaching with an aperture equivalent to 10 nm. The excitation wavelength was 295 nm (Tryptophan) and the emission set for the one found for BSA in the experiments before, 340 nm. The sample was stirred with a mixer after 10 min of exposure in order to homogenize the solution.

4.4.2 Raman reference set in solid state

A reference set of proteins was collected to train the SOM algorithm as shown in the previous chapter. No presence of water was noticed in any of the spectra, so no water subtraction was necessary. Only a baseline correction was performed as explained in the materials and methods section. The baseline corrected data were scaled by the interval method and trimmed between 1625 and 1800 cm^{-1} (Figure 4.24).

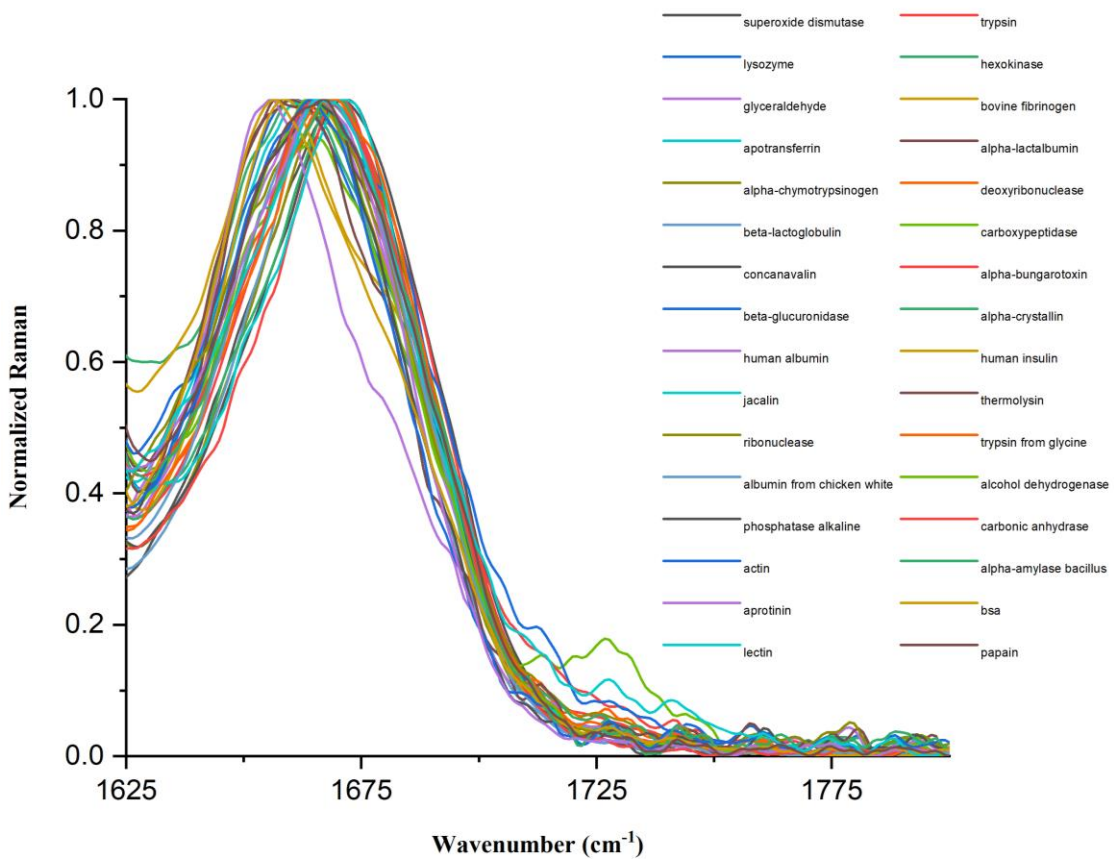

Figure 4.24. Raman spectra of 32 proteins in solid state. The spectra were normalized by the interval method and trimmed within the region of the amide I.

Exploratory analyses as those performed on solid IR data in chapter 3 were also applied to Raman solid state to determine the correlations between SS and the amide I band. The wavenumber corresponding to the max of the peaks were plotted firstly against helical content (Figure 4.25) and secondly against beta sheet (Figure 4.26).

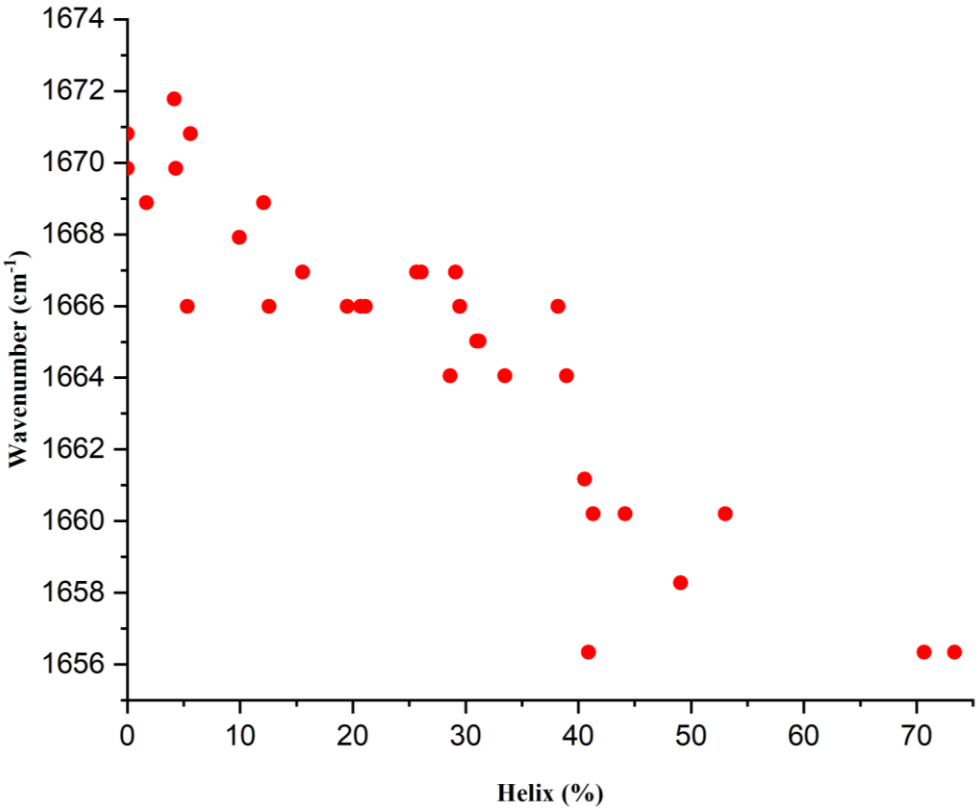

Figure 4.25. Wavenumber corresponding to the maximum of the peaks against their helical content.

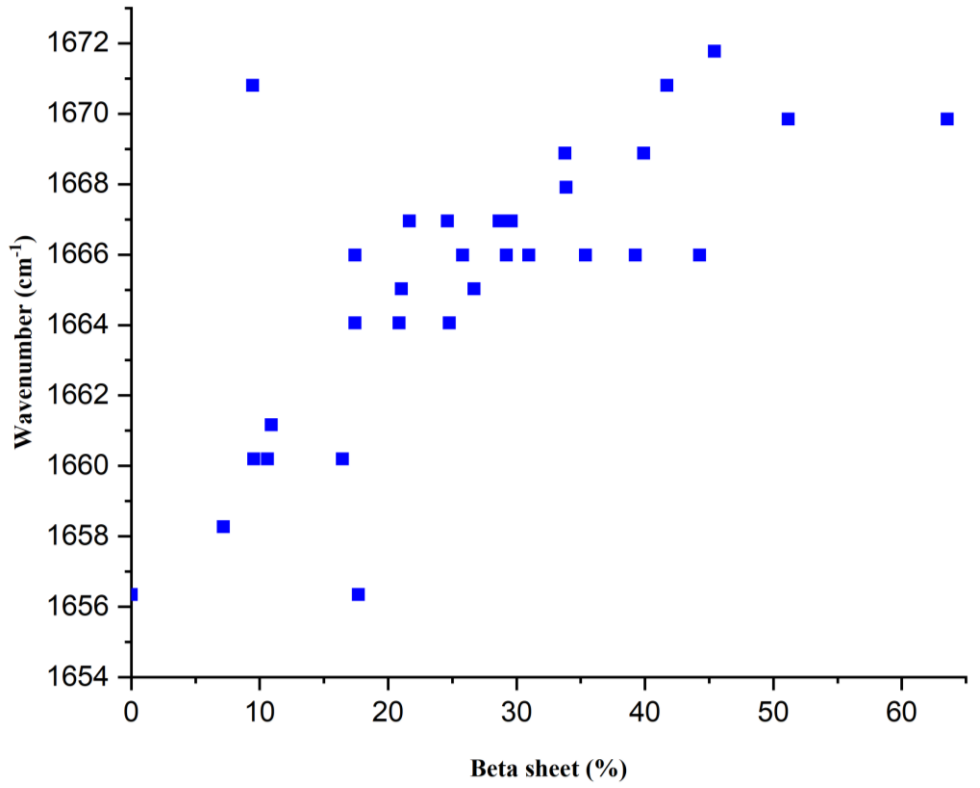

Figure 4.26. Wavenumber corresponding to the maximum of the peaks against their beta sheet content.

As it can be seen in Figures 4.25 and 4.26, the higher the helical content is the closer is the max to 1656 cm^{-1} and the higher in beta structure the closer to 1670 cm^{-1} which seems to agree with assignments given in the literature (Table 4-2).

The reference set consist of practically the same proteins used for solid IR with some few exemptions (heme and extremely fluorescing samples could not be measured by Raman and other proteins were measured instead). Figure 4.27 shows the SS coverage of the reference set.

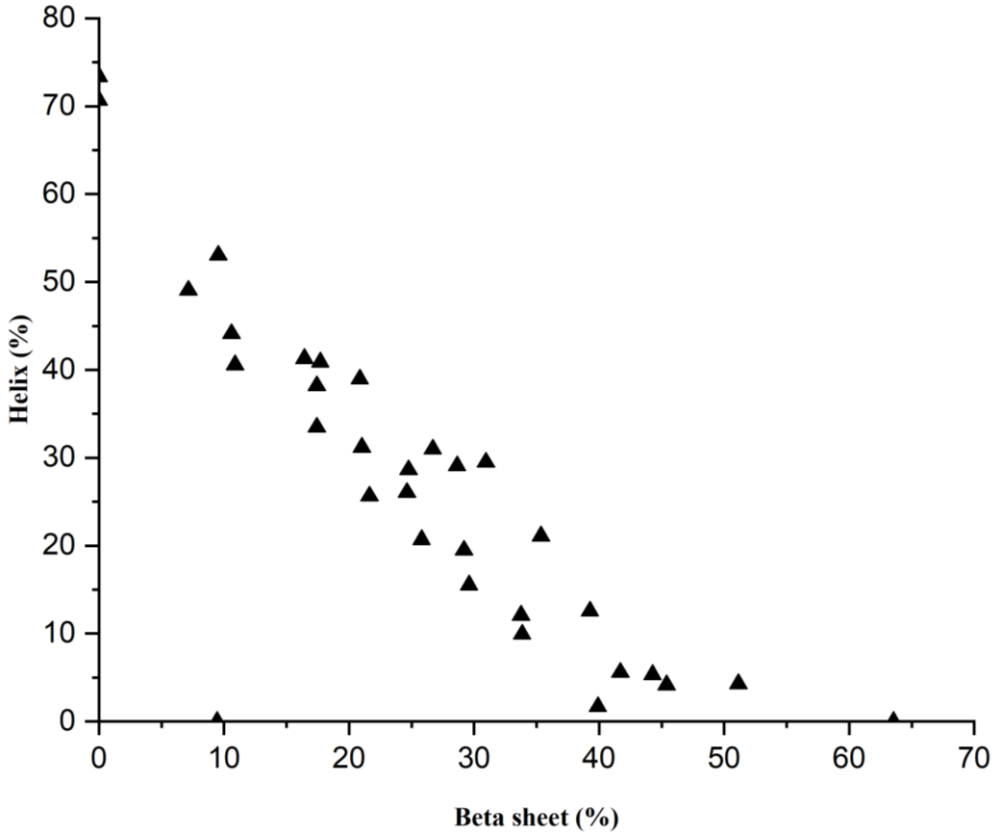

Figure 4.27. Helical vs sheet content of Raman reference set in solid state based on 2strc annotations.

It was observed that heme proteins show prominent peaks within the region of the amide I that make the analysis of SS unfeasible (Figure 4.28). These peaks are most likely due to resonance effects -Raman Resonance (RR)- from the pyrroline ring.

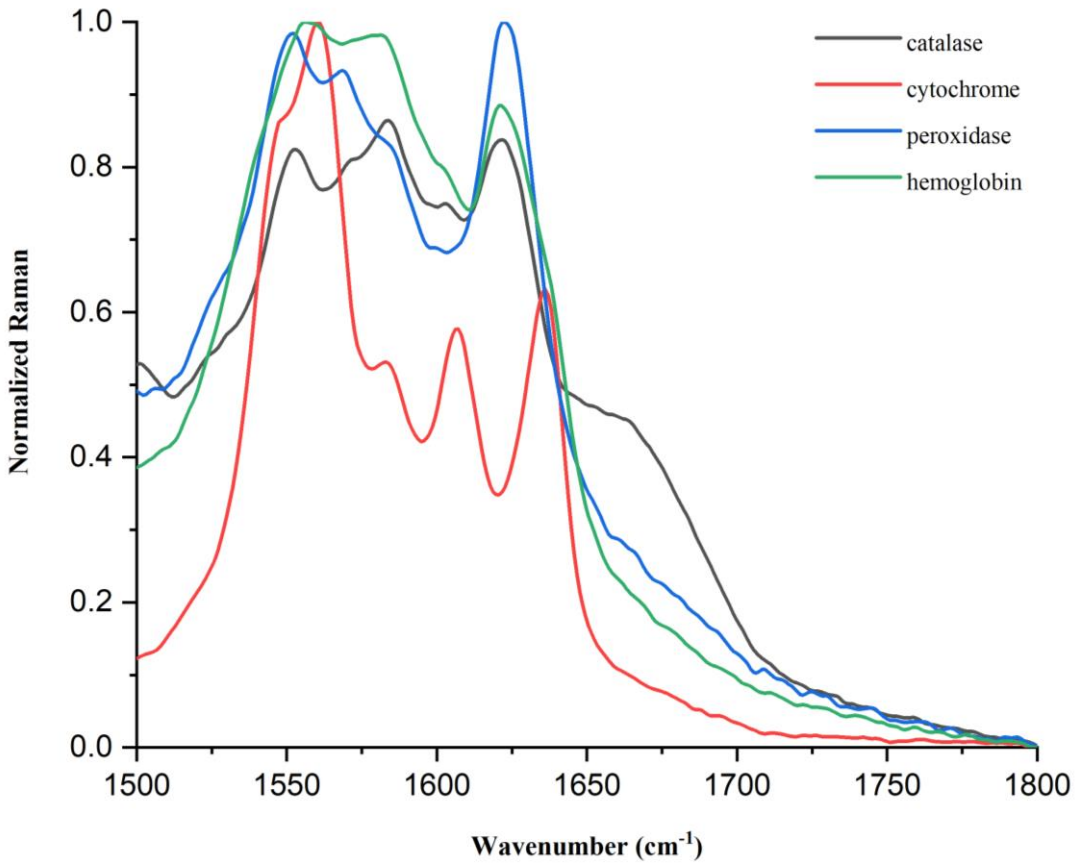

Figure 4.28. Raman spectra of heme proteins in solid state within the amide I and II regions. The spectra were normalized by the interval method but not baseline corrected.

4.4.3 Raman reference set in aqueous state

In the previous chapter it is shown that solid and aqueous IR data match for most of the proteins in the position of the maximum but do not compare in band broadening. Such differences were also found in Raman spectra (not shown) and thus a separate reference set for aqueous proteins was collected. As shown in Figure 4.16, water scatters weakly in the region of the amide I but it was still necessary to subtract it by measuring a blank separately. Because the signal recorded by the instrument is proportional to the incident power, concentration and total exposure time, the data had to be scaled by dividing it by the incident power and exposure time for the water subtraction (Equation 4.2-4.4). Finally, in order to make the different protein spectra comparable, it was necessary to divide them by the concentration (in $mg.ml^{-1}$) and multiply them by the MRW (in Da/Residue) too (Equation 4.5). As with IR, the spectra were also normalized by the interval method (Figure 4.30) and deconvolved + normalized (Figure 4.31).

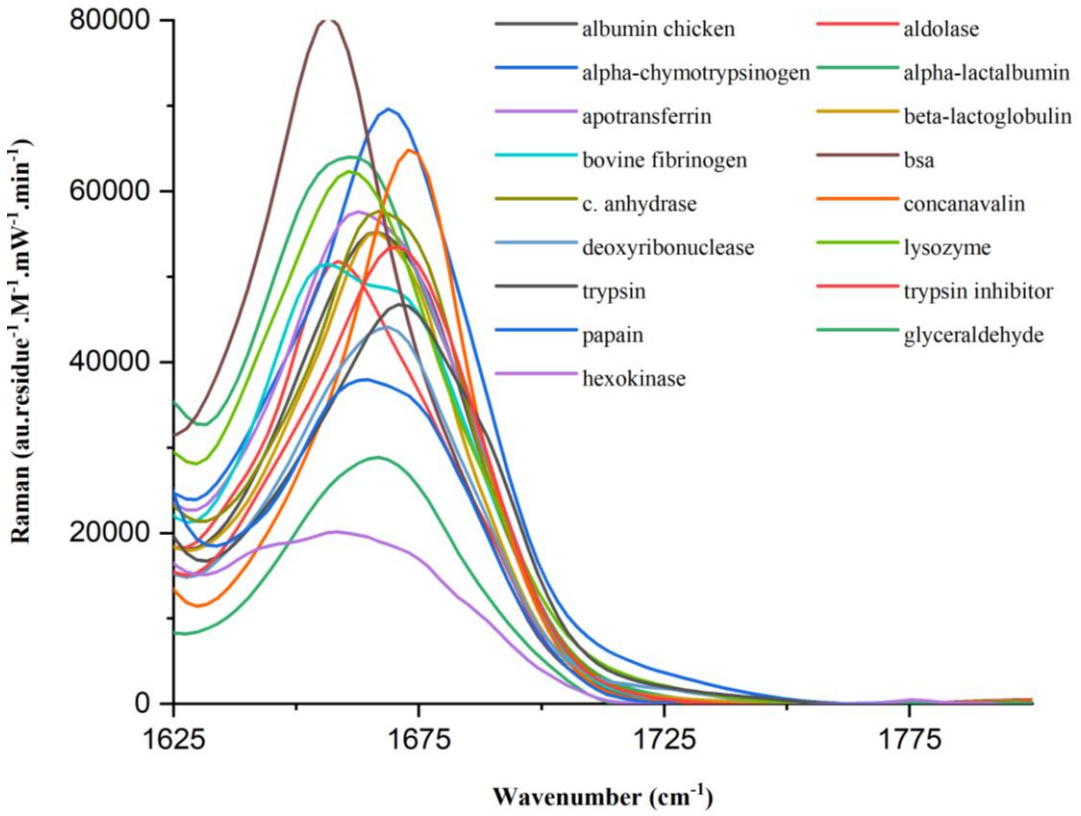

Figure 4.29. Raman spectra of the 17 proteins in solution. The spectra were water subtracted and scaled in accordance to Equations 4.2-4.5.

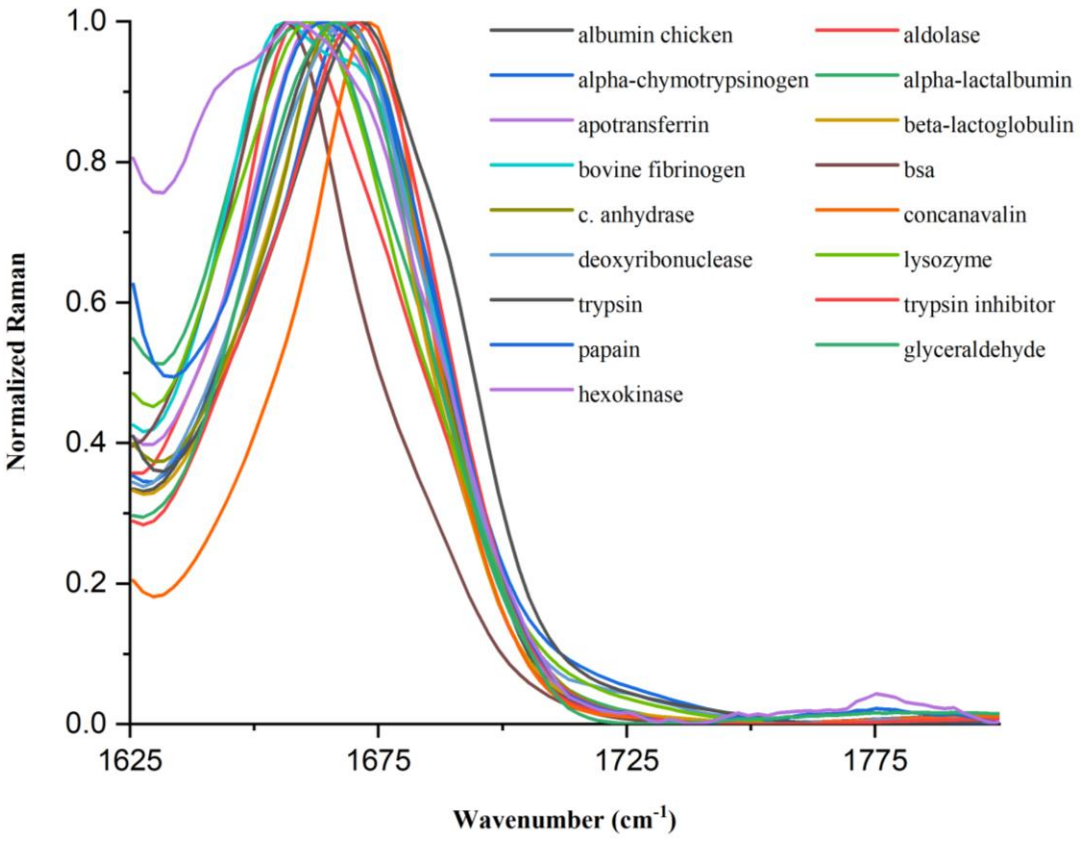

Figure 4.30. Normalized raman spectra of 17 proteins in solution.

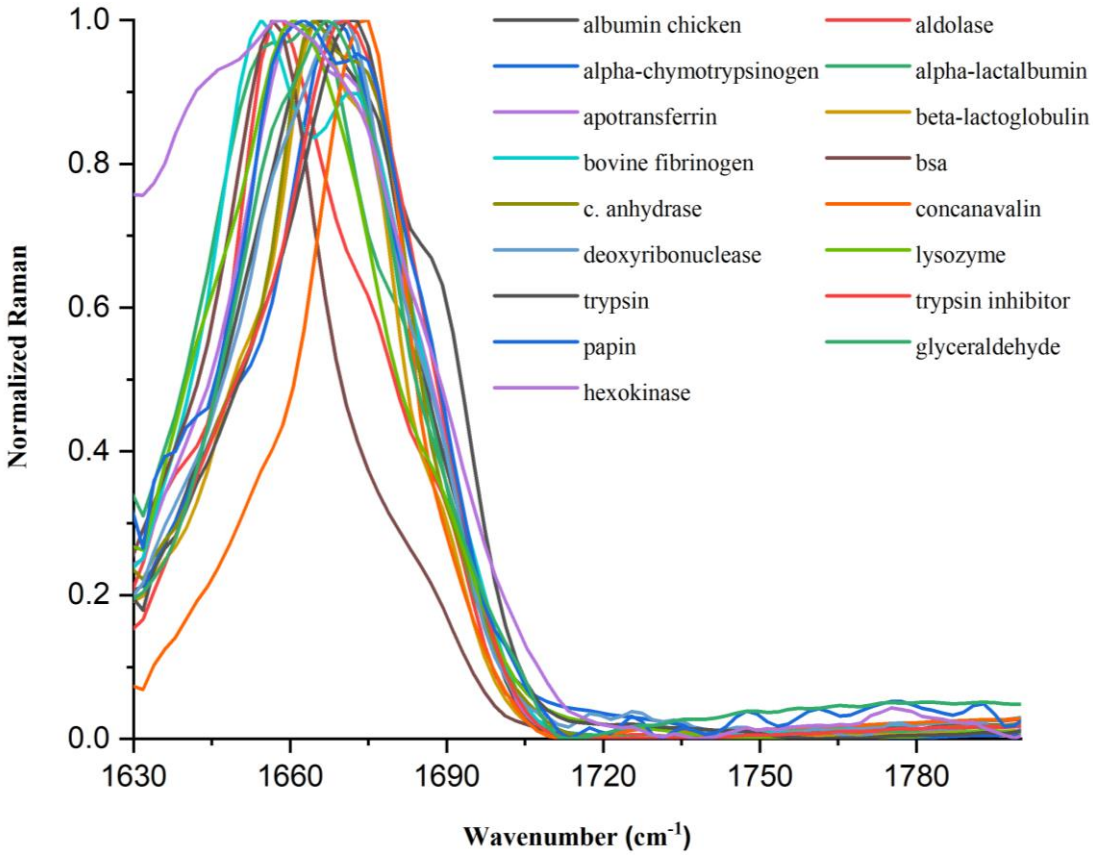

Figure 4.31. Raman spectra of 17 proteins in solution. The spectra were FSD in Origin and then normalized by the interval method. The deconvolution was done with a gamma factor of 10 and a smoothing factor of 0.25.

The reference set is made up of 17 proteins with good coverage in the range 0-50 and 10-45 sheet (Figure 4.32).

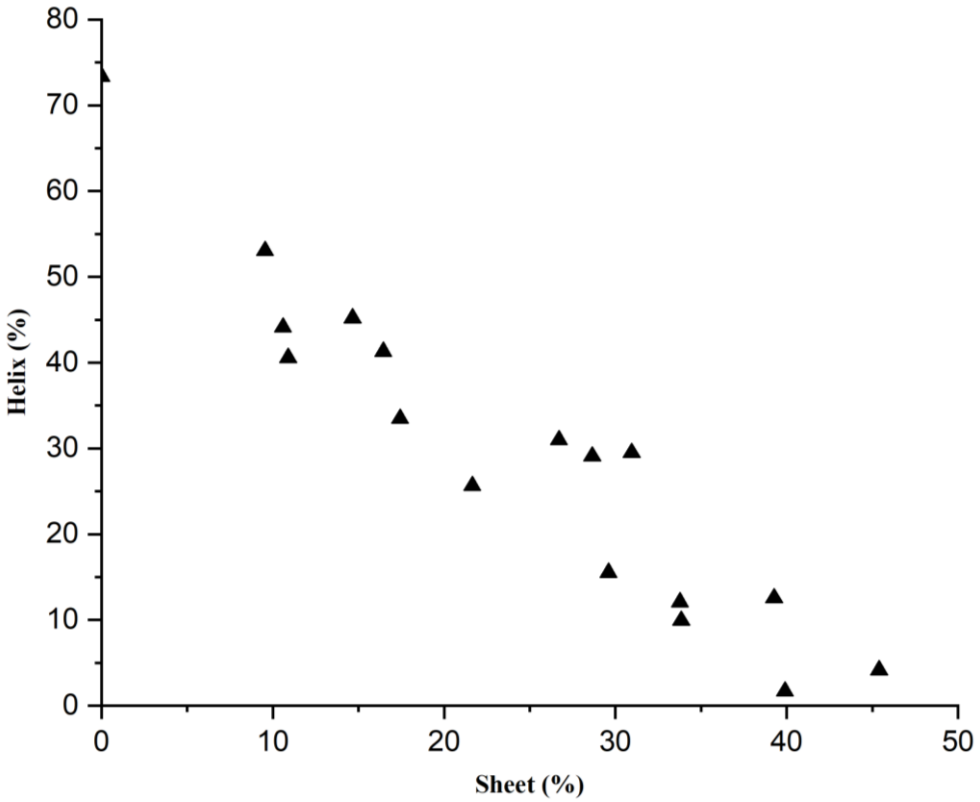

Figure 4.32. Helix vs sheet coverage of Raman reference set in aqueous state based on 2struc annotations.

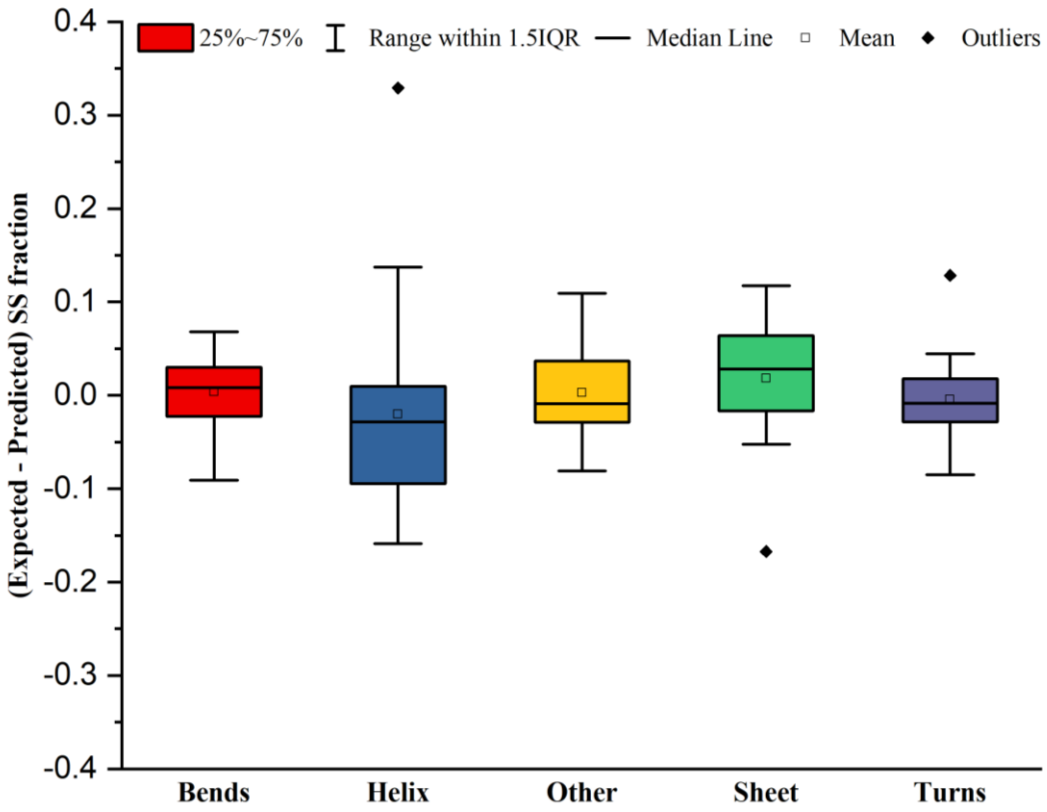

Figure 4.33. Results of the leave-one out validation of the 17-protein normalized Raman reference set in solution (Figure 4.30). The validation was run with a 40x40 map and 40000 iterations.

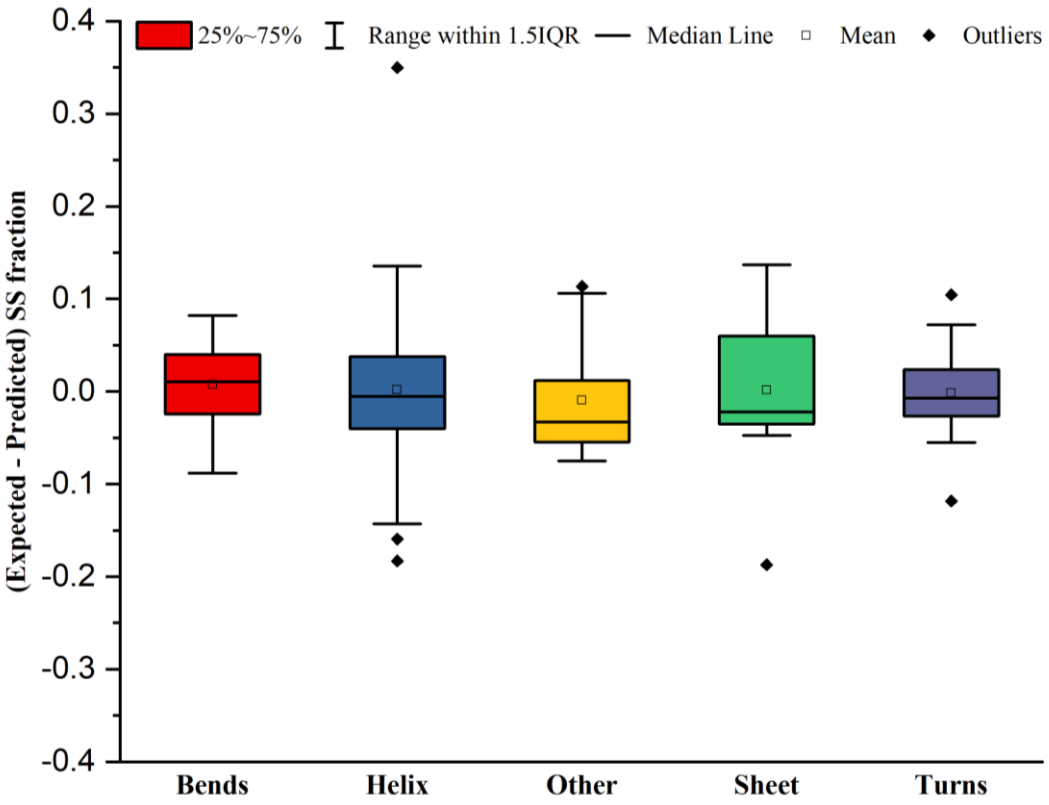

Figure 4.34. Results of the leave-one out validation of the 17-protein Raman reference set (Figure 4.31). The spectra were FSD and normalized by the interval method. The validation was run with a 40x40 map and 40000 iterations.

4.4.4 ROA reference set in aqueous state

Whereas aqueous Raman requires water subtraction and baseline correction due to fluorescence, ROA does not. However, the signal to noise ratio was significantly degraded because of fluorescence so photobleaching was still required to decrease the noise and a large number of accumulations needed in order to improve the signal to noise ratio. Unfortunately, due to time limitations, it was not possible to accumulate up to desired levels of S/N so a Savitsky-Golay (SG) filter with 3 points window and linear polynomial was applied to smooth the data. Furthermore, the baseline of some of the spectra drifted upwards so zeroing was applied within the range 1750-1850 cm^{-1} (Figure 4.35).

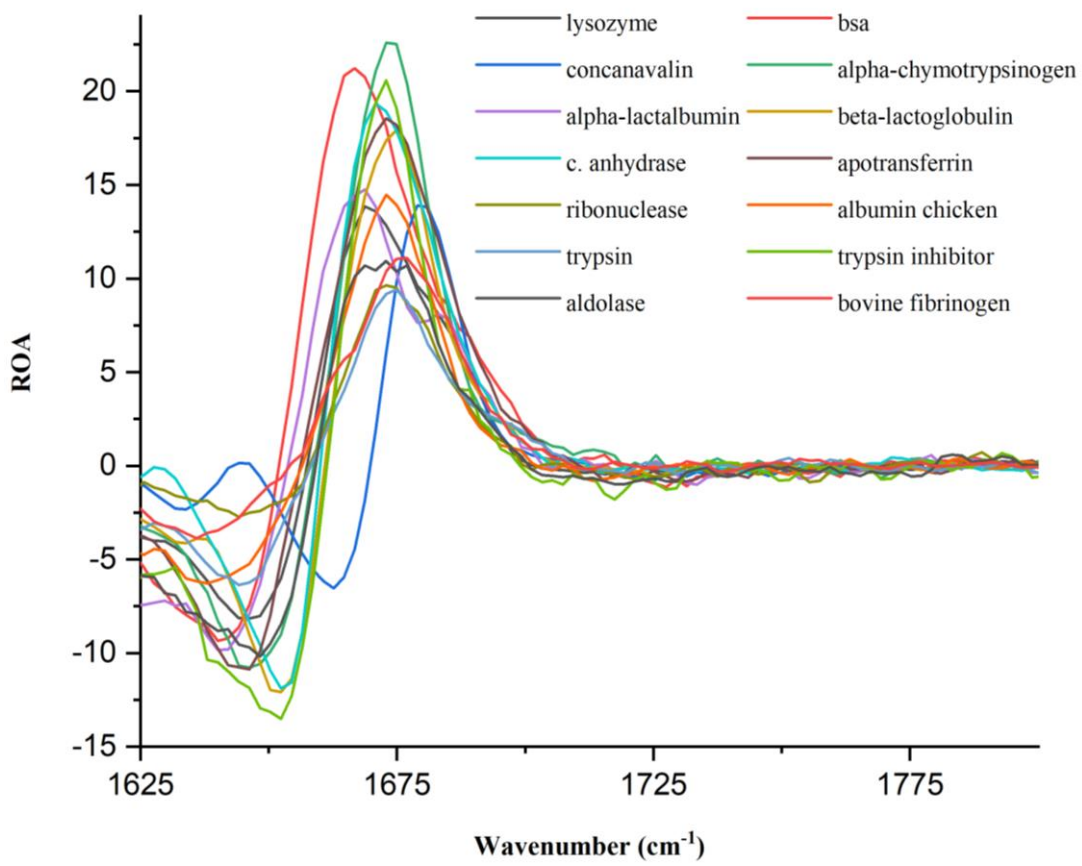

Figure 4.35. ROA spectra of 14 proteins in aqueous state. The data were scaled by incident power, exposure time, expressed in MRW concentration and zeroed as explained in the methods section.

The reference set consist of 14 proteins only. Because of the extreme background fluorescence displayed by some of the samples, the lack of solubility of others and the RR effect by the heme ones, it was impossible to cover all the helix vs sheet existing structures (note two big gaps in Figure 4.36 at ~5% and ~25% sheet content).

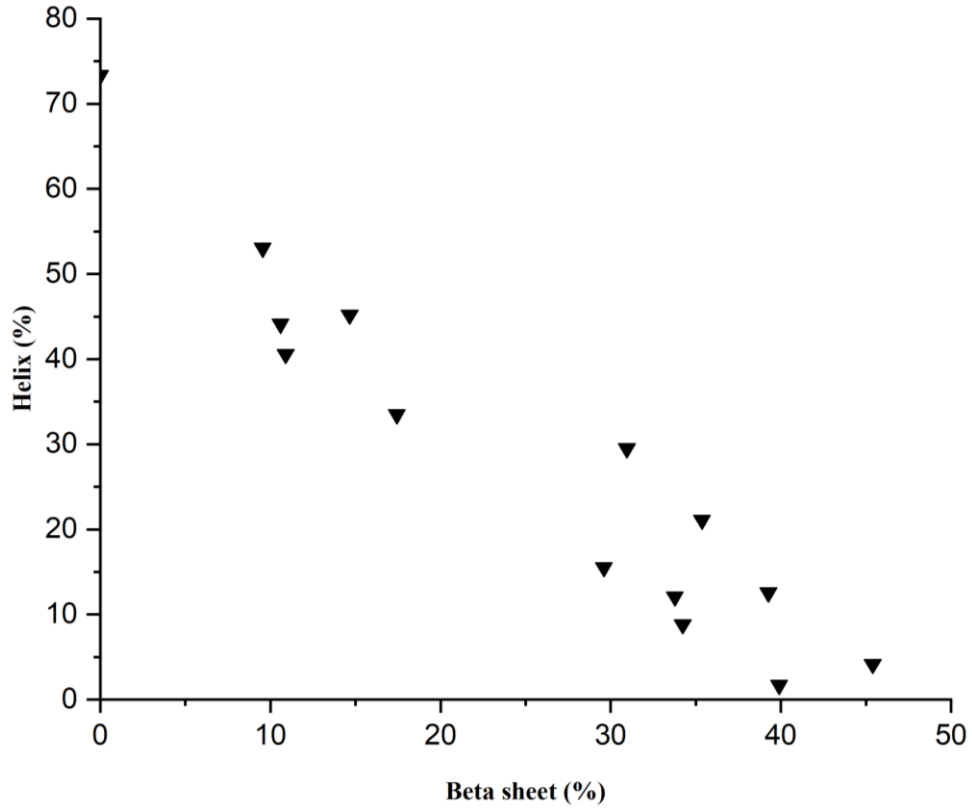

Figure 4.36. Helical vs sheet content of ROA reference set in aqueous state based on 2strc annotations.

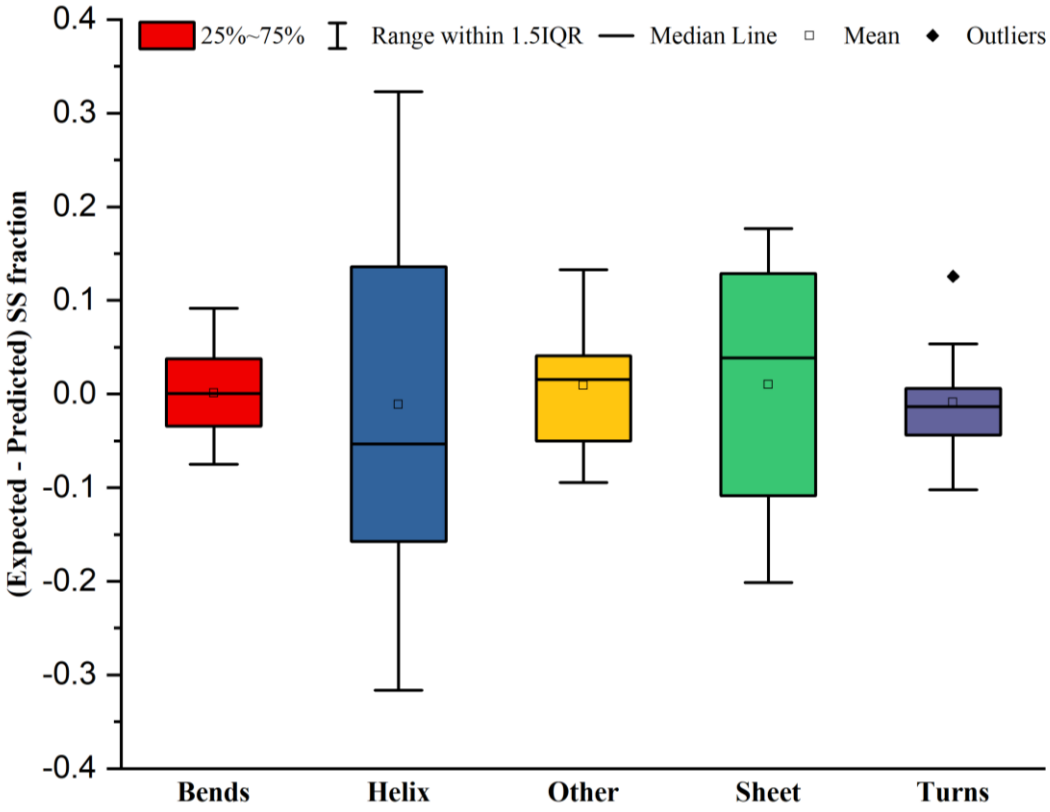

Figure 4.37. Results of the leave-one out validation of the 14- proteins ROA reference set in MRW concentration (Figure 4.35). The validation was run with a 40x40 map and 40000 iterations

In order to assess the quality of the spectra, a few proteins with the largest exposures to the laser (photobleaching + measurement) were measured by CD and fluorescence before and after Raman data collection**.**

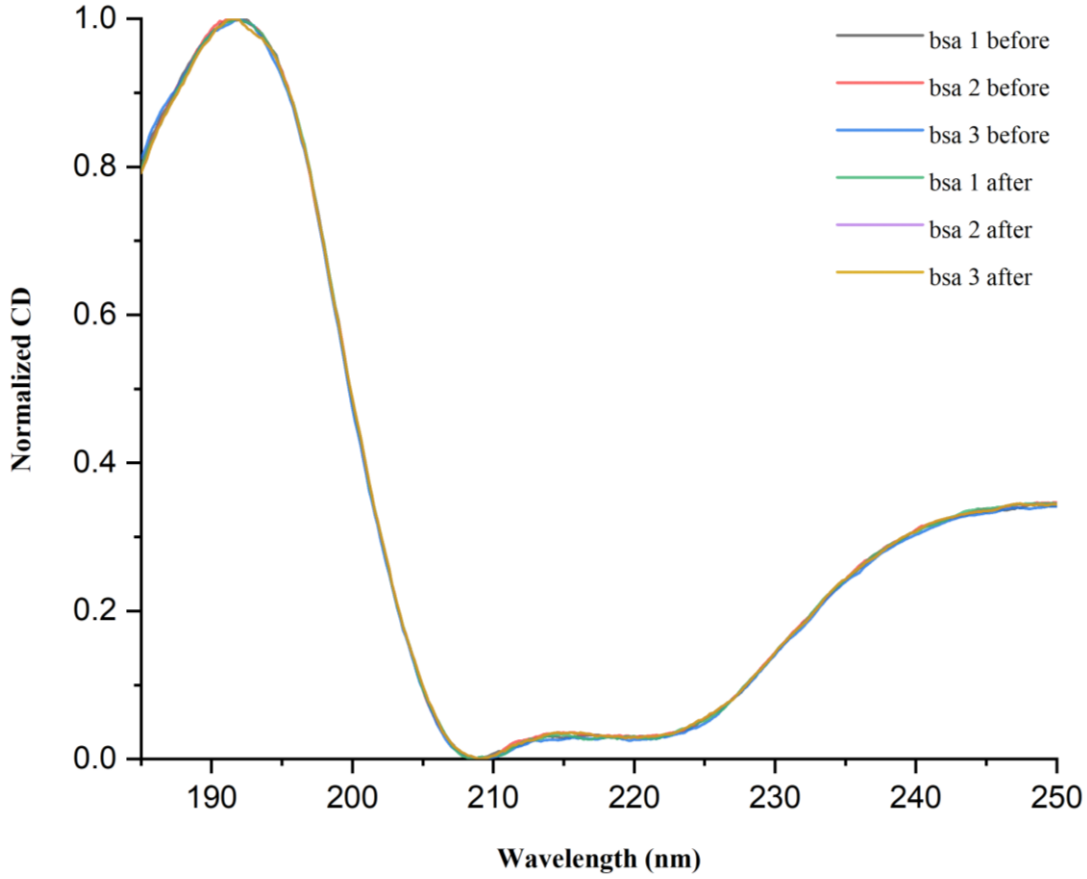

Figure 4.38. CD spectra of BSA replicates before and after long laser exposure. The spectra were water subtracted and normalized by the interval method.

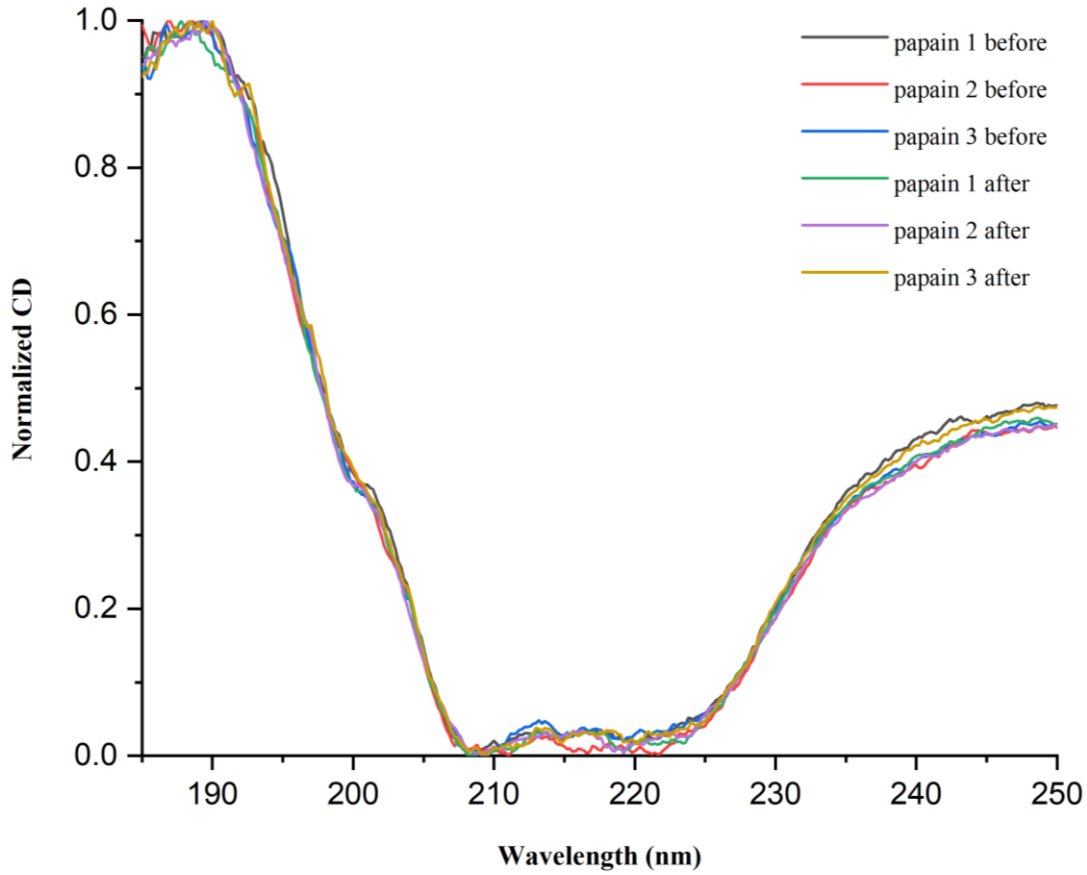

Figure 4.39. CD spectra of Papain replicates before and after long laser exposure. The spectra were water subtracted and normalized by the interval method.

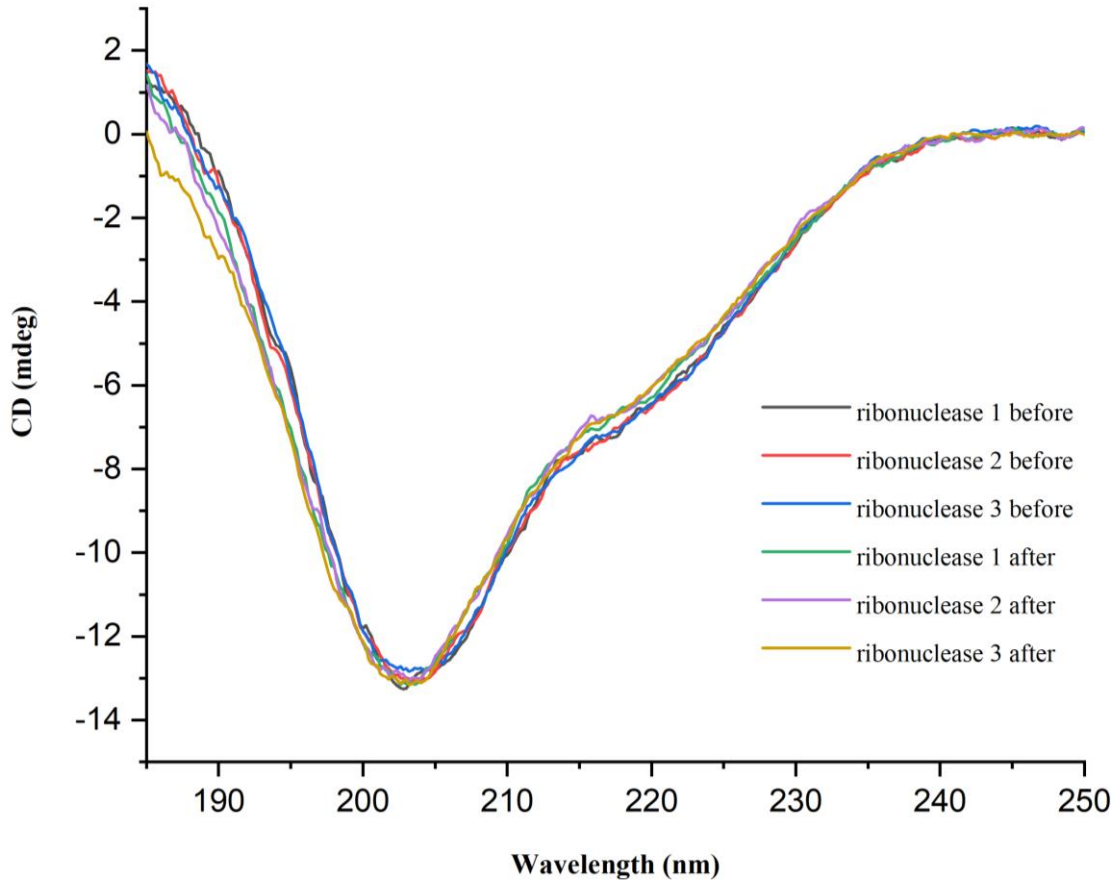

Figure 4.40. CD spectra of Ribonuclease replicates before and after long laser exposure. The spectra were water subtracted and plotted in mdeg.

4.5 CONCLUSIONS

In this chapter we report the collection of a Raman and ROA reference set spectra of proteins in aqueous state and train a neural network algorithm (SOM) with the Raman spectra processed with different strategies (normalized and deconvolved + normalized) and the ROA expressed in MRW extinction coefficient only, for prediction of SS. A leave-one-out validation was performed on each of the sets with 40x40 maps and 40000 iterations and represented with boxplots as done in chapter 3 with IR. SOM crashed every time we attempted to train it with the MRW ref set, something we blame on the magnitude disparity between the different spectra in the set. Just like with IR in chapter 3, "deconvolved + normalized" was found to provide the most accurate predictions followed very closely by "normalized" for the Raman. As for ROA, the leave-one-out validations show a very broad dispersion of the deviations from the expected values compared to those of Raman which we find difficult to blame on lack of representation (only 3 proteins less) and makes us

consider the suitability of the amide I (or the chosen range) for SS prediction of proteins instead.

On a separate note, we found in the literature and confirmed ourselves that, although higher powers speed up the photobleaching process, it produces a significant increase in temperature that could result in degradation not only of the target chromophores but also the protein and urges the need to implement mechanisms to stir and refrigerate the sample during the exposure to the laser.

5 GENERAL CONCLUSIONS AND FUTURE WORK

In this work, we investigated a set of vibrational spectroscopic techniques (IR, Raman and ROA) for SS determination of proteins based on their Amide I band by collecting large reference sets of proteins with known structure to train a neural network algorithm (SOM).

The theory of anomalous dispersion and absorption of evanescent waves were reviewed, and a procedure to convert experimental IR-ATR spectra to their transmission homologous established. Three different but related equations for the conversion -the original and two approximated versions of it- were derived and put to the test with a 50 $mg.ml^{-1}$ Lysozyme solution (absorbance ~ 0.15). No significant difference between the three conversions was appreciated meaning that for very small absorbances the most simplified version can be confidently used. The conversions were performed with two different strategies to approach the calculation of the refractive index: from experimental measurements of the transmission spectrum and iteratively from the experimental ATR spectrum, with the former being the most accurate of them.

Five reference sets were collected: IR in aqueous and solid state, Raman in aqueous and solid state and ROA in aqueous state and the aqueous ones used to trained SOM for SS prediction in the region of the Amide I band. A leave-one out validation was run for each of them and the differences between expected and predicted represented with boxplots. In IR, the best predictions were produced by the FSD + normalized ref set followed by the normalized and MRW ones. In Raman, it was impossible to validate the spectra in MRW but the normalized and FSD + normalized ones were validated and showed similar prediction power. Although FSD + normalized seemed to improve the predictions for the helical and sheet contents in both IR and Raman, it also enhances noise and thus it is only suitable for smooth spectra. Because of that, we chose only normalization as a pre-treatment. The comparison of the validations of a 21-protein IR ref set with a 47-protein one showed that the larger the number of proteins and coverage of SS is the more accurate the predictions are. This means

that, in order to improve the predictive power of SOM, the number of proteins spectra in the ref sets needs to keep growing.

Overall, Raman gave the best predictions followed by IR and ROA. Although ROA had the smallest number of proteins in the set (14), we struggle to believe that is the reason for the incredibly poor predictive capability considering the Raman ref set had only three more and yet its predictions were even better than the ones found with the 47-protein IR ref set.

In the future, more proteins will need to be added as mentioned above and the quality of the Raman spectra improved to performed FSD without enhancing noise. Also, amides II and III should be considered to train SOM for SS prediction with Raman, IR and ROA.

BIBLIOGRAPHY

1. Ang DL, Pinto-corujo M, Chmel N, Rodger A. o u r n a l a me. doi:10.1039/b000000x/In
2. Hall V, Nash A, Rodger A. SSNN, a method for neural network protein secondary structure fitting using circular dichroism data. *Anal Methods*. 2014;6(17):6721-6726. doi:10.1039/c3ay41831f
3. Ritter H, Kohonen T. Self-organizing semantic maps. *Biol Cybern*. 1989;61(4):241-254. doi:10.1007/BF00203171
4. Adams GP, Weiner LM. Monoclonal antibody therapy of cancer. *Nat Biotechnol*. 2005;23(9):1147-1157. doi:10.1038/nbt1137
5. Ausar S, Hasija, Li, Rahman. Forced degradation studies: an essential tool for the formulation development of vaccines. *Vaccine Dev Ther*. 2013:11. doi:10.2147/vdt.s41998
6. Kong J, Yu S. Fourier Transform Infrared Spectroscopic Analysis of Protein Secondary Structures Protein FTIR Data Analysis and Band Assign- ment. 2007;39(8):549-559.
7. Rygula A, Majzner K, Marzec KM, Kaczor A, Pilarczyk M. Raman spectroscopy of proteins : a review. 2013;(July):1061-1076. doi:10.1002/jrs.4335
8. Kelly SM, Jess TJ, Price NC. How to study proteins by circular dichroism. 2005;1751:119-139. doi:10.1016/j.bbapap.2005.06.005
9. Lesk AM. *Introduction to Protein Science*. Second. Oxford University press; 2010.
10. Leurs U, Mistarz UH, Rand KD. Getting to the core of protein pharmaceuticals - Comprehensive structure analysis by mass spectrometry. *Eur J Pharm Biopharm*. 2015;93:95-109. doi:10.1016/j.ejpb.2015.03.012
11. Gill DS. Protein pharmaceuticals: Discovery and preclinical development. *Adv Exp Med Biol*. 2009;655:28-36. doi:10.1007/978-1-4419-1132-2_3
12. Noh SM, Sathyamurthy M, Lee GM. Development of recombinant Chinese hamster ovary cell lines for therapeutic protein production. *Curr Opin Chem Eng*. 2013;2(4):391-397. doi:10.1016/j.coche.2013.08.002
13. Stryer L. *Biochemistry*. 4th ed. New York: Freeman; 1999.
14. Vaidyanathan K. Textbook of Biochemistry for Dental Students. *Textb Biochem Dent Students*. 2017;(January). doi:10.5005/jp/books/13106
15. Buxbaum E. *Fundamentals of {Protein} {Structure} and {Function}.*; 2007. doi:10.1007/978-0-387-68480-2
16. W K, C. S. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 1983;22:2577-2637.
17. Malkov S, Zivkovic M V., Beljanski M V., Zaric SD. Correlations of Amino Acids with Secondary Structure Types: Connection with Amino Acid Structure. 2005;(34). http://arxiv.org/abs/q-bio/0505046.
18. Krieger F, Möglich A, Kiefhaber T. Effect of proline and glycine residues on dynamics and

barriers of loop formation in polypeptide chains. *J Am Chem Soc*. 2005;127(10):3346-3352. doi:10.1021/ja042798i

19. Susi H, Serge N, Stevens L. Infrared spectra and protein conformations in aqueous solutions. *J Biol Chem*. 1967;242(23):5460-5466.

20. Fabian H, Werner M. Infrared Spectroscopy of Proteins.

21. Dong A, Caughey WS, Caughey B, Bhat KS, Coe JE. Secondary Structure of the Pentraxin Female Protein in Water Determined by Infrared Spectroscopy: Effects of Calcium and Phosphorylcholine. *Biochemistry*. 1992;31(39):9364-9370. doi:10.1021/bi00154a006

22. Barth A. Infrared spectroscopy of proteins. 2007;1767:1073-1101. doi:10.1016/j.bbabio.2007.06.004

23. Witze ES, Old WM, Resing KA, Ahn NG. <Nature Methods 2007 Witze.pdf>. 2007;4(10). doi:10.1038/1100

24. Domon B, Aebersold R. Mass spectrometry and protein analysis. *Science (80-)*. 2006;312(5771):212-217. doi:10.1126/science.1124619

25. Hunt DF, Yates Iii JR, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry (collision-activated dissociation/liquid secondary-ion mass spectrometry/apolipoprotein B). *Proc Nati Acad Sci USA*. 1986;83(September):62334237. https://www.pnas.org/content/pnas/83/17/6233.full.pdf.

26. Gawas UB, Mandrekar VK, Majik MS. *Structural Analysis of Proteins Using X-Ray Diffraction Technique*. Elsevier Inc.; 2019. doi:10.1016/b978-0-12-817497-5.00005-7

27. Poulsen FM. A brief introduction to NMR spectroscopy of proteins By Flemming M . Poulsen. 2002:1-33.

28. Yu H. Extending the size limit of protein nuclear magnetic resonance. *Proc Natl Acad Sci U S A*. 1999;96(2):332-334. doi:10.1073/pnas.96.2.332

29. Frueh DP, Goodrich AC, Mishra SH, Nichols SR. NMR methods for structural studies of large monomeric and multimeric proteins. *Curr Opin Struct Biol*. 2013;23(5):734-739. doi:10.1016/j.sbi.2013.06.016

30. Banwell CN, McCash EM. *Fundamentals of Molecular Spectroscopy*. 4th ed. London: Mc Graw Hill; 1994.

31. Hui Y, Xue X, Xuesong Z, Yan WU. Intrinsic Fluorescence Spectra of Tryptophan , Tyrosine and Phenyloalanine. 2015;(Icadme):224-233.

32. Noble JE, Bailey MJA. Chapter 8 Quantitation of Protein. Methods in Enzymology. doi:10.1016/S0076-6879(09)63008-1

33. Uniprot. https://www.uniprot.org/.

34. Bowen WJ. the Absorption Spectra and Extinction Coefficients of Myoglobin. *J Biol Chem*. 1949;179(1):235-245. http://www.jbc.org/content/179/1/235.short.

35. Colin N. Banwell and Elaine M. McCash. *Fundamentals of Molecular Spectroscopy*. Fourth. Mc Graw Hill; 1994.

36. Bengt N, Rodger A, Dafforn T. *Linear Dichroism and Circular Dichroism*. RSC Publishing; 2010. doi:10.1002/9780470027318.a5402

37. Antiquity G, Testament N, Stories M, Central PE. Part Iv Part Iv. 2019;(April):29-30.

38. Sreerama N, Woody RW. Estimation of protein secondary structure from circular dichroism spectra: Comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal Biochem*. 2000;287(2):252-260. doi:10.1006/abio.2000.4880

39. Ghisaidoobe ABT, Chung SJ. Intrinsic Tryptophan Fluorescence in the Detection and Analysis of Proteins : A Focus on Förster Resonance Energy Transfer Techniques. 2014:22518-22538. doi:10.3390/ijms151222518

40. Transitions N, States E, Perrin T. Characteristics of Fluorescence Emission. *Mol Fluoresc*. 2012;0:53-74. doi:10.1002/9783527650002.ch3

41. Townshen A. *Principles of Instrumental Analysis*. Vol 152.; 1983. doi:10.1016/S0003-2670(00)84936-3

42. Chen Y, Barkley MD. Toward Understanding Tryptophan Fluorescence in Proteins †. 1998;2960(4):9976-9982. doi:10.1021/bi980274n

43. none. Fluorescence Spectroscopy : A tool for Protein folding / unfolding Study. *Energy*. 2007:1-8.

44. Larkin VM. Introduction. *Assist Technol*. 2004;16(2):73-84. doi:10.1080/10400435.2004.10132076

45. Smith BC. *Fundamentals of Fourier Transform Infrared Spectroscopy*. 2nd ed. Boca Raton, London, N. York: CRC Press Taylor and Francis Group; 2016.

46. Yang H, Yang S, Kong J, Dong A, Yu S. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. 2015;(March). doi:10.1038/nprot.2015.024

47. Brian. SC. *Fundamental of Fourier Transform Infrared Spectroscopy Second Edition*.; 2011. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

48. Grdadolnik J. ATR-FTIR spectroscopy: Its advantages and limitations. *Acta Chim Slov*. 2002;49(3):631-642.

49. John R. Ferraro, Kazuo Nakamoto CWB. *Introductory Raman Spectroscopy*. Vol 36.; 2003. doi:10.1002/jrs.1407

50. Krimm BS, Bandekart J. Vibrational Spectroscopy and Conformation. *Adv Protein Chem*. 1986:48109. https://www.google.com/books?hl=zh-CN&lr=&id=0rke7tuwUysC&oi=fnd&pg=PA181&dq=Vibrational+spectroscopy+and+conformation+of+peptides,+polypeptides,+and+proteins.&ots=uMUdr9akIH&sig=PwhrsMxwz4TUyc9fKZjMdVval4M.

51. Budhavaram NK, Barone JR. Quantifying amino acid and protein substitution using Raman spectroscopy. *J Raman Spectrosc*. 2011;42(3):355-362. doi:10.1002/jrs.2738

52. Kitagawa T, Hirota S. Raman Spectroscopy of Proteins ORDINARY RAMAN SPECTRA. 2006;1(c). doi:10.1002/9780470027325.s8202

53. Kagan MR, Mccreery RL. Reduction of Fluorescence Interference in Raman Spectroscopy via Analyte Adsorption on Graphitic Carbon. 1994;66(23):4159-4165. doi:10.1021/ac00095a008

54. Widengren J, Rigler R. Mechanisms of photobleaching investigated by fluorescence correlation spectroscopy. 1996;4:149-157.

55. Wei D, Chen S, Liu Q. Review of Fluorescence Suppression Techniques in Raman Spectroscopy Review of Fluorescence Suppression Techniques in Raman Spectroscopy. 2015;4928. doi:10.1080/05704928.2014.999936

56. Vandenabeele P. *Practical Raman Spectroscopy*. (Wiley J, ed.). Chichester, West Sussex; 2013.

57. Requena A, zuñiga J. *Espectroscopia*. Madrid: Pearson Prentice Hall; 2004.

58. Barron LD, Hecht L, McColl IH, Blanch EW. Raman optical activity comes of age. *Mol Phys*. 2004;102(8):731-744. doi:10.1080/00268970410001704399

59. Mensch C, Barron LD, Johannessen C. Ramachandran mapping of peptide conformation using a large database of computed Raman and Raman optical activity spectra. *Phys Chem Chem Phys*. 2016;18(46):31757-31768. doi:10.1039/c6cp05862k

60. McColl IH, Blanch EW, Hecht L, Barron LD. A study of ??-helix hydration in polypeptides, proteins, and viruses using vibrational Raman optical activity. *J Am Chem Soc*. 2004;126(26):8181-8188. doi:10.1021/ja048991u

61. Buckley K, Ryder AG. Applications of Raman Spectroscopy in Biopharmaceutical Manufacturing: A Short Review. *Appl Spectrosc*. 2017;71(6):1085-1116. doi:10.1177/0003702817703270

62. Pham P V. Learn more about Recombinant DNA Technology Medical Biotechnology. 2018.

63. Chusainow J, Yang YS, Yeo JHM, et al. A study of monoclonal antibody-producing CHO cell lines: What makes a stable high producer? *Biotechnol Bioeng*. 2009;102(4):1182-1196. doi:10.1002/bit.22158

64. Fahrner RL, Knudsen HL, Basey CD, et al. Industrial purification of pharmaceutical antibodies: Development, operation, and validation of chromatography processes. *Biotechnol Genet Eng Rev*. 2001;18:301-327. doi:10.1080/02648725.2001.10648017

65. Surabattula R, Rao KRSS, Polavarapu R. An Optimized Process for Expression , Scale-Up and Purification of Recombinant Erythropoietin Produced in Chinese Hamster Ovary Cell Culture. *Res Biotechnol*. 2011;2(3):58-74.

66. Tan SC, Yiap BC. DNA, RNA, and protein extraction: The past and the present. *J Biomed Biotechnol*. 2009;2009. doi:10.1155/2009/574398

67. Et B, Acta B, Svensson H. INVESTIGATION OF SPECIFICITY IN MEMBRANE B R E A K A G E. 274(1972):447-461.

68. Woods DA, Bain CD. Total Internal Reflection Spectroscopy for Studying Soft Matter. 2013.

69. Miljković M, Bird B, Diem M. Line shape distortion effects in infrared spectroscopy. *Analyst*. 2012;137(17):3954-3964. doi:10.1039/c2an35582e

70. Harrick NJ. *Internal Reflection Spectroscopy*. John Wiley & Sonc Inc; 1967.

71. Khoshhesab ZM. *Infrared Spectroscopy*. (Theophanides Theophile, ed.). Shanghai: InTech; 2012. doi:10.5772/32009

72. Jenkins FA, White HE. *FUNDAMENTALS OF OPTICS*.; 2001. https://seanghor.files.wordpress.com/2011/11/fundamentals-of-optics_0072561912.pdf.

73. Jang WH, Miller JD. Verification of the Internal Reflection Spectroscopy Adsorption Density Equation by Fourier Transform Infrared Spectroscopy Analysis of Transferred Langmuir-Blodgett Films. *Langmuir*. 1993;9(11):3159-3165. doi:10.1021/la00035a068

74. Milosevic M. Internal Reflection and ATR Spectroscopy. 2004;39(3):365-384. doi:10.1081/ASR-200030195

75. Grant R. Fowles. *Introduction to Modern Optics*. 2nd ed. New York: Dover; 1975.

76. Romeo M, Diem M. Correction of dispersive line shape artifact observed in diffuse reflection infrared spectroscopy and absorption/reflection (transflection) infrared micro-spectroscopy. *Vib Spectrosc*. 2005;38(1-2):129-132. doi:10.1016/j.vibspec.2005.04.003

77. Ogilvie JF, Fee GJ. Equivalence of Kramers-Kronig and Fourier transforms to convert between optical dispersion and optical spectra. *Match*. 2013;69(2):249-262. http://www.scopus.com/inward/record.url?eid=2-s2.0-84898767354&partnerID=40&md5=3e2421c9792dea0c80ac0f6f23adc995.

78. Kanwal RAMP. Linear integral equations (Theory and technique), second edition. *J Comput Appl Math*. 2002;81(2):N7-N8. doi:10.1016/s0377-0427(97)89471-x

79. The MathWorks. Matlab. 2018.

80. OriginLab Corporation. Origin. 2018.

81. David J. Griffiths. *Introduction to Electrodynamics*. 4th ed. Uttar Pardesh: Pearson; 2015.

82. Milosevic M. Anatomy of ATR Absorption. *Intern Reflect ATR Spectrosc*. 2012:67-78. doi:10.1002/9781118309742.ch7

83. Fields E, The AT. F resnel Equations. 2012.

84. Averett LA, Griffiths PR, Nishikida K. Effective path length in attenuated total reflection spectroscopy. *Anal Chem*. 2008;80(8):3045-3049. doi:10.1021/ac7025892

85. Riley KF, Hobson MP, Bence SJ. *Mathematical Methods for Physics and Engineering*. 3rd ed. Delhi: Cambridge university press; 2006.

86. Arbesú A, Barth A, Romero Y, Kumar S. Biologicals Characterization of recombinant antibodies for cancer therapy by infrared spectroscopy. *Biologicals*. 2013;41(2):104-110. doi:10.1016/j.biologicals.2012.11.004

87. Wang L, Wu Y, Meersman F. Clarification of the thermally-induced pretransition of ribonuclease A in solution by principal component analysis and two-dimensional correlation infrared spectroscopy. 2006;42:201-205. doi:10.1016/j.vibspec.2006.05.010

88. Goormaghtigh E, Ruysschaert J, Raussens V. Evaluation of the Information Content in Infrared Spectra for Protein Secondary Structure Determination. 2006;90(April):2946-2957. doi:10.1529/biophysj.105.072017

89. Van De Weert M, Haris PI, Hennink WE, Crommelin DJA. Fourier transform infrared spectrometric analysis of protein conformation: Effect of sampling method and stress factors. *Anal Biochem*. 2001;297(2):160-169. doi:10.1006/abio.2001.5337

90. Byler DM, Susi H, Regional E. Examination of the Secondary Structure of Proteins by Deconvolved FTIR Spectra. 1986;25:469-487.

91. Wang Y, Boysen RI, Wood BR, Kansiz M, Mcnaughton D, Hearn MTW. Biopolymers Volume 89 / Number 11 895. 2008;89(11). doi:10.1002/bip.21022

92. Hering JA, Innocent PR, Haris PI. Towards developing a protein infrared spectra databank (PISD) for proteomics research. 2004:2310-2319. doi:10.1002/pmic.200300808

93. Zhu G, Zhu X. The growing self-organizing map for clustering algorithms in programming codes. *Proc - Int Conf Artif Intell Comput Intell AICI 2010*. 2010;3:178-182. doi:10.1109/AICI.2010.276

94. Kohonen T. The%20Self-Organizing%20Map%20(Kohonen).pdf.

95. Lasri R. Clustering and classification using a self-organizing MAP: The main flaw and the improvement perspectives. *Proc 2016 SAI Comput Conf SAI 2016*. 2016:1315-1318. doi:10.1109/SAI.2016.7556150

96. Miyazawa T, Shimanouchi T, Muzushima SI. Normal vibrations of N-methylacetamide. *J Chem Phys*. 1958;29(3):611-616. doi:10.1063/1.1744547

97. Elliot A, Ambrose EJ. Structure of Sythetic Polypetides. *Nature*. 1950;(4206):921-922.

98. Miyazawa T. Perturbation treatment of the characteristic vibrations of polypeptide chains in various configurations. *J Chem Phys*. 1960;32(6):1647-1652. doi:10.1063/1.1730999

99. Dwivedi AM, Krimm S. Vibrational Analysis of Peptides, Polypeptides, and Proteins. XI. β-Poly(l-alanine) and Its N-Deuterated Derivative. *Macromolecules*. 1982;15(1):186-193. doi:10.1021/ma00229a036

100. Moore WH, Krimm S. Transition dipole coupling in Amide I modes of , 3 polypeptides. 1975;72(12):4933-4935.

101. tiffany ML, Krimm S. Circular dichroism of poly-L-proline in an unordered conformation. *Biopolymers*. 1968;6(12):1767-1770. doi:10.1002/bip.1968.360061212

102. Rüegg M, Metzger V, Susi H. Computer analyses of characteristic infrared bands of globular proteins. *Biopolymers*. 1975;14(7):1465-1471. doi:10.1002/bip.1975.360140712

103. Susi BH, Byler DM. different FeS clusters , one each of 2Fe , 3Fe , and 4Fe , provides the first [13] R e s o l u t i o n - E n h a n c e d F o u r i e r T r a n s f o r m I n f r a r e d Spectroscopy of Enzymes ing the secondary structure of polypeptides and proteins . A. *Methods*. 1985;130(1950):290-311. doi:10.1083/jcb.1862iti1

104. Dong A, Huang P, Caughey WS. Redox-Dependent Changes in. ^{2}-extended Chain and Turn Structures of Cytochrome C in Water Solution Determined by Second Derivative Amide I Infrared Spectra. *Biochemistry*. 1992;31(1):182-189. doi:10.1021/bi00116a027

105. Wen ZQ. Raman Spectroscopy of Protein Pharmaceuticals. *Pharm Sci*. 2007;96(11):2861-2878. doi:10.1002/jps

106. Peters J, Luczak A, Ganesh V, Park E, Kalyanaraman R. Protein Secondary Structure Determination Using Drop Coat Deposition Confocal Raman Spectroscopy. *Spectroscopy*. 2016;31(10):31-39.

107. Yu T -J, Lippert JL, Peticolas WL. Laser Raman studies of conformational variations of poly-L-lysine. *Biopolymers*. 1973;12(9):2161-2176. doi:10.1002/bip.1973.360120919

108. Lippert JL, Tyminski D, Desmeules PJ. Determination of the Secondary Structure of Proteins by Laser Raman Spectroscopy. 1975:7075-7080. doi:10.1021/ja00438a057

109. Takeuchi H. Raman structural markers of tryptophan and histidine side chains in proteins. *Biopolym - Biospectroscopy Sect*. 2003;72(5):305-317. doi:10.1002/bip.10440

110. Williamst RW. Estimation of Protein Secondary Structure from the Laser Raman Amide ! Spectrum. 1983:581-603.

111. Kane O, Biochemistry J, Bussian BM, Sander C. How To Determine Protein Secondary Structure in Solution by Raman Spectroscopy : Practical Guide and Test Case DNase I. 1989;(1981):4271-4277. doi:10.1021/bi00436a023

112. Bandekar J. Amide modes and protein conformation. *Biochim Biophys Acta (BBA)/Protein Struct Mol*. 1992;1120(2):123-143. doi:10.1016/0167-4838(92)90261-B

113. Maiti NC, Apetri MM, Zagorski MG, Carey PR, Anderson VE. Raman Spectroscopic Characterization of Secondary Structure in Natively Unfolded Proteins : r -Synuclein. 2004:2399-2408. doi:10.1021/ja0356176

114. Macdonald AM, Wyeth P. On the use of photobleaching to reduce fluorescence background in Raman spectroscopy to improve the reliability of pigment identification on painted textiles. *J Raman Spectrosc*. 2006. doi:10.1002/jrs.1510

115. Diaspro A, Chirico G, Usai C, Ramoino P. Photobleaching. 2010;(September 2016). doi:10.1007/978-0-387-45524-2

116. Zhu F, Isaacs NW, Hecht L, Barron LD. Raman optical activity: A tool for protein structure analysis. *Structure*. 2005;13(10):1409-1419. doi:10.1016/j.str.2005.07.009

117. Mensch C, Barron LD, Johannessen C. Ramachandran mapping of peptide conformation using a large database of computed Raman and Raman optical activity spectra. *Phys Chem Chem Phys*. 2016;18(46):31757-31768. doi:10.1039/C6CP05862K

118. Wilson G, Ford SJ, Cooper A, Hecht L, Wen ZQ, Barron LD. Vibrational raman optical activity of α-lactalbumin: Comparison with lysozyme, and evidence for native tertiary folds in molten globule states. *J Mol Biol*. 1995;254(4):747-760. doi:10.1006/jmbi.1995.0652

119. Tuma R. Raman spectroscopy of proteins: From peptides to large assemblies. *J Raman Spectrosc*. 2005;36(4):307-319. doi:10.1002/jrs.1323

120. Stynes HC, Ibers JA. et al., ’12. 1972;293(14):5127-5128.

121. Sane SU, Cramer SM, Przybycien TM. A Holistic Approach to Protein Secondary Structure Characterization Using Amide I. 1999;272:255-272.

122. Miura T, Thomas GJ. *Proteins: Structure, Function, and Engineering*. Vol 24. New York: Springer Sciece+Business Media; 1995. doi:10.1039/c3cs60094g

123. Blanch EW, Morozova-Roche LA, Hecht L, Noppe W, Barron LD. Raman optical activity characterization of native and molten globule states of equine lysozyme: Comparison with hen lysozyme and bovine α- lactalbumin. *Biopolym - Biospectroscopy Sect*. 2000;57(4):235-248. doi:10.1002/1097-0282(2000)57:4<235::AID-BIP5>3.0.CO;2-H

124. McColl IH, Blanch EW, Hecht L, Kallenbach NR, Barron LD. Vibrational Raman Optical Activity Characterization of Poly(L-proline) II Helix in Alanine Oligopeptides. *J Am Chem Soc*. 2004;126(16):5076-5077. doi:10.1021/ja049271q

125. Blanch EW, Morozova-Roche LA, Cochran DAE, Doig AJ, Hecht L, Barron LD. Is polyproline II helix the killer conformation? A Raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme. *J Mol Biol*. 2000;301(2):553-563. doi:10.1006/jmbi.2000.3981

126. Jacob CR, Luber S, Reiher M. Calculated raman optical activity signatures of tryptophan side chains. *ChemPhysChem*. 2008;9(15):2177-2180. doi:10.1002/cphc.200800448

127. Briggs J, Panfili PR. Quantitation of DNA and Protein Impurities in Biopharmaceuticals. *Anal*

Chem. 1991;63(9):850-859. doi:10.1021/ac00009a003

128. Kostamovaara J, Tenhunen J, Kögler M, Nissinen I, Nissinen J, Keränen P. Fluorescence suppression in Raman spectroscopy using a time-gated CMOS SPAD. *Opt Express*. 2013;21(25):31632. doi:10.1364/OE.21.031632

129. McPhie P, Lakowicz JR. *Principles of Fluorescence Spectroscopy*.; 2000. doi:10.1007/978-0-387-46312-4

130. Praktikum P, Quenching F. Fluorescence Quenching Studies. *Phys Prakt*. 2016;(1):1-14. https://www.chem.uzh.ch/de/study/download/year2/che211.html.

131. Zi J, Michalska A. Vibrational Spectroscopy Photobleaching as a useful technique in reducing of fluorescence in Raman spectra of blue automobile paint samples. 2014;74:6-12. doi:10.1016/j.vibspec.2014.06.007

132. Irish DD, Adams EM. Apparatus and Methods. *Am Ind Hyg Assoc Q*. 1940;1(1):1-5. doi:10.1080/00968204009343768

133. Analysis of human skin tissue by Raman microspectroscopy: Dealing with the background. *Vib Spectrosc*. 2012;61:124-132. doi:10.1016/j.vibspec.2012.03.009

APPENDICES

A MATLAB CODES

A.1 Conversion from ATR to transmission from a transmission spectrum

```
clear
%% REFRACTIVE INDEX CALCULATION ACROSS AN ABSORPTION BAND by means
of KK transform
%% Variables for user
pathATR='ATR_lys__WATERSUBTR_VAPOURSUBTR_ZEROID.xlsx';
pathTRANS='TRANS_lys__ZEROID_SCALED_AVERAGE_1(50 faked).xlsx';
pathTRANSprot='TRANS_lys__ZEROID_SCALED_AVERAGE__WATERSUBTR_VAPOURSU
BTR_ZEROID.xlsx';
lowerlim=1500;
upperlim=1800;
a=[lowerlim:upperlim];
lowerlimfit=1500;
upperlimfit=1800;
plotting_checks=1;
angle=4.18
refractive_crys=2.403
Protein_atr_concentration=0.003079
Protein_trans_concentration=0.006158
%% ATR and TRANSMISSION spectra to import and trim
[ATR]=ATRdata(pathATR,lowerlim,upperlim,plotting_checks,a)
[wavenumber,TRANS,TRANS1]=TRANSdata(pathTRANS,lowerlimfit,upperlimfi
t,plotting_checks)
[TRANSprot,x_axis]=TRANSprot(pathTRANSprot,lowerlim,upperlim,plottin
g_checks)
%% Fitting transmission bands to express as a function for the
purpose of calculation of refractive index
[coeff,sumdiff,Iteration,COEFF,A,U,W,INTERCEPT,yfit]=fitting(wavenum
ber,TRANS,plotting_checks)
%% Refractive index calculation
[pathlength,Refractive_Sample,u]=refractive(a,A,U,W,INTERCEPT,TRANS1
,Iteration)
%% Conversion from ATR to transmission
[pathlength_atr_averaged]=conversion(yfit,TRANS,Refractive_Sample,a,
ATR,x_axis,wavenumber,TRANSprot,pathlength,angle,refractive_crys,Pro
tein_atr_concentration,Protein_trans_concentration,Iteration,A,W,U,I
NTERCEPT)
```

```
function [ATR]=ATRdata(pathATR,lowerlim,upperlim,plotting_checks,a)
%% First import the ATR experimental data
ATR=xlsread(pathATR);
%% Select region of interest 1500-1780 cm-1 in the original spectrum
%%to plot along with the fitted one in the end and zero it
n=0;
for i=1:length(ATR);
    u=ATR(i,1);
    n=n+1;
   if u>lowerlim
       break
   end
```

```
end

k=0;
for i=1:length(ATR);
    y=ATR(i,1);
    k=k+1;
   if y>upperlim
       break
   end
end

TRIMMED_DATA=ATR(n:k,:);
ATR=TRIMMED_DATA(:,2)';
x_axis=TRIMMED_DATA(:,1)';

% % Build a vector with the same length out of the fitting to
represent
% %%the ATR spectrum in
equation(iteration)=ATR./(dp.*REFRACTIVE_INDEX)
ATR=interp1(x_axis,ATR,a,'spline')
%% check plotts
if plotting_checks==1;
figure(1)
plot(a,ATR)
ylabel('abs')
xlabel('wavenumber')
axis([lowerlim upperlim 0 max(ATR)+0.005])
end


function
[wavenumber,TRANS,TRANS1]=TRANSdata(pathTRANS,lowerlimfit,upperlimfi
t,plotting_checks)
%% First import the TRANSMISSION experimental data OF THE PROTEIN IN
WATER
TRANS1=xlsread(pathTRANS)
%% Select region of interest 1500-1780 cm-1 in the original spectrum
to plot along with the fitted one in the end and zero it
n=0;
for i=1:length(TRANS1);
    u=TRANS1(i,1);
    n=n+1;
   if u>lowerlimfit
       break
   end
end

k=0;
for i=1:length(TRANS1);
    y=TRANS1(i,1);
    k=k+1;
   if y>upperlimfit
       break
   end
end

TRIMMED_DATA=TRANS1(n:k,:);
TRANS=TRIMMED_DATA(:,2);
wavenumber=TRIMMED_DATA(:,1);

%% plotting checks
```

```
if plotting_checks==1
figure(2)
plot(wavenumber,TRANS)
ylabel('abs')
xlabel('wavenumber')
axis([lowerlimfit upperlimfit 0 max(TRANS)+0.005])
end
end

function
[TRANSprot,x_axis]=TRANSprot(pathTRANSprot,lowerlim,upperlim,plottin
g_checks)
%% First import the TRANSMISSION experimental data
TRANS2=xlsread(pathTRANSprot)
%% Select region of interest 1500-1780 cm-1 in the original
spectrumto plot along with the fitted one in the end and zero it
n=0;
for i=1:length(TRANS2);
    u=TRANS2(i,1);
    n=n+1;
   if u>lowerlim
       break
   end
end

k=0;
for i=1:length(TRANS2);
    y=TRANS2(i,1);
    k=k+1;
   if y>upperlim
       break
   end
end

TRIMMED_DATA=TRANS2(n:k,:);
TRANSprot=TRIMMED_DATA(:,2);
x_axis=TRIMMED_DATA(:,1);

%% plotting checks
if plotting_checks==1
figure(2)
plot(x_axis,TRANSprot)
ylabel('abs')
xlabel('wavenumber')
axis([lowerlim upperlim 0 max(TRANSprot)+0.005])
end
end

function
[coeff,sumdiff,Iteration,COEFF,A,U,W,INTERCEPT,yfit]=fitting(wavenum
ber,TRANS,plotting_checks,a)
%% Fitting

x=wavenumber
y=TRANS


[pks,locs] = findpeaks(y,x)
for i=1:10
u=locs(1)-5;
v=locs(2)-5;
```

```
myfittype=fittype('(A1*w1/(w1^2+(x-u1)^2)+A2*w2/(w2^2+(x-
u2)^2)+A3*w3/(w3^2+(x-u3)^2)+A4*w4/(w4^2+(x-u4)^2)+A5*w5/(w5^2+(x-
u5)^2))+b');
myfit=fit(x,y,myfittype,'StartPoint',[20 20 20 40 40 0 u+i u+3*i
u+5*i v-i v+3*i 20 20 20 40 40],'Lower',[0 0 0 0 0 0 1500 1500 1500
1500 1500 0 0 0 0 0]);%[20 20 20 40 40 0 u+i u+3*i u+5*i v+i v+3*i
20 20 20 40 40]
yfit=feval(myfit,x);

sumdiff(i)=sum((yfit-y).^2);
coeff(i,:)=coeffvalues(myfit)

if plotting_checks==1
    figure(25)
    subplot(2,5,i)
    plot(myfit,x,y)
    figure(26)
    plot(sumdiff);
    axis on
end
end

Iteration=find(sumdiff==min(sumdiff))
 %% Pile coefficients into separate vectors by nature
COEFF=coeff(Iteration,:);
A=zeros(1,5)
for i=1:length(A)
A(i)=COEFF(1,i)
end
U=zeros(1,5)
for i=1:length(U)
U(i)=COEFF(1,i+6)
end
W=zeros(1,5)
for i=1:length(W)
W(i)=COEFF(1,i+11)
end
INTERCEPT=COEFF(1,6)


end



function
[pathlength,Refractive_Sample,u]=refractive(a,A,U,W,INTERCEPT,TRANS1
,Iteration)
%% Next step was to add up all the peaks that I fitted:
    %sum A(i)*W(i)/(W(i)+(x-U(i))^2 +INTERCEPT

        %A==intensity related
        %W==width related
        %U==position of the peak
        %INTERCEPT==intercept
syms x
L=0;

for     i=1:length(A)
        L=(A(i).*W(i))./((x-U(i)).^2+W(i)^2)+L
        %I plotted the result
```

```
fplot(x,L,[1500,1800])

end

Jtrans=L+INTERCEPT;


%% Then I multiplied the Lorenztians by the other term and
% %integrated by the Cauchy Principal Value as follows:
%% Find the pathlength of the transmission spectrum
n=0;
for i=1:length(TRANS1);
    u=TRANS1(i,1);
    n=n+1;
   if u>1900
       break
   end
end

k=0;
for i=1:length(TRANS1);
    y=TRANS1(i,1);
    k=k+1;
   if y>2400
       break
   end
end

transtrimmed=TRANS1(n:k,:);
xx=transtrimmed(:,1);
yy=transtrimmed(:,2);
MAX=max(yy);
pathlength=(0.1/1.9254879)*MAX;
%% Calculation of refractive index by KK integral

for i=1:length(a)
E(i)=[(1)./(x.^2-a(i)^2)];
Y(i)=Jtrans.*E(i);
u(i)=int(Y(i),x,0,inf,'PrincipalValue',true);
cpv(i)=((log(10)*(10/pathlength))/(2*pi^2))*u(i);
n(i)=1.33+cpv(i);
%the next step was just to take the real part:
Refractive_Sample(i)=real(n(i))
end
end



function
[pathlength_atr_averaged]=conversion(yfit,TRANS,Refractive_Sample,a,
ATR,x_axis,wavenumber,TRANSprot,pathlength,angle,refractive_crys,Pro
tein_atr_concentration,Protein_trans_concentration,Iteration,A,W,U,I
NTERCEPT)
%% WORK OUT THE DEPTH OF PENETRATION
dp=((1./a).*(1./(2*pi*refractive_crys*((sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2)).^0.5))

%% POLARIZATION TERMS
s=4*(cos(pi/angle)).^2./(1-(Refractive_Sample/refractive_crys).^2);
```

```
p=(4*(cos(pi/angle)).^2*((sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2)+4*(cos(pi/angle))^2*(sin(pi/
angle))^2)./((1-
(Refractive_Sample/refractive_crys).^2).*((1+(Refractive_Sample/refr
active_crys).^2).*(sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2));
t=(p+s)/2

%% FIRST TERM OF THE TAYLOR EXPANSION OF THE FIRST TERM OF THE FIRST
TAYLOR EXPANSION
pathlength_atr=((Refractive_Sample/refractive_crys).*dp*(1/cos(pi/an
gle)).*t)
pathlength_atr_averaged=mean(pathlength_atr)
figure(66)
plot(a,pathlength_atr)

extinction_coeff_simulated_TRANS=2*ATR./(Protein_atr_concentration*(
Refractive_Sample/refractive_crys).*dp*(1/cos(pi/angle)).*t);

extinction_coeff_experimental_TRANS=TRANSprot/(Protein_trans_concent
ration*pathlength)

extinction_coeff_experimental_ATR=(max(extinction_coeff_experimental
_TRANS)/max(ATR))*ATR


%plotting of generated transmission along with original atr and
experimental transmission
figure(17)
plot(a,extinction_coeff_simulated_TRANS,'r.','LineWidth',1)
hold on
plot(a,extinction_coeff_experimental_ATR,'k--','LineWidth',3)
plot(x_axis,extinction_coeff_experimental_TRANS,'g','LineWidth',2)
legend({'Calculated transmission' 'Experimental ATR','Experimental
transmission'},'FontSize',8,'Position',[0.68 0.73 0.1
0.1],'FontWeight','bold','FontName','Times')
xlabel('Wavenumber (cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
ylabel('Extinction coefficient (M.cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
% title('Harrick')
axis([1500 1800 0 max(extinction_coeff_experimental_TRANS)+500])
set(gca,'YTick',[0:1000:max(extinction_coeff_experimental_TRANS)])
annotation('textbox',[0.27,0.68,0.1,0.1],'String',"Amide
II",'LineStyle','none','FontWeight','bold')
annotation('textbox',[0.55,0.85,0.1,0.1],'String',"Amide
I",'LineStyle','none','FontWeight','bold')
legend boxoff
box off
saveas(gcf,'Harrick.svg')
%% FIRST TERM OF THE FIRST TAYLOR EXPANSION

extinction_coeff_simulated_TRANS=2*(1-10.^(-
ATR))./(log(10)*Protein_atr_concentration*(Refractive_Sample/refract
ive_crys).*dp*(1/cos(pi/angle)).*t)

%plotting of generated transmission along with original atr and
experimental transmission
figure(18)
plot(a,extinction_coeff_simulated_TRANS,'m.','LineWidth',1)
hold on
```

```
plot(a,extinction_coeff_experimental_ATR,'k--','LineWidth',3)
plot(x_axis,extinction_coeff_experimental_TRANS,'g','LineWidth',2)
legend({'Calculated transmission' 'Experimental ATR','Experimental
transmission'},'FontSize',8,'Position',[0.68 0.73 0.1
0.1],'FontWeight','bold','FontName','Times')
xlabel('Wavenumber (cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
ylabel('Extinction coefficient (M.cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
% title('Milosevic')
axis([1500 1800 0 max(extinction_coeff_experimental_TRANS)+500])
set(gca,'YTick',[0:1000:max(extinction_coeff_experimental_TRANS)])
annotation('textbox',[0.27,0.68,0.1,0.1],'String',"Amide
II",'LineStyle','none','FontWeight','bold')
annotation('textbox',[0.55,0.85,0.1,0.1],'String',"Amide
I",'LineStyle','none','FontWeight','bold')
legend boxoff
box off
saveas(gcf,'Milosevic.svg')

%% FULL EQUATION

extinction_coeff_simulated_TRANS=-
((2./(log(10)*Protein_atr_concentration*dp)).*log(1-(1-10.^(-
ATR))./(t.*(Refractive_Sample/refractive_crys)*(1/cos(pi/angle)))))

%plotting of generated transmission along with original atr and
%experimental transmission

figure(19)
plot(a,extinction_coeff_simulated_TRANS,'b.','LineWidth',2)
hold on
plot(a,extinction_coeff_experimental_ATR,'k--','LineWidth',3)
plot(x_axis,extinction_coeff_experimental_TRANS,'g','LineWidth',2)
legend({'Calculated transmission' 'Experimental ATR','Experimental
transmission'},'FontSize',8,'Position',[0.68 0.73 0.1
0.1],'FontWeight','bold','FontName','Times')
xlabel('Wavenumber (cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
ylabel('Extinction coefficient (M.cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
%title('Full equation (mine)')
axis([1500 1800 0 max(extinction_coeff_experimental_TRANS)+500])
set(gca,'YTick',[0:1000:max(extinction_coeff_experimental_TRANS)])
annotation('textbox',[0.27,0.68,0.1,0.1],'String',"Amide
II",'LineStyle','none','FontWeight','bold')
annotation('textbox',[0.55,0.85,0.1,0.1],'String',"Amide
I",'LineStyle','none','FontWeight','bold')
legend boxoff
box off
saveas(gcf,'Mine.svg')
%csvwrite('data.csv',[a' TRANSPROT' Refractive_Sample'])


figure(Iteration)
L(1)=plot(wavenumber,TRANS,'k-.','LineWidth',3)
hold on
L(2)=plot(wavenumber,yfit,'r-','LineWidth',1)
xlabel('Wavenumber (cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
```

```
ylabel('Extinction coefficient (M.cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
axis([1500 1800 0 1.5])
set(gca,'YTick',[0:0.2:1.5])

yyaxis right
L(3)=plot(a,Refractive_Sample,'k-','LineWidth',2)
ylabel('Refractive
index','FontSize',10,'FontWeight','bold','FontName','Times')
axis([1500 1800 1.05 1.4])
set(gca,'YTick',[1.05:0.05:1.4])


for i=1:length(A)
LOR=(A(i).*W(i))./((a-U(i)).^2+W(i)^2)+INTERCEPT
yyaxis left
L(4)=plot(a,LOR,'b-','LineWidth',2)
axis([1500 1800 0 1.5])
set(gca,'YTick',[0:0.2:1.5])
end

legend(L,{'Experimental transmission','Fitted curve','Refractive
index','Fit peaks'},'FontSize',8,'Position',[0.28 0.82 0.1
0.1],'FontWeight','bold','FontName','Times')
legend boxoff
box off
saveas(gcf,'graph.svg')
end
```

A.2 Conversion from ATR to transmission from an ATR spectrum iteratively

```
 clear
%% REFRACTIVE INDEX CALCULATION ACROSS AN ABSORPTION BAND by meansof
KK transform
%% Variables for user
pathATR_PROT='ATR_lys__WATERSUBTR_VAPOURSUBTR_ZEROID.xlsx';
pathATR_WP='lys_vapoursubtra.xlsx';
pathTRANS_prot='TRANS_lys__ZEROID_SCALED_AVERAGE__WATERSUBTR_VAPOURS
UBTR_ZEROID.xlsx';
lowerlim=1500;
upperlim=2300;
a=[lowerlim:upperlim];
plotting_checks=1;
angle=4.18
refractive_crys=2.403
Protein_atr_concentration=0.003079
Protein_trans_concentration=0.003079*2
NumIter=1; %% ATR and TRANSMISSION
%% Import data
[ATR_PROT]=ATR_PROT(pathATR_PROT,lowerlim,upperlim,plotting_checks,a
)
[ATR_PW,ATR_PW1]=ATR_PROT_WATER(pathATR_WP,lowerlim,upperlim,plottin
g_checks,a)
[TRANSprot,x_axis]=TRANS_prot(pathTRANS_prot,lowerlim,upperlim,plott
ing_checks)
%% First Calculation
```

```
[coeff,Iteration,COEFF,A,U,W,INTERCEPT]=fitting(a,ATR_PW1,plotting_c
hecks)%% Refractive index calculation
[pathlength,Refractive_Sample]=refractive(a,A,U,W,INTERCEPT,ATR_PW)
[e_exp_TRANS_prot,e_exp_ATR_PROT,corrected_ATR_WATER,ce_sim_TRANS_wa
ter_prot]=conversion(ATR_PW1,yfit,Refractive_Sample,a,ATR_PROT,x_axi
s,wavenumber,TRANSprot,pathlength,angle,refractive_crys,Protein_atr_
concentration,Protein_trans_concentration,Iteration,A,W,U,INTERCEPT)
%% Iterative
iterative(ATR_PW1,a,ce_sim_TRANS_water_prot,plotting_checks,correcte
d_ATR_WATER,x_axis,ATR_PROT,e_exp_TRANS_prot,e_exp_ATR_PROT,refracti
ve_crys,angle,Protein_atr_concentration,NumIter)



 function
[ATR_PROT]=ATR_PROT(pathATR_PROT,lowerlim,upperlim,plotting_checks,a
)
%% First import the ATR experimental data
ATR=xlsread(pathATR_PROT)
%% Select region of interest 1500-1780 cm-1 in the original
spectrumto plot along with the fitted one in the end and zero it
n=0;
for i=1:length(ATR);
    u=ATR(i,1);
    n=n+1;
   if u>lowerlim
       break
   end
end

k=0;
for i=1:length(ATR);
    y=ATR(i,1);
    k=k+1;
   if y>upperlim
       break
   end
end

TRIMMED_DATA=ATR(n:k,:);
ATR=TRIMMED_DATA(:,2)';
x_axis=TRIMMED_DATA(:,1)';

ATR_PROT=interp1(x_axis,ATR,a,'spline');
%% plotting checks
if plotting_checks==1
figure(1)
plot(a,ATR_PROT)
ylabel('abs')
xlabel('wavenumber')
axis([lowerlim upperlim 0 max(ATR_PROT)+0.005])
end
end


function
[ATR_PW,ATR_PW1]=ATR_PROT_WATER(pathATR_WP,lowerlim,upperlim,plottin
g_checks,a)
%% First import the water + prot atr experimental data
ATR_PW=xlsread(pathATR_WP)
```

```
%% Select region of interest 1500-1780 cm-1 in the original
spectrumto plot along with the fitted one in the end and zero it
n=0;
for i=1:length(ATR_PW);
    u=ATR_PW(i,1);
    n=n+1;
   if u>lowerlim
       break
   end
end

k=0;
for i=1:length(ATR_PW);
    y=ATR_PW(i,1);
    k=k+1;
   if y>upperlim
       break
   end
end

TRIMMED_DATA=ATR_PW(n:k,:);
ATR_PW1=TRIMMED_DATA(:,2);
wavenumber=TRIMMED_DATA(:,1);

ATR_PW1=interp1(wavenumber,ATR_PW1,a,'spline');


%% plotting checks
if plotting_checks==1
figure(2)
plot(a,ATR_PW1)
ylabel('abs')
xlabel('wavenumber')
axis([lowerlim upperlim 0 max(ATR_PW1)+0.005])
end
end
```

```
 function
[TRANSprot,x_axis]=TRANS_prot(pathTRANS_prot,lowerlim,upperlim,plott
ing_checks)
%% First import the TRANSMISSION experimental data
TRANS2=xlsread(pathTRANS_prot)
%% Select region of interest 1500-1780 cm-1 in the original
spectrumto plot along with the fitted one in the end and zero it
n=0;
for i=1:length(TRANS2);
    u=TRANS2(i,1);
    n=n+1;
   if u>lowerlim
       break
   end
end

k=0;
for i=1:length(TRANS2);
    y=TRANS2(i,1);
    k=k+1;
   if y>upperlim
```

```
       break
   end
end

TRIMMED_DATA=TRANS2(n:k,:);
TRANSprot=TRIMMED_DATA(:,2);
x_axis=TRIMMED_DATA(:,1);

%% plotting checks
if plotting_checks==1
figure(2)
plot(x_axis,TRANSprot)
ylabel('abs')
xlabel('wavenumber')
axis([lowerlim upperlim 0 max(TRANSprot)+0.005])
end
end


 function
[coeff,Iteration,COEFF,A,U,W,INTERCEPT]=fitting(a,ATR_PW1,plotting_c
hecks)
%% Fitting

x=a';
y=ATR_PW1';
myfittype=fittype('A1*w1/(w1+(x-u1)^2)+A2*w2/(w2+(x-
u2)^2)+A3*w3/(w3+(x-u3)^2)+A4*w4/(w4+(x-u4)^2)+A5*w5/(w5+(x-
u5)^2)+A6*w6/(w6+(x-u6)^2)+A6*w6/(w6+(x-u6)^2)+A7*w7/(w7+(x-
u7)^2)+b');
%%Search for valid initial values
[pks,locs] = findpeaks(y,x)
for i=1:30
u=locs(1)-5;
v=locs(2)-5;
o=locs(4)-5;
myfit=fit(x,y,myfittype,'StartPoint',[20 20 20 40 40 60 60 -3 u+i
u+3*i u+5*i v-i v+3*i o+i o+i*3 20 20 20 40 40 60 60],'Lower',[0 0 0
0 0 0 0 0 1500 1500 1500 1600 1600 1800 1800 0 0 0 0 60 120 120])

yfit=feval(myfit,x);
sumdiff(i)=sum(abs(yfit-y));
coeff(i,:)=coeffvalues(myfit);

if plotting_checks==1
    figure(5)
    subplot(3,10,i)
plot(myfit,x,y)
figure(111)
plot(sumdiff);
axis on
end
end

Iteration=find(sumdiff==min(sumdiff))
 %% Pile coefficients into separate vectors by nature
COEFF=coeff(Iteration,:);
A=zeros(1,7)
for i=1:length(A)
A(i)=COEFF(1,i)
end
```

```
U=zeros(1,7)
for i=1:length(U)
U(i)=COEFF(1,i+8)
end
W=zeros(1,7)
for i=1:length(W)
W(i)=COEFF(1,i+15)
end
INTERCEPT=COEFF(1,8)
end

function
[pathlength,Refractive_Sample]=refractive(a,A,U,W,INTERCEPT,ATR_PW)
%% Next step was to add up all the peaks that I fitted:
    %sum A(i)*W(i)/(W(i)+(x-U(i))^2 +INTERCEPT

        %A==intensity related
        %W==width related
        %U==position of the peak
        %INTERCEPT==intercept
syms x
L=0;

for     i=1:length(A)
        L=(A(i).*W(i))./((x-U(i)).^2+W(i))+L

end

Jtrans=L+INTERCEPT;

%I plotted the result
figure(12)
fplot(x,Jtrans,[1500,2300])

%% Find the pathlength of the atr spectrum
n=0;
for i=1:length(ATR_PW);
    u=ATR_PW(i,1);
    n=n+1;
   if u>1900
       break
   end
end

k=0;
for i=1:length(ATR_PW);
    y=ATR_PW(i,1);
    k=k+1;
   if y>2400
       break
   end
end

trimmed=ATR_PW(n:k,:);
yy=trimmed(:,2);
MAX=max(yy);
pathlength=(0.1/1.921653)*MAX;
%% Calculation of refractive index by KK integral

for i=1:length(a)
E=[(1)./(x.^2-a(i)^2)];
```

```
Y=Jtrans.*E;
u=int(Y,x,0,inf,'PrincipalValue',true);
cpv(i)=((log(10)*(10/pathlength))/(2*pi^2))*u;
n(i)=1.33+cpv(i);
%the next step was just to take the real part:
Refractive_Sample(i)=real(n(i))
end
end



function
[e_exp_TRANS_prot,e_exp_ATR_PROT,corrected_ATR_WATER,ce_sim_TRANS_wa
ter_prot]=conversion(ATR_PW1,yfit,Refractive_Sample,a,ATR_PROT,x_axi
s,wavenumber,TRANSprot,pathlength,angle,refractive_crys,Protein_atr_
concentration,Protein_trans_concentration,Iteration,A,W,U,INTERCEPT)

%% WORK OUT THE DEPTH OF PENETRATION
dp=((1./a).*(1./(2*pi*refractive_crys*((sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2)).^0.5))

%% POLARIZATION TERMS
s=4*(cos(pi/angle)).^2./(1-(Refractive_Sample/refractive_crys).^2);
p=(4*(cos(pi/angle)).^2*((sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2)+4*(cos(pi/angle))^2*(sin(pi/
angle))^2)./((1-
(Refractive_Sample/refractive_crys).^2).*((1+(Refractive_Sample/refr
active_crys).^2).*(sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2));
t=(p+s)/2

%%


e_exp_TRANS_prot=TRANSprot/(0.0065392*Protein_trans_concentration)

e_exp_ATR_PROT=(max(e_exp_TRANS_prot)/max(ATR_PROT))*ATR_PROT


%% FULL EQUATION

e_sim_TRANS_prot=-
((2./(log(10).*dp.*Protein_atr_concentration)).*log(1-(1-10.^(-
ATR_PROT)).*(t.*(Refractive_Sample/refractive_crys)*(1/cos(pi/angle)
))))

ce_sim_TRANS_water_prot=-((2./(log(10).*dp)).*log(1-(1-10.^(-
ATR_PW1)).*(t.*(Refractive_Sample/refractive_crys)*(1/cos(pi/angle))
)))

corrected_ATR_WATER=[a' ce_sim_TRANS_water_prot']


%plotting of generated transmission along with original atr and
%experimental transmission

figure(19)
plot(a,e_sim_TRANS_prot,'b.','LineWidth',2)
hold on
plot(a,e_exp_ATR_PROT,'k--','LineWidth',3)
plot(x_axis,e_exp_TRANS_prot,'g','LineWidth',2)
```

```
legend({'Calculated transmission' 'Experimental ATR','Experimental
transmission'},'FontSize',8,'Position',[0.68 0.73 0.1
0.1],'FontWeight','bold','FontName','Times')
xlabel('Wavenumber (cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
ylabel('Extinction coefficient (M.cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
%title('Full equation (mine)')
axis([1500 1800 0 max(e_exp_TRANS_prot)+500])
set(gca,'YTick',[0:1000:max(e_exp_TRANS_prot)])
annotation('textbox',[0.27,0.68,0.1,0.1],'String',"Amide
II",'LineStyle','none','FontWeight','bold')
annotation('textbox',[0.55,0.85,0.1,0.1],'String',"Amide
I",'LineStyle','none','FontWeight','bold')
legend boxoff
box off
% saveas(gcf,'Mine.svg')
% %csvwrite('data.csv',[a' TRANSPROT' Refractive_Sample'])
%
%
% figure(Iteration)
% L(1)=plot(wavenumber,TRANS,'k-.','LineWidth',3)
% hold on
% L(2)=plot(wavenumber,yfit,'r-','LineWidth',1)
% xlabel('Wavenumber (cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
% ylabel('Extinction coefficient (M.cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
% axis([1500 1800 0 1.5])
% set(gca,'YTick',[0:0.2:1.5])
%
% yyaxis right
% L(3)=plot(a,Refractive_Sample,'k-','LineWidth',2)
% ylabel('Refractive
index','FontSize',10,'FontWeight','bold','FontName','Times')
% axis([1500 1800 1.05 1.4])
% set(gca,'YTick',[1.05:0.05:1.4])
%
%
% for i=1:length(A)
% LOR=(A(i).*W(i))./((a-U(i)).^2+W(i)^2)+INTERCEPT
% yyaxis left
% L(4)=plot(a,LOR,'b-','LineWidth',2)
% axis([1500 1800 0 1.5])
% set(gca,'YTick',[0:0.2:1.5])
% end
%
% legend(L,{'Experimental transmission','Fitted curve','Refractive
index','Fit peaks'},'FontSize',8,'Position',[0.28 0.82 0.1
0.1],'FontWeight','bold','FontName','Times')
% legend boxoff
% box off
% saveas(gcf,'graph.svg')
end



 function
iterative(ATR_PW1,a,ce_sim_TRANS_water_prot,plotting_checks,correcte
d_ATR_WATER,x_axis,ATR_PROT,e_exp_TRANS_prot,e_exp_ATR_PROT,refracti
ve_crys,angle,Protein_atr_concentration,NumIter)
```

```
for i=1:NumIter
%% Iterative fitting
[coeff,Iteration,COEFF,A,U,W,INTERCEPT]=fitting2(a,ce_sim_TRANS_wate
r_prot,plotting_checks)
[pathlength,Refractive_Sample]=refractive7(a,A,U,W,INTERCEPT,correct
ed_ATR_WATER)
[corrected_ATR_WATER,ce_sim_TRANS_water_prot]=conversion2(ATR_PW1,x_
axis,ATR_PROT,Refractive_Sample,a,e_exp_TRANS_prot,e_exp_ATR_PROT,ce
_sim_TRANS_water_prot,refractive_crys,angle,Protein_atr_concentratio
n)
iter=num2str(i+1)
label=strcat('iteration_',iter)
dlmwrite(strcat(label,'.txt'),ce_sim_TRANS_water_prot,'delimiter','\
t')
end
end



function
[coeff,Iteration,COEFF,A,U,W,INTERCEPT]=fitting2(a,ce_sim_TRANS_wate
r_prot,plotting_checks)
%% Fitting

x=a';
y=ce_sim_TRANS_water_prot';
myfittype=fittype('A1*w1/(w1+(x-u1)^2)+A2*w2/(w2+(x-
u2)^2)+A3*w3/(w3+(x-u3)^2)+A4*w4/(w4+(x-u4)^2)+A5*w5/(w5+(x-
u5)^2)+A6*w6/(w6+(x-u6)^2)+A6*w6/(w6+(x-u6)^2)+A7*w7/(w7+(x-
u7)^2)+b');
%%Search for valid initial values
[pks,locs] = findpeaks(y,x)
for i=1:30
u=locs(1)-5;
v=locs(2)-5;
% o=locs(4)-5;
myfit=fit(x,y,myfittype,'StartPoint',[20 20 20 40 40 60 60 -3 u+i
u+3*i u+5*i v-i v+3*i 2100+i 2100+i*3 20 20 20 40 40 60
60],'Lower',[0 0 0 0 0 0 0 0 1500 1500 1500 1600 1600 1800 1800 0 0
0 0 60 120 120])

yfit=feval(myfit,x);
sumdiff(i)=sum(abs(yfit-y));
coeff(i,:)=coeffvalues(myfit);

if plotting_checks==1
    figure(5)
    subplot(3,10,i)
plot(myfit,x,y)
figure(111)
plot(sumdiff);
axis on
end
end

Iteration=find(sumdiff==min(sumdiff))
 %% Pile coefficients into separate vectors by nature
COEFF=coeff(Iteration,:);
A=zeros(1,7)
for i=1:length(A)
A(i)=COEFF(1,i)
```

```
end
U=zeros(1,7)
for i=1:length(U)
U(i)=COEFF(1,i+8)
end
W=zeros(1,7)
for i=1:length(W)
W(i)=COEFF(1,i+15)
end
INTERCEPT=COEFF(1,8)
end



function
[pathlength,Refractive_Sample]=refractive7(a,A,U,W,INTERCEPT,correct
ed_ATR_WATER)
%% Next step was to add up all the peaks that I fitted:
    %sum A(i)*W(i)/(W(i)+(x-U(i))^2 +INTERCEPT

        %A==intensity related
        %W==width related
        %U==position of the peak
        %INTERCEPT==intercept
syms x
L=0;

for     i=1:length(A)
        L=(A(i).*W(i))./((x-U(i)).^2+W(i))+L

end

Jtrans=L+INTERCEPT;

%I plotted the result
figure(12)
fplot(x,Jtrans,[1500,2300])
%% Then I multiplied the Lorenztians by the other term and
% %integrated by the Cauchy Principal Value as follows:

%% Find the pathlength of the atr spectrum
n=0;
for i=1:length(corrected_ATR_WATER);
    u=corrected_ATR_WATER(i,1);
    n=n+1;
   if u>1900
       break
   end
end

k=0;
for i=1:length(corrected_ATR_WATER);
    y=corrected_ATR_WATER(i,1);
    k=k+1;
   if y>2299
       break
   end
end

trimmed=corrected_ATR_WATER(n:k,:);
xx=trimmed(:,1);
```

```
yy=trimmed(:,2);
MAX=max(yy);
pathlength=(0.1/1.921653)*MAX;
%% Calculation of refractive index by KK integral

for i=1:length(a)
E=[(1)./(x.^2-a(i)^2)];
Y=Jtrans.*E;
u=int(Y,x,0,inf,'PrincipalValue',true);
cpv(i)=((log(10)*(10/pathlength))/(2*pi^2))*u;
n(i)=1.33+cpv(i);
%the next step was just to take the real part:
Refractive_Sample(i)=real(n(i))
end
end



function
[corrected_ATR_WATER,ce_sim_TRANS_water_prot]=conversion2(ATR_PW1,x_
axis,ATR_PROT,Refractive_Sample,a,e_exp_TRANS_prot,e_exp_ATR_PROT,ce
_sim_TRANS_water_prot,refractive_crys,angle,Protein_atr_concentratio
n)

%% WORK OUT THE DEPTH OF PENETRATION
dp=((1./a).*(1./(2*pi*refractive_crys*((sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2)).^0.5))

%% POLARIZATION TERMS
s=4*(cos(pi/angle)).^2./(1-(Refractive_Sample/refractive_crys).^2);
p=(4*(cos(pi/angle)).^2*((sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2)+4*(cos(pi/angle))^2*(sin(pi/
angle))^2)./((1-
(Refractive_Sample/refractive_crys).^2).*((1+(Refractive_Sample/refr
active_crys).^2).*(sin(pi/angle)).^2-
(Refractive_Sample/refractive_crys).^2));
t=(p+s)/2

%% FULL EQUATION

e_sim_TRANS_prot=-
((2./(log(10).*dp.*Protein_atr_concentration)).*log(1-(1-10.^(-
ATR_PROT))./(t.*(Refractive_Sample/refractive_crys)*(1/cos(pi/angle)
))))

ce_sim_TRANS_water_prot=-((2./(log(10).*dp)).*log(1-(1-10.^(-
ATR_PW1))./(t.*(Refractive_Sample/refractive_crys)*(1/cos(pi/angle))
)))

corrected_ATR_WATER=[a' ce_sim_TRANS_water_prot']



%plotting of generated transmission along with original atr and
%experimental transmission

figure(19)
plot(a,e_sim_TRANS_prot,'b.','LineWidth',2)
hold on
plot(a,e_exp_ATR_PROT,'k--','LineWidth',3)
```

```
plot(x_axis,e_exp_TRANS_prot,'g','LineWidth',2)
legend({'Calculated transmission' 'Experimental ATR','Experimental
transmission'},'FontSize',8,'Position',[0.68 0.73 0.1
0.1],'FontWeight','bold','FontName','Times')
xlabel('Wavenumber (cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
ylabel('Extinction coefficient (M.cm)^{-
1}','FontSize',10,'FontWeight','bold','FontName','Times')
%title('Full equation (mine)')
axis([1500 1800 0 max(e_exp_TRANS_prot)+500])
set(gca,'YTick',[0:1000:max(e_exp_TRANS_prot)])
annotation('textbox',[0.27,0.68,0.1,0.1],'String',"Amide
II",'LineStyle','none','FontWeight','bold')
annotation('textbox',[0.55,0.85,0.1,0.1],'String',"Amide
I",'LineStyle','none','FontWeight','bold')
legend boxoff
box off
saveas(gcf,'Mine.svg')
%csvwrite('data.csv',[a' TRANSPROT' Refractive_Sample'])

end
```

A.3 IR data processing

```
clear
%%USER PARAMETERS
PATH='C:\Users\u1566630\Desktop\Water TRANS AND
ATR\ATR\BSA_REP_1,3,7_ORIGINALS\'
Sample_label='bsa_'
Water_label='water_bsa_'
Vapour_label='-vapour_250_scans'
ext='.txt'

d=[1 3 7] %%number of replicates
vactors=[-0.01:0.001:0.01]%%factors for the subtraction of vapour
factors=[0.985:0.001:1.0]%%factors for the subtraction of water

water_correction=1
%------------
vapour_correction=1
%-------
zeroing=1
Protein=1
Water=0
%--------
Scaling=0
%---------
trimming=1
lowerlim=1000;
upperlim=4000;
%%


for j=1:length(d)
Sample_label_ext=strcat(Sample_label,num2str(d(j)),ext)
Sample_file =importdata(strcat(PATH,Sample_label_ext),'\t',19);
SAMPLE=Sample_file.data;
```

```
S=SAMPLE(:,2);
x=SAMPLE(:,1);

Water_label_ext=strcat(Water_label,num2str(d(j)),ext)
Water_file = importdata(strcat(PATH,Water_label_ext),'\t',19);
WATER=Water_file.data;
W=WATER(:,2);

Vapour_label_ext=strcat(Vapour_label,ext)
Vapour_file=importdata(strcat(PATH,Vapour_label_ext),'\t',19)
VAPOUR=Vapour_file.data;
V=VAPOUR(:,2)

%% POTENTIAL CORRECTIONS TO PERFORM

if pathlength==1
    [p]=path_length(S,WS,x,lactors)
end
if water_correction==1
[S]=Subtraction_of_water(S,W,x,factors)
end
if vapour_correction==1
[S]=Subtraction_of_vapour(V,S,x,vactors)
end
if zeroing==1
[S]=Zeroing(x,S,Protein,Water)
end
if Scaling==1
[S]=Scaling2(S,x)
end
if Conversion==1
 [e]=conversion_to_extinction_coeff(S,concentration,p)
end
if trimming==1
[S,x]=trim(S,x,lowerlim,upperlim)
end
TRIMMED(:,j)=S
plot(x,S)
MEAN=(mean(TRIMMED'))'
plot(x,MEAN)
end
%xlswrite(strcat(PATH,Sample_label,'_water_corrected_vap_corrected.x
lsx'),[x,MEAN])

function [p]=path_length(S,WS,x,lactors)
%% define range to fit and factors to correct the water subtraction
and trim data withing that range
lowerlim=1750;
upperlim=2300;
%
n=0;
for i=1:length(x)
    u=x(i)
    n=n+1
    if u>lowerlim
        break
    end
end
k=0;
for  i=1:length(x)
```

```
    v=x(i)
    k=k+1
    if v>upperlim
        break
    end
end
trimmed_y=S(n:k);
trimmed_x=x(n:k);
trimmed_z=WS(n:k);

%% iterative fitting of the baseline for the different factors
established above
for j=1:length(lactors)
trimmed_sub=trimmed_y-trimmed_z*lactors(j);
%
myfittype=fittype('(a+b*x)');
myfit=fit(trimmed_x,trimmed_sub,myfittype,'StartPoint',[-0.0012
7.99e-07])
figure(6)
plot(myfit,trimmed_x,trimmed_sub)
yfit = feval(myfit,trimmed_x);
hold on
xlabel('Wavenumber (cm^{-1})','FontSize',8)
ylabel('Abs','FontSize',8)
legend off
diff=trimmed_sub-yfit;
d(j)=sum(diff.^2);
legend({'Experimental data','Fitted
data'},'FontSize',8,'Position',[0.73 0.45 0.1 0.1])
legend boxoff
box off
end
hold off
%% find the factor with the smallest sum of residuals
MIN=min(d);
ITERATION=find(d==MIN);
LACTOR=lactors(ITERATION)
p=100*LACTOR %(100 um of thickness)

end

function [S]=Subtraction_of_water(S,W,x,factors)
%% define range to fit and factors to correct the water subtraction
and trim data withing that range
lowerlim=1850;
upperlim=2600;
%
n=0;
for i=1:length(x)
    u=x(i)
    n=n+1
    if u>lowerlim
        break
    end
end
k=0;
for  i=1:length(x)
    v=x(i)
    k=k+1
    if v>upperlim
        break
```

```
    end
end
trimmed_y=S(n:k);
trimmed_x=x(n:k);
trimmed_z=W(n:k);

%% iterative fitting of the baseline for the different factors
established above
for j=1:length(factors)
trimmed_sub=trimmed_y-trimmed_z*factors(j);
%
myfittype=fittype('(a+b*x+c*x^2)');
myfit=fit(trimmed_x,trimmed_sub,myfittype,'StartPoint',[-0.0012
7.99e-07 -1e-10])
figure(1)
plot(myfit,trimmed_x,trimmed_sub)
yfit = feval(myfit,trimmed_x);
hold on
xlabel('Wavenumber (cm^{-1})','FontSize',8)
ylabel('Abs','FontSize',8)
legend off
diff=trimmed_sub-yfit;
d(j)=sum(diff.^2);
legend({'Experimental data','Fitted
data'},'FontSize',8,'Position',[0.73 0.45 0.1 0.1])
legend boxoff
box off
end
hold off
%% find the factor with the smallest sum of residuals
MIN=min(d);
ITERATION=find(d==MIN);
FACTOR=factors(ITERATION)
%% Perform the subtraction with the right factor
S=S-W*FACTOR
figure(2)
plot(x,S)
hold on
end

function [S]=subtraction_of_vapour(V,S,x,vactors)
%% SUBTRACTION OF WATER VAPOUR
%% define range to fit and factors to correct the water subtraction
and trim data withing that range
lowerlim=3800; %% 1720-1800 for proteins %% 1500-1550 for water%%
3800-3900 for any other sample
upperlim=3900;

n=0;
for i=1:length(x)
    u=x(i)
    n=n+1
    if u>lowerlim
        break
    end
end
k=0;
for  i=1:length(x)
    v=x(i)
    k=k+1
```

```
    if v>upperlim
        break
    end
end
trimmed_y=S(n:k);
trimmed_x=x(n:k);
trimmed_v=V(n:k);


%% iterative fitting of the baseline for the different factors
established above

for j=1:length(vactors)

trimmed_sub=trimmed_y-trimmed_v*vactors(j);
%
myfittype=fittype('(a+c*x+d*x^2)');
myfit=fit(trimmed_x,trimmed_sub,myfittype,'StartPoint',[-0.0012 -
7.9e-7 1e-11])
figure(3)
plot(myfit,trimmed_x,trimmed_sub)
hold on
xlabel('Wavenumber (cm^{-1})','FontSize',8)
ylabel('Abs','FontSize',8)
yfit = feval(myfit,trimmed_x);
legend off
% axis([lowerlim upperlim -0.0006 0.0006])
diff=trimmed_sub-yfit;
d(j)=sum(diff.^2);
legend({'Experimental data','Fitted data'},'FontSize',8)
legend boxoff
box off
end
hold off

%% find the factor with the smallest sum of residuals
MIN=min(d);
ITERATION=find(d==MIN);
VACTOR=vactors(ITERATION)

%% Perform the subtraction with the right factor

S=S-VACTOR*V
figure(4)
plot(x,S)
hold on
plot(x,S)
ylabel('Abs','FontSize',8)
xlabel('Wavenumber (cm^{-1})','FontSize',8)
box off
end


function [S]=Zeroing(x,S,Protein,Water)
if Protein==1
lowerlim=1800;
upperlim=1900;

n=0;
```

```
for i=1:length(x)
    u=x(i)
    n=n+1
    if u>lowerlim
        break
    end
end
k=0;
for  i=1:length(x)
    v=x(i)
    k=k+1
    if v>upperlim
        break
    end
end
trimmed_y=S(n:k);
trimmed_x=x(n:k)

average=mean(trimmed_y)
S=S-average
end
if Water==1
lowerlim=4000;
upperlim=5000;

n=0;
for i=1:length(x)
    u=x(i)
    n=n+1
    if u>lowerlim
        break
    end
end
k=0;
for  i=1:length(x)
    v=x(i)
    k=k+1
    if v>upperlim
        break
    end
end
trimmed_y=S(n:k);
trimmed_x=x(n:k)

MIN=min(trimmed_y)
S=S-MIN
end
end
```

```
function [S]=Scaling2(S,x)
lowerlim=1600;
upperlim=1800;

n=0;
for i=1:length(x)
    u=x(i)
```

```
    n=n+1
    if u>lowerlim
        break
    end
end
k=0;
for  i=1:length(x)
    v=x(i)
    k=k+1
    if v>upperlim
        break
    end
end
trimmed_y=S(n:k);
trimmed_x=x(n:k)

MAX=max(trimmed_y)
S=S/MAX
end

function[e]=conversion_to_extinction_coeff(S,concentration,p)
e=S*10000/(concentration*p) %the pathlength was in um so we have to
divide it by 10000 to convert to cm
end


function [S,x]=trim(S,x,lowerlim,upperlim)
%% TRIM
n=0;
for i=1:length(x)
    u=x(i)
    n=n+1
    if u>lowerlim
        break
    end
end
k=0;
for  i=1:length(x)
    v=x(i)
    k=k+1
    if v>upperlim
        break
    end
end
S=S(n:k);
x=x(n:k);
end
```

A.4 Raman data processing

```
PATH='C:\Users\u1566630\Desktop\'
Sample_label='Original_trimmed'
ext='.xlsx'
u=[2342 1850 1508 1145]

SAMPLE=xlsread(strcat(PATH,Sample_label));
```

```
for i=2:19
S=SAMPLE([1:759],i);
B=SAMPLE([1:759],20);
SS=SAMPLE([1:767],i);
x=SAMPLE([1:759],1);
S=S/(SS(767)*SS(761))-1*B/(1070*800);
for j=1:length(u)
l=abs(x-u(j))
MIN=min(l)
d=l-MIN
I=find(d==0);
w(j)=x(I);
z(j)=S(I);
 end
  %
fitting_S= fit(w',z','cubicspline');
figure(3)
plot(fitting_S,w,z);
hold on
plot(x,S)
baseline= feval(fitting_S,x)
baselined=S-baseline
In_MRW=baselined*SS(765)/(SS(763))
Y(:,i-1)=In_MRW;
figure(5)
plot(x,baselined)
hold on
end

% % scaled= S*MRW/(Power*conc*Totalexp)

xlswrite(strcat(PATH,Sample_label,'corrected.xlsx'),[x,Y])
```

B ADDITIONAL INFORMATION CHAPTER 3

B.1 Exploratory analysis of the 47-protein reference set in solution with PCA in the region of the amide I band

A set of IR spectra of proteins with their respective SS annotations were analysed by (Principal Component Analysis) PCA in order to visualize groups based on SS content and correlate the region of the spectrum where the Amide I is to structures: helix, sheet, turns, bends and irregular (random coil).

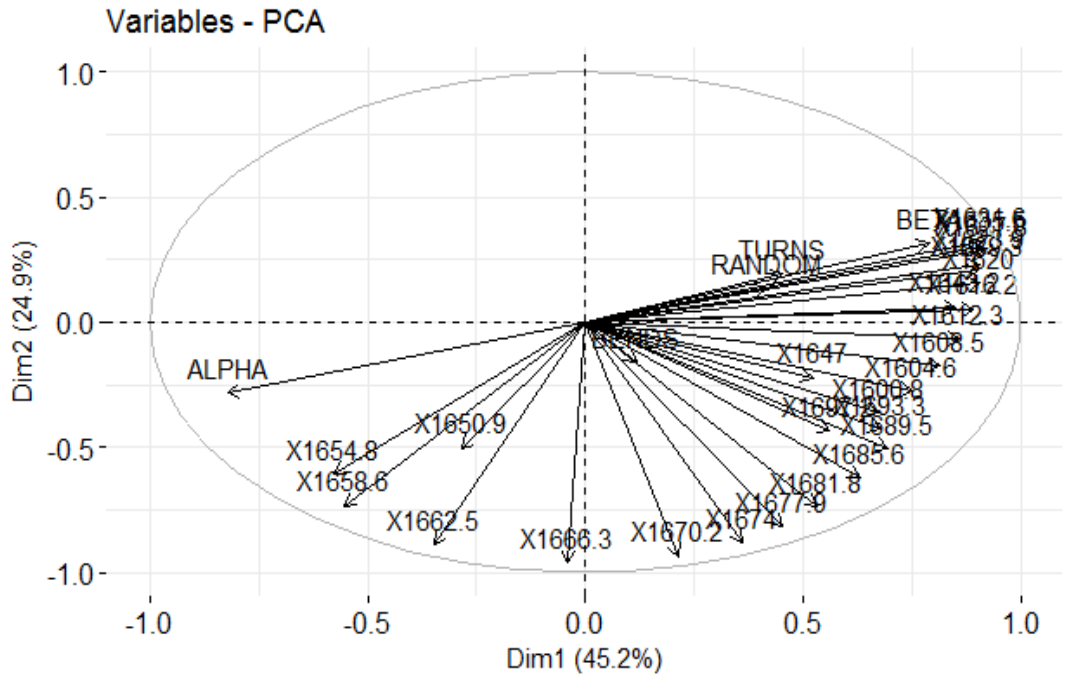

Figure 0.1. PCA 1 vs PCA2 loadings plot.

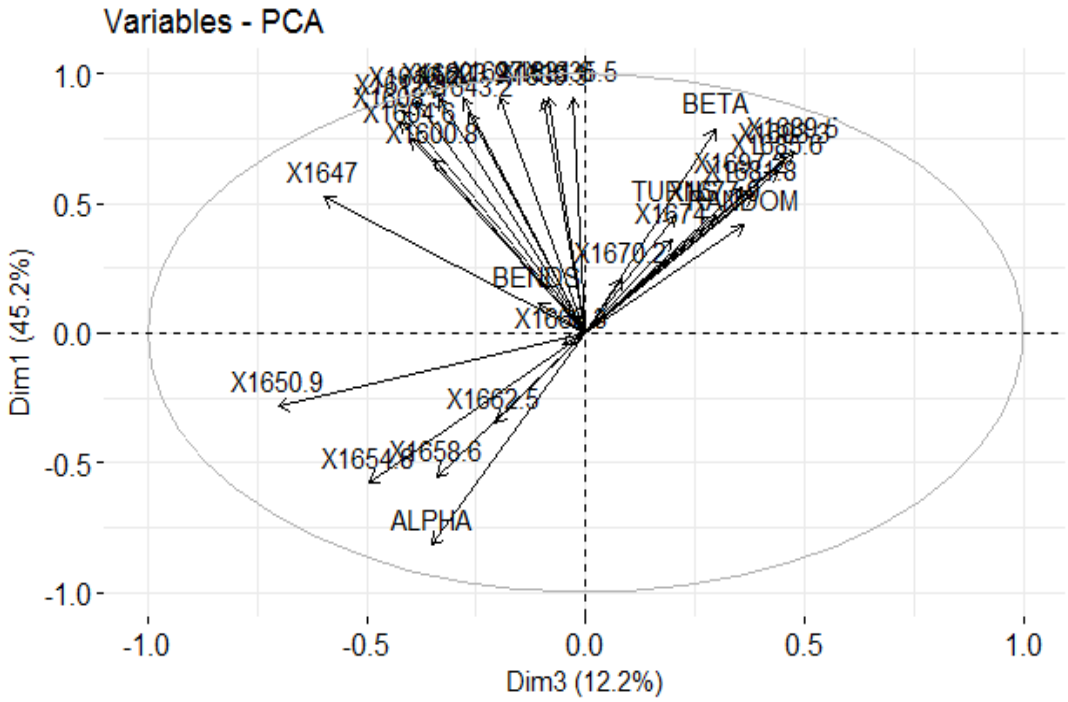

Figure 0.2. PCA 1 vs PCA 3 loadings plot.

PCA 1 and 2 explain about 70% of the overall variance. Figure 0.1 shows helix and sheet are inversely correlated. Helix is strongly associated with 1654.8 and 1658.6 cm^{-1} whereas beta structures are more correlated to 1620–1640 cm^{-1}. Figure 0.2

shows PCA 3, which increases the total explained variance to ~82%. In this figure sheet seems to be mostly explained by wavelengths between 1685 and 1700 cm^{-1} which in the literature is assigned to antiparallel sheet.

B.2 Fourier Self-Deconvolution (FSD) + band fitting

Band fitting is the technique most commonly used to estimate protein secondary structures from IR data. It involves fitting Gaussians, Lorentzians or a combination of both on the presumption that the amide I consist of many different vibrational modes that are indicative of a different secondary structure. The application of this technique usually involves a previous enhancement of resolution. In order to resolve the peaks within the amide I, the bands were FSD (explained below) deconvolved. The first issue to be considered is whether noise is apparent in the spectrum. The difference between the spectra in Figure 0.3 can be explained by the difference in the number of accumulations *n* and how they relate to the signal-to-noise ratio (*S/N)*

$$\frac{S}{N} = \sqrt{n}\frac{Sx}{Nx} \tag{0.1}$$

where *n* is the number of accumulations, S_x the mean of the signal detected and N_x the noise of the detector. Thus, the signal to noise ratio, for a certain concentration, depends on the number of scans. This means, the larger the number of scans the smaller is the spectral noise.

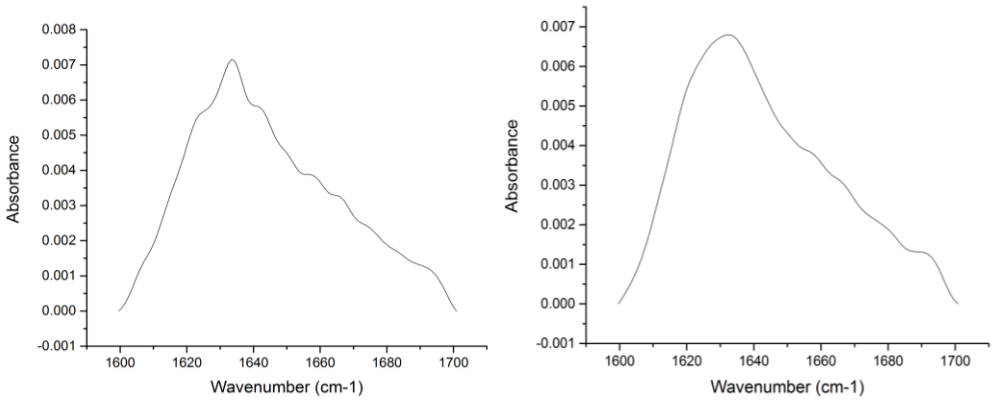

Figure 0.3. Spectrum of a 2 $mg.ml^{-1}$ concanavalin in SPB collected with 50 (left) and 600 (right) accumulations. The spectra were collected with a Specac 6-bounce ATR unit.

Fourier Self-Deconvolution is a resolving technique based on the deconvolution theorem which is as follows

$$E(v) = G(v) * E'(v)$$
$$f[E(v)] = f[G(v) * E'(v)] = cte \cdot f[G(v)] \cdot f[E'(v)]$$
$$E'(v) = f^{-1}\frac{f[E(v)]}{cte \cdot f[G(v)]}$$

(0.2)

where *E(v)* is the spectrum of interest, *G(v)* the detector function which broadens the peaks, *E'(v)* the original function with no broadening and *f* stands for Fourier transformed.

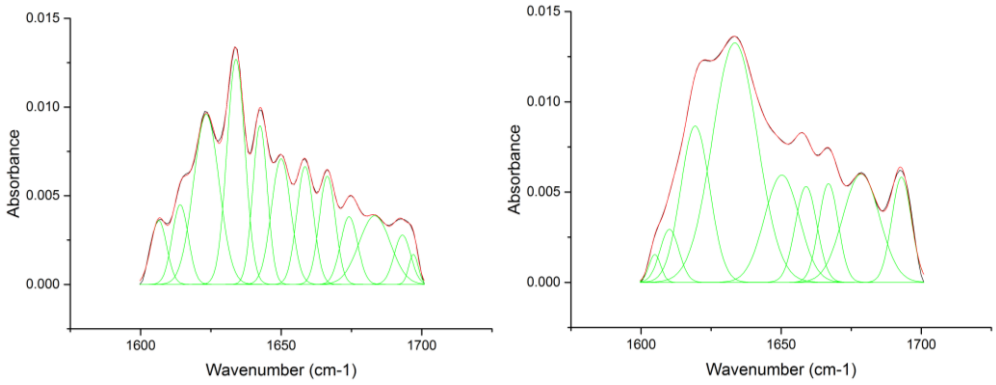

Figure 0.4. Deconvolution and band fitting of the spectra in Figure 0.3 with gamma factor 8 and 0.12 smoothing.

There are two parameters that need to be chosen so that they give the optimal resolution of the peaks. The first one is called the 'gamma factor' which we deduce is the factor *w* in the Lorentzian function (Equation 2.70) and the second a 'smoothing factor' to attenuate the effect from the noise which is prominent at this level of absorbance. The fittings showed in Figure 0.4 were performed in Origin[80] with the following procedure:

Firstly, the 'baseline' was flattened and zeroed by linear interpolation between 1600 and 1700 cm^{-1}. Secondly, the band was Fourier Self-Deconvolved. Then, the deconvolved bands were zeroed again, and the second derivatives calculated in order to find the positions of the peaks. These positions were used by the software as an initial value that was optimized throughout many iterations along with the bandwidth and the area. In some cases, it was necessary to set some boundaries

(limiting bandwidth to between 0 and 20 cm^{-1} for example, fixing positions calculated by the second derivative and setting positive values for the areas) in order to prevent unreal solutions (solutions that give an outstanding goodness of fit but that make no sense from a fundamental point of view). Assignments are shown in Table 0.1, which were made by comparing to the ones in the literature (Table 3-2). The fraction of each structure was obtained by adding up the areas of those bands that stand for a particular SS and dividing their sum by the total area.

The problem for us with the band-fitting approach to secondary structure determination is illustrated in Figures 0.4. The spectral noise in Figure 0.4 (left) produced peaks that could not be distinguished from the actual intrinsic bands. As we wished to push the instrumental limits of the technique and we cannot produce noise-free spectra, we were uncomfortable with this fitting approach.

Table 0-1. Secondary structure predictions of band fitted concanavalin.

Peak wavenumber	**Secondary structure Assignment**	**100 scans**	**600 scans**
1624 ±1.0	β-sheet	18.36476	----------
1627 ± 2.0	β-sheet	----------	----------
1633 ± 2.0	β-sheet	17.2226	35.38799
1638 ± 2.0	β-sheet		----------
1642 ± 1.0	β-sheet	9.86228	----------
1648 ± 2.0	Random	10.37308	12.32993
1656 ± 2.0	α-helix	8.36185	6.8125
1663 ± 3.0	3_{10} -helix		6.33292
1667 ± 1.0	β-turn	7.43749	----------
1675 ± 1.0	β-turn	4.93433	----------
1680 ± 2.0	β-turn		13.1077
1685 ± 2.0	βturn	9.11843	----------
1691 ± 2.0	β-sheet	3.43539	6.711
1696 ± 2.0	β-sheet	1.18954	----------

B.3 Predictions with 47-protein reference set

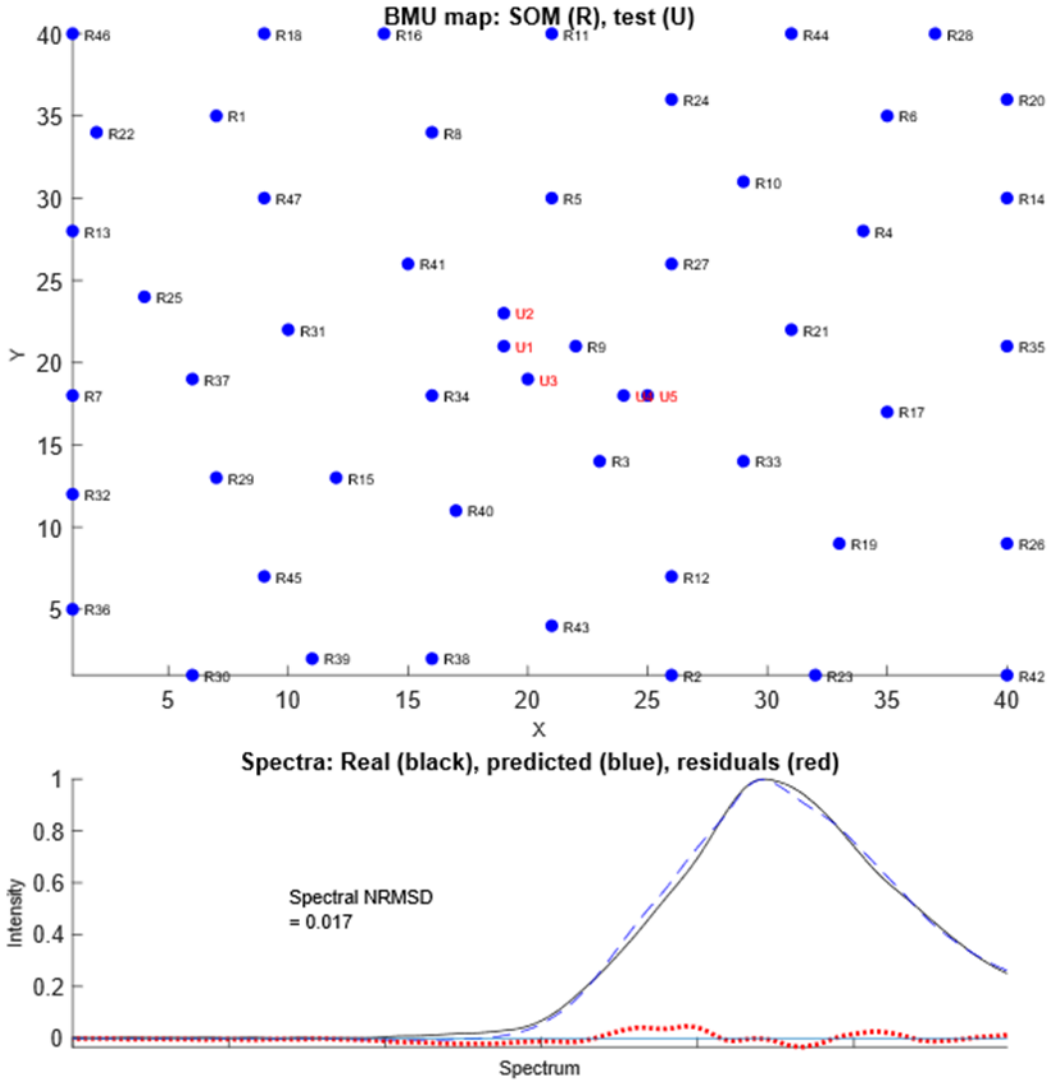

Figure 0.5. SOM prediction for the 50 mg.ml^{-1} IR-ATR spectrum of Lysozyme by means of the 47 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm-1 and normalized).

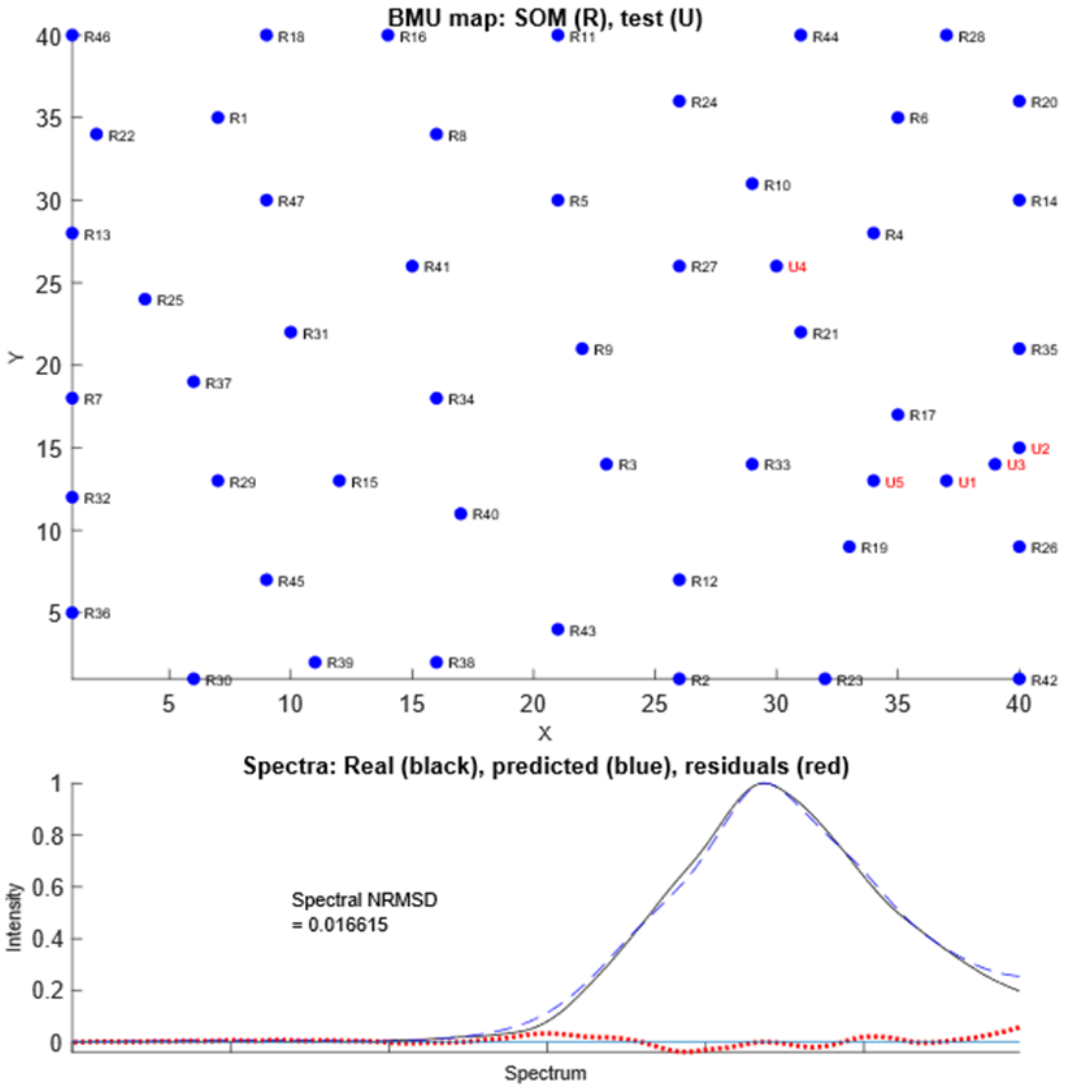

Figure 0.6. SOM prediction for the 50 mg.ml^{-1} corrected IR-ATR spectrum of Lysozyme by means of the 47 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm^{-1} and normalized).

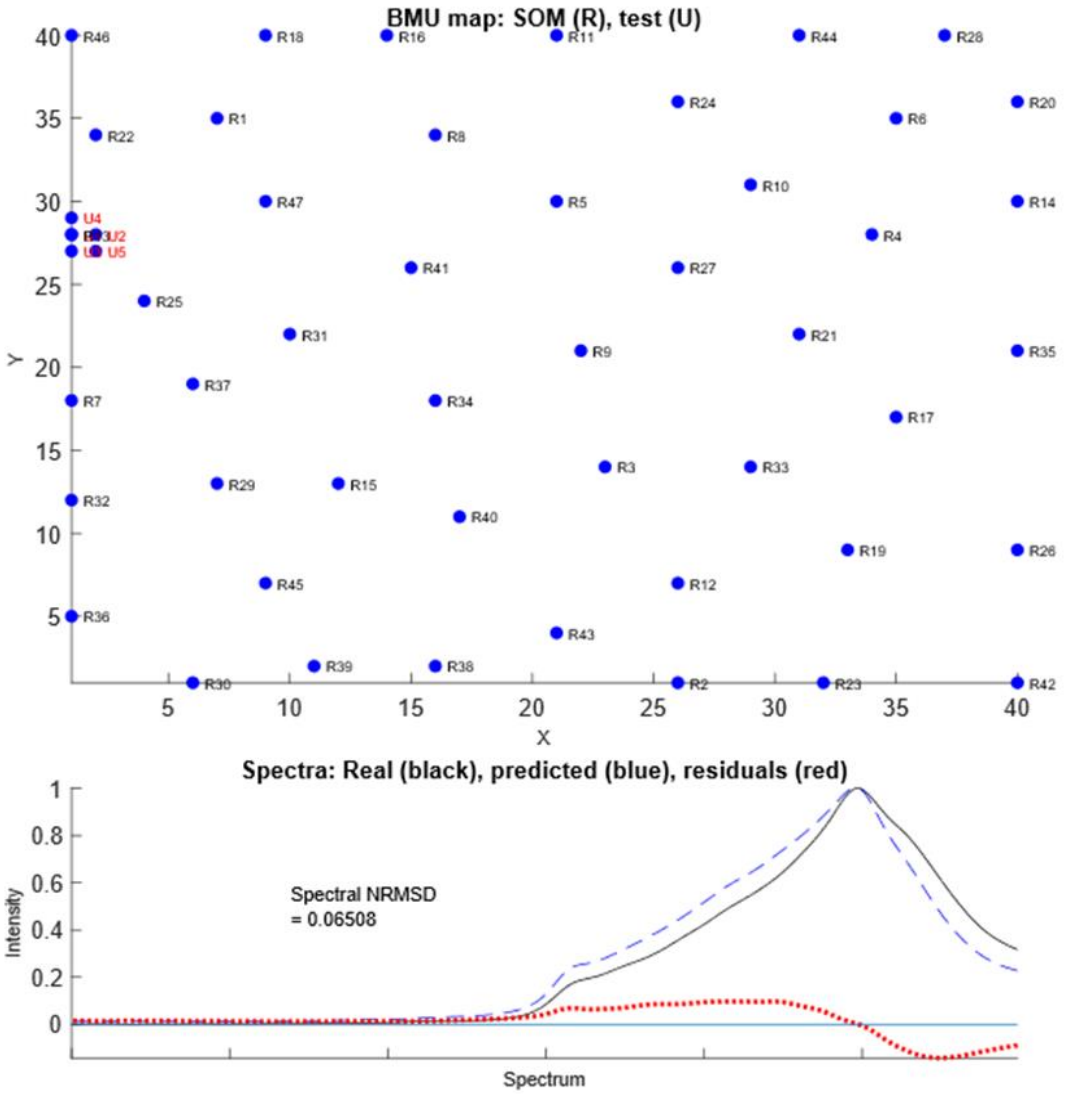

Figure 0.7. SOM prediction for the 50 mg.ml^{-1} IR-ATR spectrum of Concanavalin by means of the 47 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm-1 and normalized).

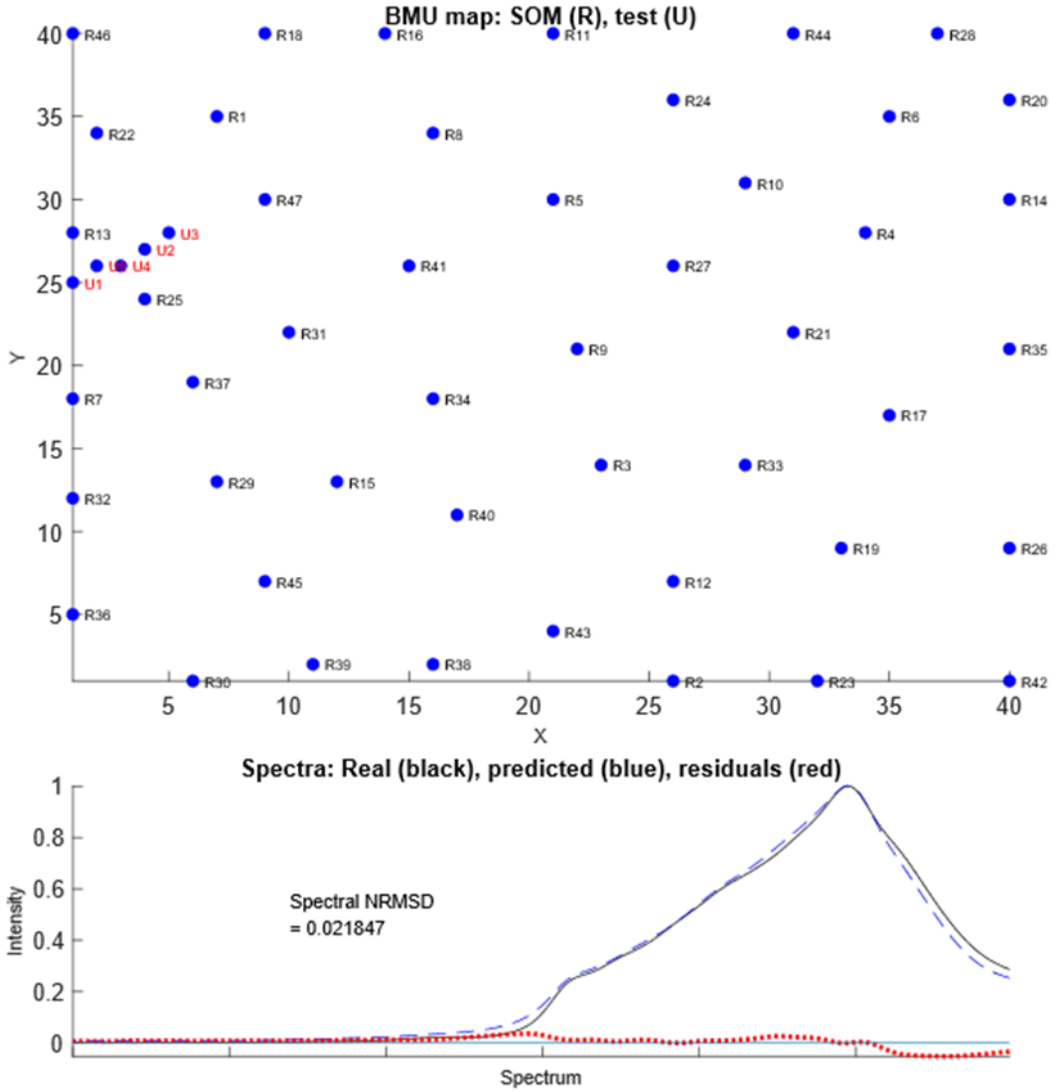

Figure 0.8. SOM prediction for the 50 mg.ml^{-1} corrected IR-ATR spectrum of Concanavalin by means of the 47 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm-1 and normalized).

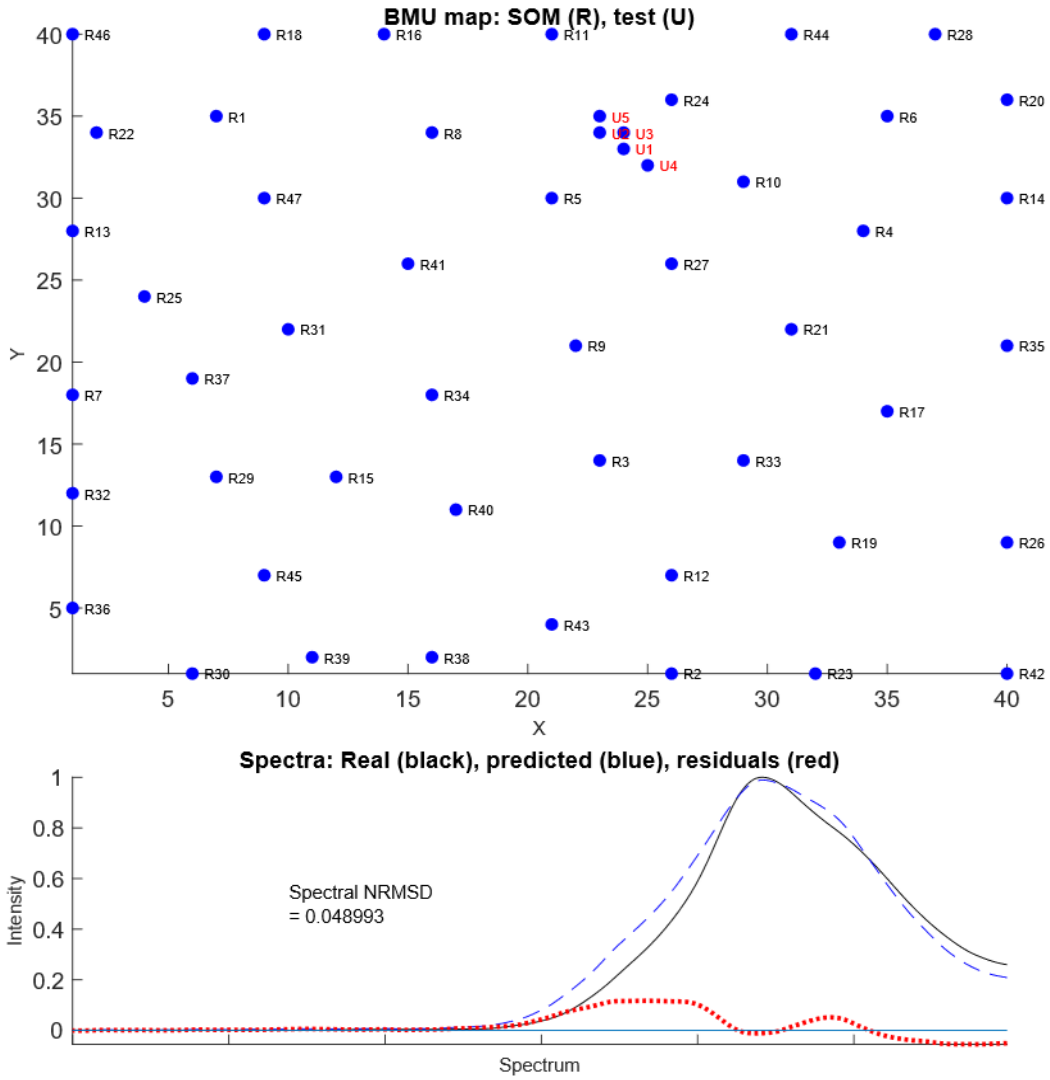

Figure 0.9. SOM prediction for the 50 mg.ml^{-1} IR-ATR spectrum of Bsa by means of the 47 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm-1 and normalized).

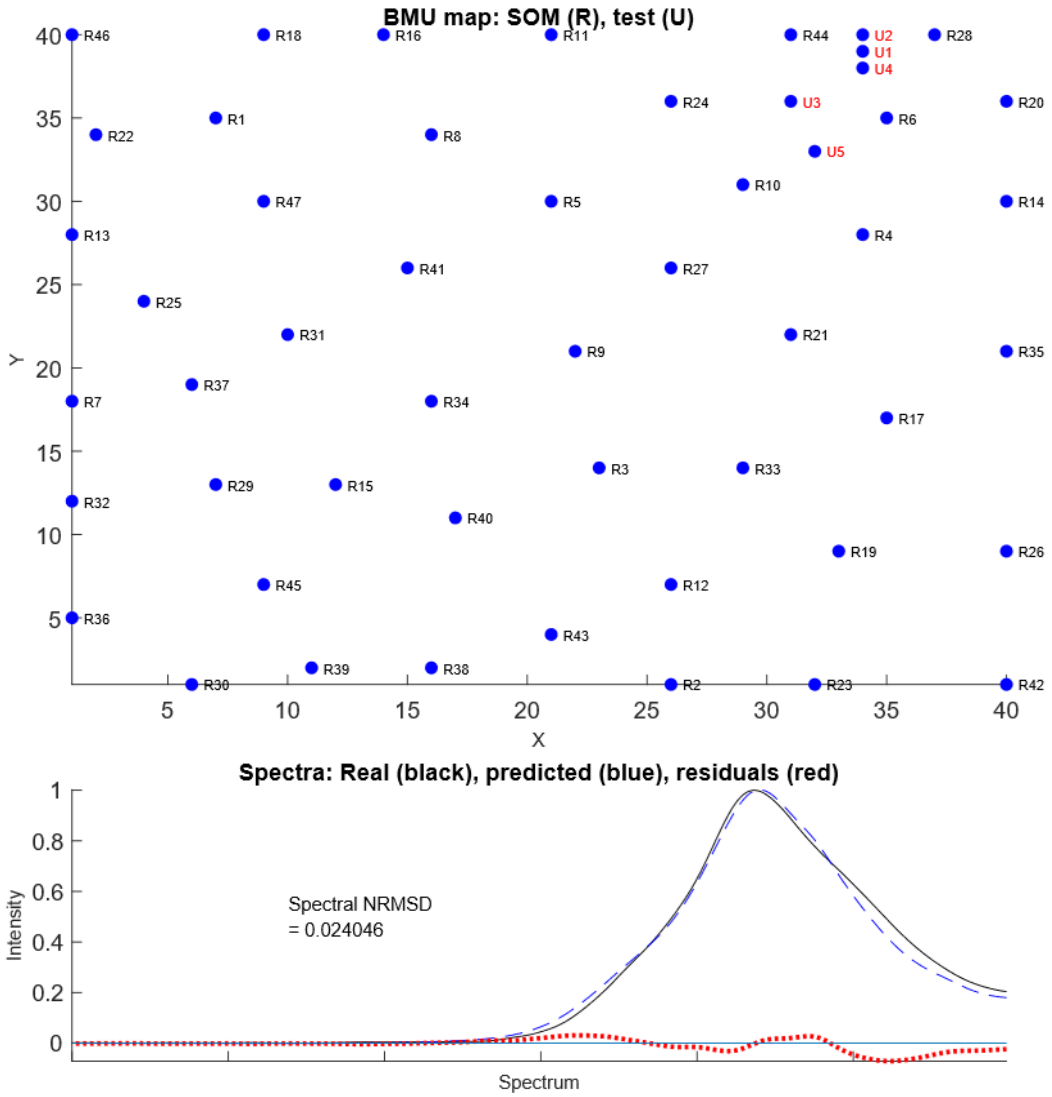

Figure 0.10. SOM prediction for the 50 mg.ml^{-1} corrected IR-ATR spectrum of Bsa by means of the 47 proteins ref set. The map was trained with 40x40 nodes, 40000 iterations and 5 BMU and the spectrum processed like the proteins in the training set (trimmed from 1600 to 1800 cm-1 and normalized).

C ADDITIONAL INFORMATION CHAPTER 4

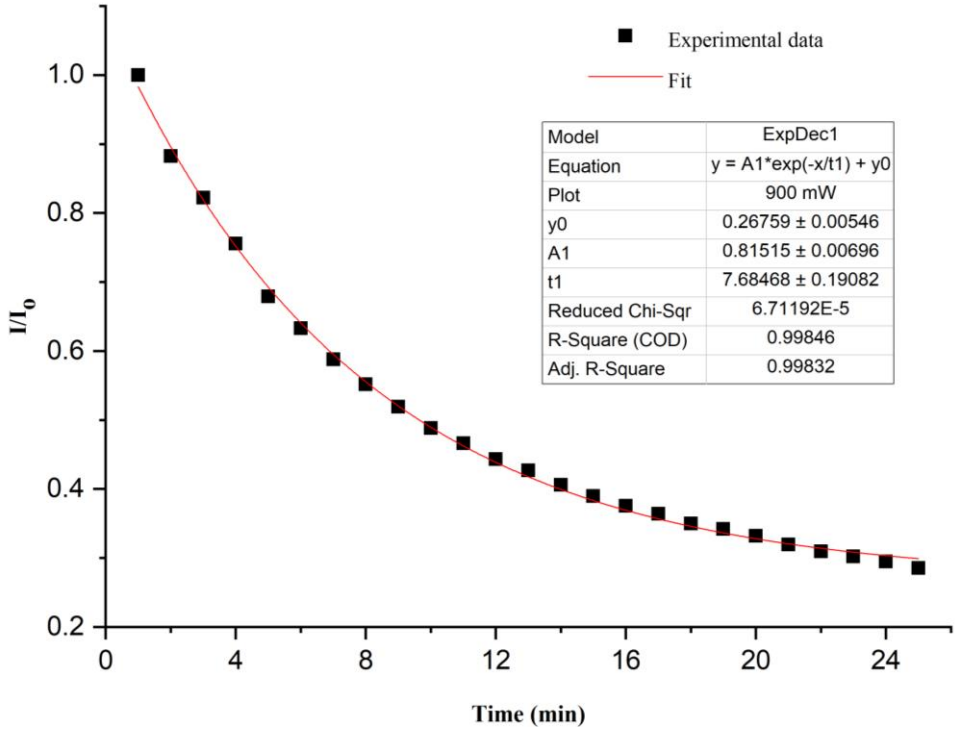

Figure 0.11. Fitting curve plots of 900 mW photobleaching curve with a one-exponential model. The algorithm used for the fitting was Levenberg Marquardt.

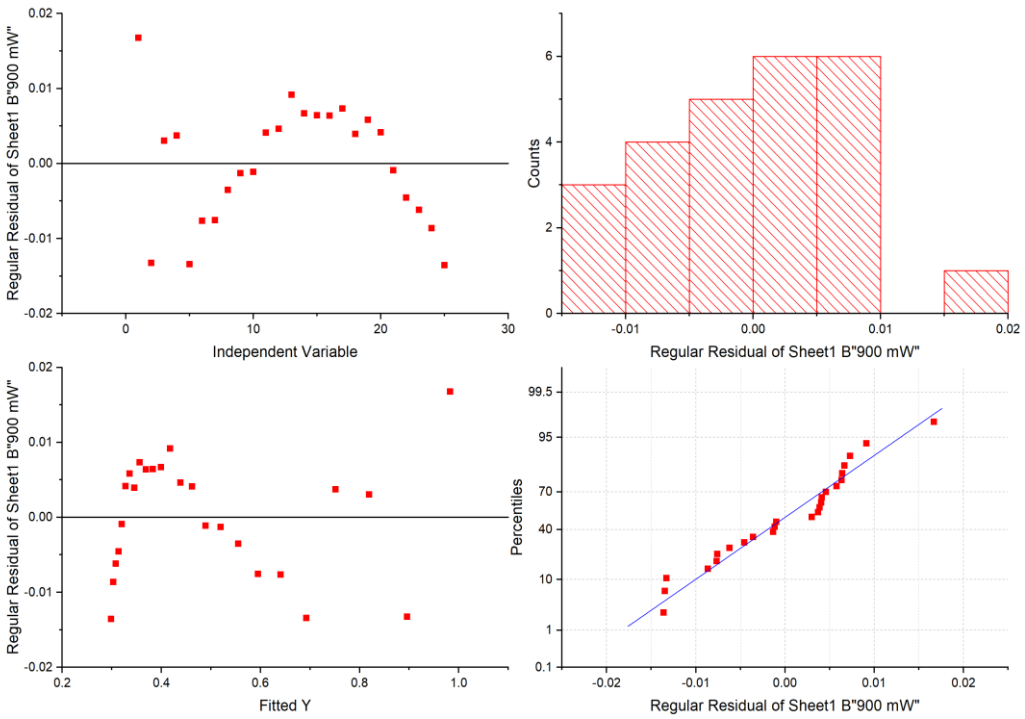

Figure 0.12. Residuals plots corresponding to Figure 0.11

D SUPPLEMENTARY INFORMATION ON PROTEIN REFERENCE SETS AND LIST OF AMINO ACIDS

Table 0-2. Proteins integrating the reference sets of the different techniques and aggregation state.

Protein	IR aq	IR solid	Raman aq	Raman solid	ROA aq
Albumin from chicken egg	X		X	X	X
Alcohol dehydrogenase				X	
Aldolase from rabbit tail	X		X		X
Actin from bovine muscle		X		X	
α-chymotrypsinogen from bovine pancreas	X	X	X	X	X
α-lactalbumin	X	X	X	X	X
Aprotinin	X	X	X	X	
Avidin					
α-bungarotoxin		X		X	
α-amylase from bacillus licheniformis		X		X	
α-crystallin from bovine eye				X	
β-lactoglobulin from bovine	X	X	X	X	X
β-glucuronidase				X	
Bovine serum albumin	X	X	X	X	X
Bovine fibrinogen	X		X	X	X
Carbonic anhydrase isozyme II from bovine erythrocytes	X	X	X	X	X
Carboxypeptidase				X	
Concanavalin A from canavalia ensiformis	X	X	X	X	X
Catalase from bovine liver		X			
Cytochrome c from bovine heart		X			
D-amino acid oxidase		X			

	Deoxyribonuclease I from bovine pancreas	**X**	**X**	**X**	**X**	
	Glyceraldehyde-3-phosphate dehydrogenase from rabbit muscle	**X**	**X**	**X**	**X**	
	Haemoglobin from bovine*, human**	**X**	**X***			
	Human insulin		**X**		**X**	
	Human serum albumin		**X**		**X**	
	Human apo transferrin	**X**	**X**		**X**	**X**
	Hexokinase from saccharomyces	**X**	**X**	**X**	**X**	
	Jacalin		**X**		**X**	
	Lectin from phaseolus vulgaris		**X**		**X**	
	Lysozyme from chicken	**X**	**X**	**X**	**X**	**X**
	Myoglobin		**X**			
	Papain from papaya latex	**X**	**X**	**X**	**X**	
	Peroxidase from horseradish	**X**	**X**			
	Phosphatase alkaline from bovine intestine		**X**		**X**	
	Ribonuclease A from bovine	**X**	**X**		**X**	**X**
	Superoxide dismutase from bovine erythrocytes		**X**		**X**	
	Thermolysin from geobacillus				**X**	
	Trypsin from bovine pancreas	**X**	**X**	**X**	**X**	**X**
	Trypsin inhibitor from glycine	**X**	**X**	**X**	**X**	**X**

Table 0-3. List of amino acids from CRC Handbook of Chemistry 2010.