Computational Studies of the Dynamics and Spectroscopy of Peptides.

Rachel E. Hill, MSci.

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

July, 2016

Abstract

Proteins play a crucial role in almost all biological processes. Developing a complete understanding of the link between their structure, dynamics and function is goal of many areas of scientific research. One tool with which the protein structure has been investigated is infrared (IR) spectroscopy. IR is a useful probe of protein structure because the amide I region (1600-1700 cm^{-1}) is sensitive to the secondary structure elements such as α helices and β-sheets; different secondary structures give rise to different signatures in the IR spectrum. The drawback of traditional IR is that the amide I region is often broad and featureless and thus difficult to interpret. Two-dimensional infrared spectroscopy (2DIR) can improve on the structural sensitivity of IR by spreading the transitions over a second frequency domain resulting in off-diagonal peaks that quantify the coupling between molecular vibrations. The development of the technique has been greatly aided by computational calculations of 2DIR spectra, such as from molecular dynamics (MD) simulations. In this thesis we apply the exciton method to calculate IR and 2DIR of Leu-enkephalin, a pentapeptide that is involved in the mediation of pain in the body by binding to opioid receptors. The calculated IR show qualitative agreement with both previously calculated and experimental spectra. Previous calculations gave results only for a single structure in the gas phase, and we have expanded this work to including spectra from MD simulations. Our work contributes calculated spectra to aid further understanding of the dynamics of Leu-enkephalin, which may help in the search for more effective opioid analgesics. In addition to enkephalin, we investigated four variants of the

enoyl-acyl carrier protein reductase enzyme InhA found in *Mycobacterium Tuberculosis*, the bacterium responsible for tuberculosis, which is a threat to global health. In particular, we investigate both wild type variants and mutant known to exhibit resistance to isoniazid, one of the front-line treatments for tuberculosis. Due to the size of the protein, it is currently not a suitable candidate for theoretical 2DIR calculations. Instead we used data extracted from the one exciton Hamiltonian to probe the structural dynamics of the different variants. Our work supports previous experimental results, in particular work that had suggested the importance of a 20 residue binding loop in mediating isoniazid resistance, and suggest several residues as potential candidates for isotope-labelled 2DIR experiments. The work in this thesis provides a starting point for further investigation of the dynamics and calculated spectroscopy of both Leu-enkephalin and InhA.

For my mum, who told me I could be whatever I wanted when I grew up as long as I tried hard enough.

Acknowledgements

Thanks must go first and foremost to my supervisor, Professor Jonathan Hirst, without whom this would not have been possible. I am immensely grateful for his seemingly endless patience and excellent advice and encouragement. Thanks also to Doctor David Robinson for his advice and support, particularly with the simulation side of things. I also wish to thank Professor Tony Parker and Doctor Neil Hunt, particularly for their helpful discussion with regards to 2DIR and its practical workings and applications. My gratitude also goes to my two examiners: Professor Ian Williams and Doctor Nicholas Besley.

Thanks also to the University of Nottingham and the Science and Technology Facilities Council for funding; the School of Chemistry, the University of Nottingham High Performance Computing service and Midplus for resources.

My gratitude goes to my colleagues in A47 and 48, past and present for discussions, distractions and occasional trips to the pub. I am especially thankful to John, Fouad and Pritesh for never seeming to mind me coming to them with complaints about my computer.

Thanks must also go to my GP, Doctor Tim Baker, and his colleagues at Cripps Health Centre. Without them to help me manage my mental health I

most certainly wouldn't have gotten to this stage.

To the late Doctor James Bullock, whom I knew as Jamibu. Thanks for being my inspiration when things were at their toughest. I wish you could be here.

Last, but certainly not least, I also wish to thank my family and friends. To Sam, who has borne the brunt of my frustration and been endlessly patient with the long working hours throughout "the PhD years". Thank you for marrying me in the middle of one of the hardest things I will ever do. Thanks also to my Mum, to whom this thesis is dedicated, Dad, Gran, and my brothers Andrew and Christopher. You have helped provide advice, encouragement, distractions and acted as sounding boards for all the complaints that couldn't go to Sam or Jonathan. Sarah, Susan, Jay, Vex, John, Pinaz, Mick, Rebecca, James T, Emily, James B, Mel and everyone else I don't have space to mention; thank you for helping to keep me (mostly) sane.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviations	
2DIR	Two-dimensional Infrared
2DNMR	Two-dimensional Nuclear Magnetic Resonance
ABNR	Adopted Basis Newton Raphson
ACP	Acyl-carrier Protein
BPTI	Bovine Pancreatic Trypsin Inhibitor
CD	Circular Dichroism
DNA	Deoxyribonucleic Acid
FRET	Fluorescence Resonance Energy Transfer
FTIR	Fourier Transform Infrared
GBSA	Generalized Born/Surface Area
GPU	Graphical Processing Unit
HBC	Hydrogen Bond-Specific Charges
HEWL	Hen Egg White Lysozyme
HIV	Human Immunodeficiency Virus
HP35	Chicken Villin Headpiece 35
IAPP	Islet Amyloid Polypeptide
IR	Infrared
MD	Molecular Dynamics
mRNA	Messenger Ribonucleic Acid
NADH	Nicotinamide Adenine Dinucleotide
NEE	Nonlinear Exciton Equations
NISE	Numerical Integration of the Schrödinger Equation
NMR	Nuclear Magnetic Resonance
NR	Newton Raphson
OLPS-AA	Optimized Potentials for Liquid Simulations-All Atom
PDB	Protein Data Bank
PME	Particle Mesh Ewald
QM	Quantum Mechanical
RMSD	Root Mean Square Deviation
SD	Steepest Descent
SPC	Simple Point Charge
SPC/E	Extended Simple Point Charge
UB	Urey-Bradley
UV	Ultraviolet
VMD	Visual Molecular Dynamics

Chapter 1

Introduction

1.1 Proteins

Proteins are biological macromolecules that play a crucial role in almost all biological processes. Within living organisms proteins are responsible for a wide variety of functions and fall into a number of important classes. Proteins in the human body are involved in catalysis of metabolic reactions e.g. breakdown of food within the digestive system; DNA replication requires polymerase enzymes to "unzip" the DNA chain as well as ribosomes to translate the mRNA to an amino acid sequence for the generation of new proteins; structural roles such as keratin in hair and nails; cell signalling and stimulus response both as signal molecules and the receptors for signal molecules; molecular transport such as ions through cell membranes.

Proteins are long chain polymers built from amino acid residues. There are twenty naturally occurring (L) amino acids. These polypeptides do not remain as loose chains, instead fold into unique three-dimensional structures. These three-dimensional structures are of key importance for the protein's ability to carry out its biological function. Hence one key area of protein research includes structure determination via techniques such as X-ray crystallography and Nuclear Magnetic Resonance (NMR). X-ray crystallography has been indispensable in the elucidation of protein structure, providing almost 93,000

crystal structures in the Protein Data Bank.[1] It's major drawback however, is that solution phase structure is not obtainable via X-ray crystallography, requiring other tools to investigate this important aspect of protein structure and function. Other key areas of research include investigations into folding mechanisms; identification of small molecule drug targets that bind to specific protein binding sites; inter- and intra-molecular interactions, particularly in the cases of proteins interacting with DNA and with each other. Within almost all living organisms, the process for producing a protein molecule is virtually the same. Proteins are coded for by the organism's DNA, with each amino acid being coded for by a sequence of three nucleotides, known as codons. Proteins are both produced from and act on DNA.

1.1.1 Structure

Protein structure is generally divided into four levels, called the primary, secondary, tertiary and quaternary structures. These levels of structures are described below.

The primary structure of a protein—and one of the main ways in which they are differentiated from one another—is the sequence of amino acid residues they comprise. An amino acid is a small, organic molecule consisting of an amine group, a carboxylic acid and a side chain specific to each individual amino acid. The general structure is shown in figure 1.1.

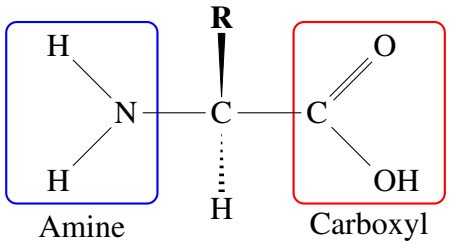


Figure 1.1: General structure of an amino acid

There are twenty naturally occurring amino acids that generally make up the primary structure of a protein. These vary in structure by the side chain **R**

and fall into three general categories; hydrophobic (e.g. alanine), charged (e.g. lysine) and polar (e.g.serine). Glycine (Gly), is usually considered by itself owing to the fact its side chain consists simply of a single hydrogen. The exact nature of side chain **R**—structure, pK_a, hydrophobicity etc.—has an enormous impact on the way the peptide chain folds into the three-dimensional protein.

The individual amino acids are joined together in a long chain via a peptide bond, which occurs between the amine terminus of one amino acid and the carboxyl terminus of the other. The bond is formed via the loss of one water molecule. The peptide bond is planar and relatively rigid, owing to the delocalisation of electrons between the CO and NH. The rigidity of the peptide bond has consequences for the secondary structure of proteins.

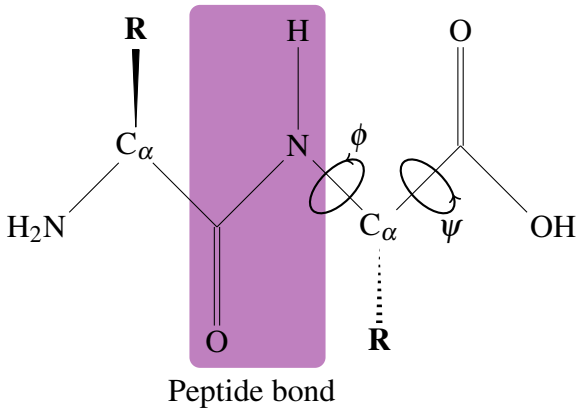


Figure 1.2: Dipeptide showing peptide bond (highlighted in purple), phi (ϕ) and psi (ψ) dihedral angles.

The next level of protein structure is the secondary structure, which encompasses the 3D structure that the amino acid chain forms locally. Protein secondary structure elements are generally held together via hydrogen bonds interactions. The most common types are α helices and β sheets but other secondary structure elements do exist. Random coils refer to residues that do not adopt α-helical nor β-sheet or other definable conformations. Random coils are important as they commonly form links between the other types of secondary structure elements.

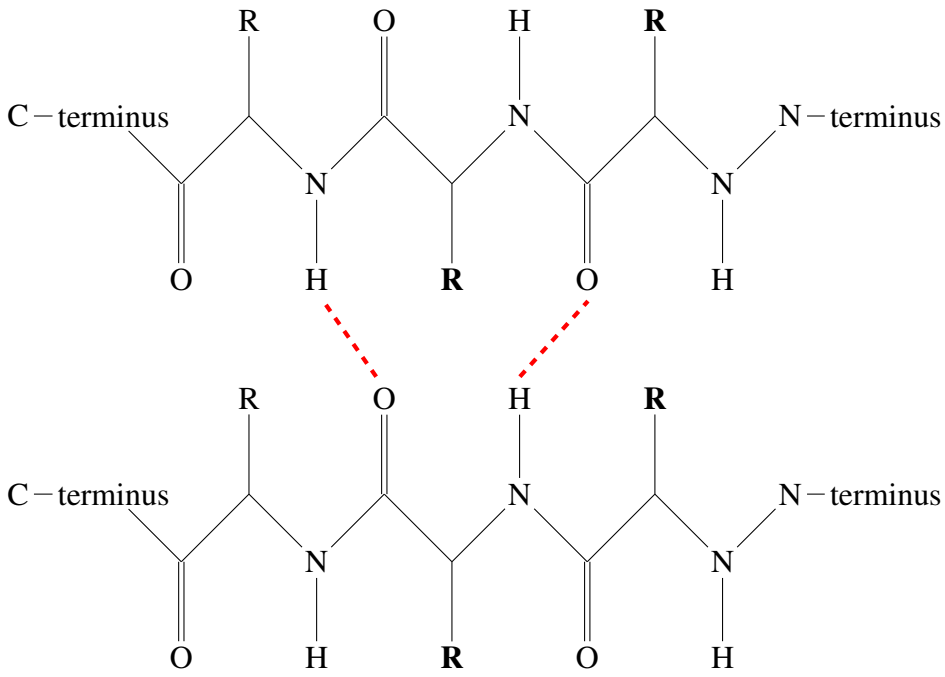
The polypeptide backbone can be described completely by the dihedral

angles ϕ and ψ, shown in figure 1.2. ϕ describes the angle of rotation about the $C_\alpha-N$ bond and ψ the rotation about $C_\alpha-C$. Different combinations of these angles give rise to different secondary structure elements. α helices are defined by dihedral angles of around (-60°,-45°). Generally α-helices adopt dihedral angles such that the sum of the ψ angle of one residue and the ϕ angle of the next residue is approximately 105°.

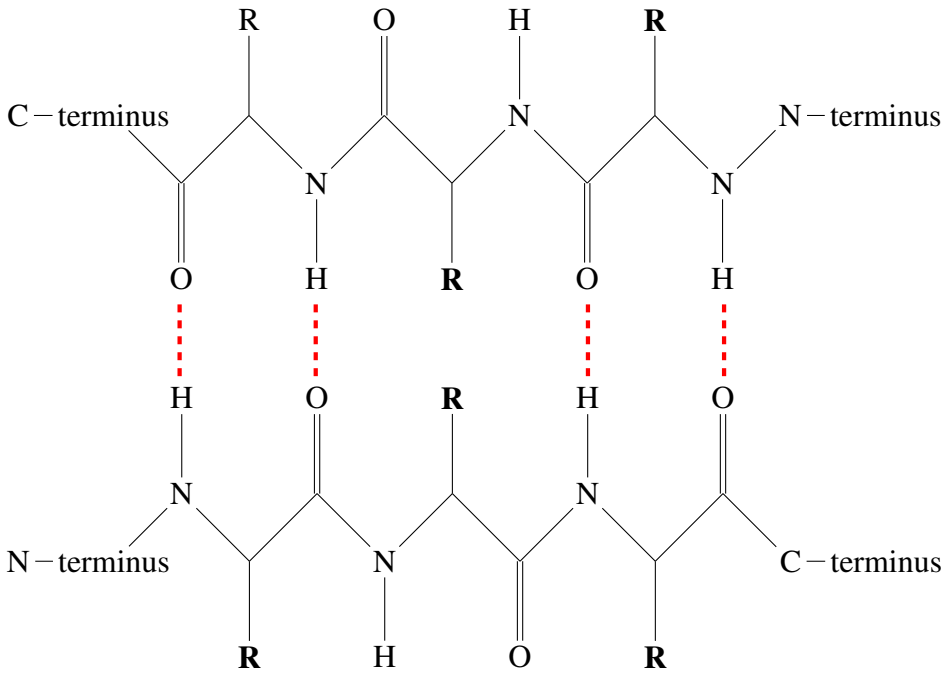
Also of great importance to this secondary structure element is the hydrogen bonding pattern. In an α-helix the N−H of one residue is hydrogen bonded to the C=O of the residue four residues previous, giving a hydrogen bonding pattern of $(i, i+4)$. The combination of the dihedral angles and the hydrogen bonding pattern give rise to a helix with four residues per full 360° turn, meaning that the minimum length for this type of helix is four residues. There also exists a tighter helix with an $(i, i+3)$ hydrogen bonding pattern known as the 3_{10} helix and a looser one with an $(i, i+5)$ hydrogen bonding arrangement, called the π-helix.

The other secondary structure element commonly found in proteins is the β-sheet, consisting of two or more β-strands held together by inter-strand hydrogen bonds. The sheet is defined by dihedral angles of around (-135°,135°), though the actual values can vary greatly. β-sheets can be described either as parallel or anti-parallel; in a parallel sheet the individual strands are aligned N-terminus to N-terminus, whereas in an anti-parallel sheet the strands are aligned N-terminus to C-terminus. Both types of β-sheet are shown in figure 1.3. Unlike in the α-helix, the hydrogen bonds present in a β-sheet do not have to be between local or consecutive residues; in fact β-sheets are commonly formed from strands made up of residues throughout the amino acid chain.

Different types of amino acids have a tendency to form different secondary structure elements. For example methionine, alanine, leucine and glutamate and lysine are commonly found in α-helices; large aromatic residues—such as tyrosine, phenylalanine and tryptophan—and branched amino acids—like



(a) Parallel



(b) Anti-parallel

Figure 1.3: Hydrogen bonding in parallel and anti-parallel β-sheets. Hydrogen bonds shown in red.

threonine, valine and isoleucine—often occur in β-sheets. Proline has a tendency to break secondary structure, especially α-helices due to its inability to form hydrogen bonds and because of steric hindrance due to its cyclised side chain. Proline will also break β-sheets as the allowed dihedral angles for proline lie outside of those generally found in β-sheets.

One useful method for examining a protein's secondary structure is a Ramachandran plot,[2] in which the phi and psi dihedral angles for each residue are plotted against each other. By doing this for a large number of proteins, Ramachandran found that certain combinations of dihedral angles are conformationally favoured. Other regions of the Ramachandran plot are only sparsely populated, due to the steric interactions preventing those particular combinations from occurring. The favoured regions correspond to the major secondary structure elements in a protein. For example, in figure 1.4, region A corresponds to the α-helix, region B to the β-sheet and region C to the comparatively rare left-handed α-helix. Most amino acids will stay within these regions, with the notable exceptions being glycine and proline. Since glycine has only a hydrogen side chain, it has much more conformational freedom than other residues, and as such can adopt phi-psi angle combinations not available to other residues. Proline, on the other hand, being a secondary amine, is much more conformationally restricted than other amino acids, which is why it has a tendency to break secondary structure, with the exception of the polyproline helices.

Ramachandran's original plot[2] modelled the atoms of the amino acids as hard spheres defined by their van der Waals radii. For this plot, the allowed and favoured regions were restricted only to conformations where the electron clouds did not overlap. More recent modelling calculations have been done to update the work of Ramachandran and provide newer, more accurate favourable and allowed regions. One example of this work is by Richardson,[3] which was used to generate figure 1.4.

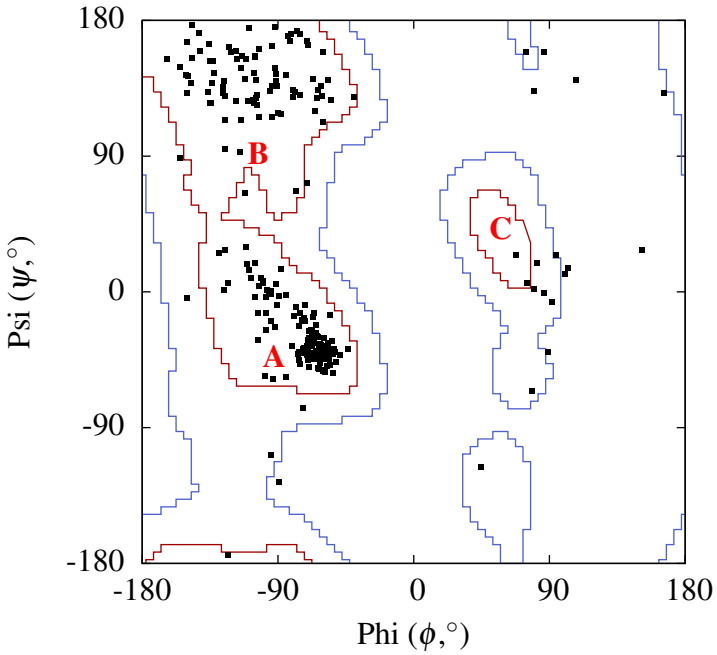


Figure 1.4: Ramachandran plot for a variant of the Enoyl-Acyl Carrier Protein Reductase in *Mycobacterium Tuberculosis* (PDB code: 1zid) after minimization. Favoured regions (red) and allowed regions (blue) taken from Richardson *et al.* 2003.[3] A= α-helix region; B= β-sheet region; C=left handed helix region.

The third level of protein structure is the tertiary structure, which is the overall three-dimensional structure of the protein and is vital to the function of the protein. The dominant factor in determining the native structure of a folded protein is the thermodynamic stability of the structure. It is assumed that the native state will be the lowest energy structure. There are a number of different ways in which the amino acids—particularly the side chains—may interact with one another, and these interactions play a role in determining and maintaining the tertiary structure. Hydrogen bonding, important in maintaining the stability of secondary structure elements, also plays a role in stabilising the tertiary structure. Also important are disulfide bonds between sulfur containing cysteine residues, which can form covalent linkages between two portions of the protein. The disulfide bonds bias the protein towards its native state and can aid in protein folding. Hydrophobic interactions also play an important role in protein folding and formation of the tertiary structures. Residues with hydrophobic side-chains will tend to cluster together near the centre of the protein especially

in aqueous environments, and this contributes to maintenance the tertiary structure. The role of hydrophobic side chains in protein folding is discussed further in section 1.1.2.

There are three broad groups of protein tertiary structure, and all are related to the protein's function. Globular proteins tend to be spherical in shape and are somewhat soluble with their hydrophobic residues in the centre. Water solubility is important for these proteins as they tend to be enzymes or transporters. Fibrous proteins tend to form long filaments which are largely water insoluble. These proteins are usually found in structural roles rather than regulatory, such as keratin in human hair. The third class of structure is the membrane proteins, which are usually responsible for transport in and out of the cell. In the case of these proteins the hydrophobic residues will usually settle within the membrane itself, and the hydrophilic residues either in the cytoplasm of the cell or in the external aqueous medium. Membrane proteins will tend to have a large concentrations of one particular secondary structure element, which sit within the membrane. Depending on the protein, they may have one or more transmembrane domains. Protein structure is very closely related to function and misfolding of the tertiary structure can result in a number of diseases, such as Alzheimer's or Parkinson's disease.[4]

There is a fourth level of protein structure, the quaternary structure, which occurs when a protein comprises multiple subunits. The quaternary structure then describes the three-dimensional shape and arrangement of the subunits, and is stabilised in much the same way as the tertiary structure. One example of a protein with quaternary structure is haemoglobin, which is made up of four haem containing subunits; haemoglobin is a heterotetramer, comprising two α subunits and two β sub units. Another example is HIV protease, which exists as a homodimer of two identical protein subunits.

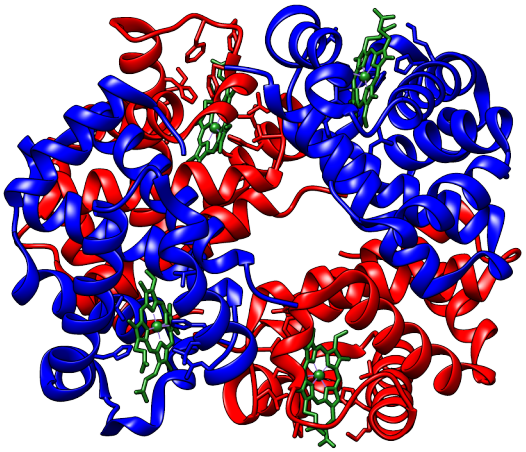


Figure 1.5: Quaternary structure of haemoglobin with α subunits shown in red, β subunits shown in blue and haem units shown in green.

1.1.2 Protein Folding

The tertiary structure of more than 100,000 proteins[1] are now known and well characterised. However, the process by which an amino acid chain folds into its native conformation is currently poorly understood. It is a widely held hypothesis[5] that the native state of a protein is the conformation that has the lowest energy i.e. is the most thermodynamically stable. This suggests that the only determinant of three-dimensional shape is the amino acid chain itself. This is evidenced by the spontaneous reforming of tertiary structure after a protein has been denatured. While the primary structure of a protein can be determined from gene sequencing and the tertiary structures of many proteins have been determined via X-ray crystallography or NMR, the exact mechanism by which the primary structure folds into the tertiary structure is currently unknown. Random sampling of all possible conformations would take an astronomical amount of time—even for sampling on a picosecond time scale—and protein folding takes place in seconds or less.[5]

Determining the detailed mechanism(s) of protein folding is therefore at the forefront of protein research, protein folding having become a field of research in and of itself.[6] Solving "the protein folding problem" would essentially allow prediction of native state conformation from the gene that

encodes the protein. There are a number of theories as to how protein folding takes place but it is now generally accepted that the folding process involves funnel-shaped energy landscapes; random thermal motions lead to conformational changes that move the protein energetically downhill towards the native structure.[4,6] This view of protein folding means that there are a large number of high-energy, unfolded structures and only a few low-energy folded structures and that the dynamics of both unfolded and natively folded proteins are of critical importance. A number of factors appear to contribute to driving the protein towards its folded state,[6] including hydrogen bonds, van der Waals and electrostatic interactions, backbone angle preferences and hydrophobic interactions where hydrophobic residues forms a nucleus around which the rest of the secondary structure can develop. This "hydrophobic effect" is a non-specific attraction between the (non-polar) hydrophobic residues. It differs from other "elementary" forces (such as van der Waals) in that it is completely dependant on environment.

Also considered important are the formation of disulfide bridges, and it is possible that early formation of the secondary structure elements may drive the protein towards its native confirmation. Other proteins reach their native state with the assistance of chaperone proteins, but this is mostly a catalytic process and all the information necessary to deduce the final form lies within the amino acid sequence. Uncovering the dynamical processes that lead to the formation of protein tertiary structure could also help to unlock potential treatments for a number of diseases which arise as a result of aggregation of proteins that have misfolded such as prion diseases like Creutzfeldt-Jakob disease. Determining at which step of the pathway the misfold occurs could lead to new treatments for these diseases, and potentially even allow for the development of preventatives.

1.1.3 Protein-Ligand Interactions

Protein-ligand interactions are of fundamental importance in cellular metabolism since it is via these interactions that proteins carry out their myriad functions. Detailed knowledge of how these interactions work on a micro-and macroscopic level is therefore required. Some examples of protein-ligand interactions include: antigen-antibody binding; enzyme-substrate interactions; ligand binding to structural proteins; protein-DNA binding; protein-saccharide, protein-protein, and protein-peptide interactions.

In biological systems the term 'ligand' can have many different meanings—in its broadest sense a ligand is simply any molecule which interacts with any given molecule, which in this thesis will be taken to mean a protein molecule. This vast range of size and type of ligand makes it difficult to understand and draw conclusions about their behaviour and biophysical properties, including the nature of their interactions with proteins.

In the context of protein function, a ligand is simply a molecule that is capable of a (usually reversible) non-covalent interaction with the protein which modulates the protein's biological role in some controllable manner. The non-covalent nature of these interactions means that they have some similarities with the interactions that hold hold together the secondary secondary structure of a protein, namely; hydrogen-bonding, van der Waals and charge interactions in particular are important for ligand binding, as is complementarity with the active or binding site. These interactions are what provides the specificity of a particular ligand for a given protein.

One of the earliest models of protein-ligand binding is the lock and key model, proposed by Fischer in the 1890s.[7] In this model, the enzyme has an active site which is like the keyhole of a lock, and the ligand is the key that fits in this keyhole. The lock and key model was proposed before it was known that proteins are dynamic entities and does not take into account the inherent flexibility of the folded protein.

A more up to date model is the induced fit model[8] which takes into account this flexibility and the conformational changes that occur upon ligand binding. It is analogous to the lock and key model, but instead of the lock and key fitting together perfectly, both change conformation slightly to improve the fit.

1.1.4 Dynamics and Function

Rather than being static entities, proteins are capable of adopting numerous related conformations that are dependent on the protein's environment. Protein flexibility and dynamics play important roles in carrying out their functions.[9–11] Instead of a single static conformation, proteins will populate ensembles of often quite similar structures and transitions between these states occur on a variety of time scales, from the picosecond to the millisecond. It seems, therefore, that determining the tertiary or quaternary structure of a protein is not an end in and of itself, but a stepping stone to full understanding of protein behaviour. Figure 1.6 shows an ensemble of related structures from a vacuum MD simulation of Leu-enkephalin.

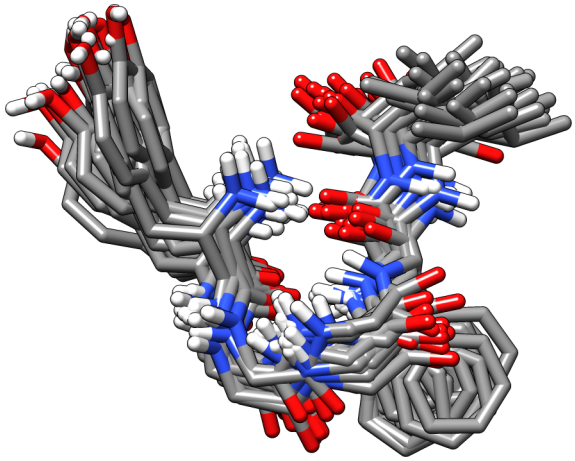


Figure 1.6: An ensemble of 20 closely related structures for Leu-enkephalin in vacuum, taken from an MD simulation.

Instead of there being a single static structure at the end of the protein folding pathway, the protein folding pathway is a dynamic equilibrium between a number of related conformations, and the interchange between these

structures can occur on a number of time scales. The kinetics and thermodynamics of the system will ensure that the majority of molecules will inhabit a conformation close to the global minimum, but exchange between these related conformers will occur at biological temperatures.

It is now generally accepted that primary structure determines tertiary structure, which in turn determines dynamics and hence function, but as yet no direct measurement has been recorded that establishes the exact relationship between dynamics, structure and function. This is due, in part, to the fact that protein behaviour occurs across numerous time-scales, from the picosecond to the second: changes in secondary structure happen over nanoseconds to milliseconds; the movement of a large side chain occurs on the order of tens of picoseconds and rapid changes in local structure and environment as a result solvent interactions can happen in just a few picoseconds. It is difficult to carry out a single experiment that probes all these time-scales simultaneously. However, much work has been conducted to flesh out the picture of protein dynamics on a picosecond time scale. A range of tools is therefore required in order to investigate the full range of protein motions on various time-scales. In the picosecond range, Two-dimensional Infrared Spectroscopy (2DIR) may be used to probe the behaviour of proteins in solution. Experimental 2DIR measurements of proteins may be complemented by calculated spectra derived from molecular dynamics simulations. A combination of the two may allow the nature of the relationship between structure, dynamics and function on a picosecond to nanosecond time-scale to investigated in greater detail.

1.1.5 Molecular Dynamics Studies of Proteins

Experimental techniques for elucidating information on protein structure, such as X-ray crystallography and NMR, are well established. However, useful as these techniques are, they are not without their drawbacks. Many proteins are difficult to crystallise, meaning the structure cannot be found by

crystallography, and while NMR spectra are easier to measure, the information it generates is much more open to interpretation. Both techniques offer only snapshots of information on the system and lack the ability to follow the dynamics in real time.

Molecular dynamics (MD) simulations, based on propagation of Newton's equation of motion, provide information about a protein at the atomic level, and on time scales not currently accessible by experiment. It is possible to follow the dynamics of a protein's function—be it as a channel protein or a catalytic enzyme—essentially in real time. The main limitation of MD is that the accuracy is greatly dependent on the quality of the force field used to describe the system. Despite this limitation, MD simulations have been used extensively to investigate proteins, and are an indispensable complement to experimental techniques.

The first MD simulation of a protein system was carried out in 1976 on the bovine pancreatic trypsin inhibitor (BPTI), a 58 residue peptide that inhibits the action of the digestive enzyme trypsin.[12,13] Since then the rapid development of computer hardware and software has allowed longer and larger simulations to be carried out. Microseconds of simulation time is now commonplace, and the field is rapidly approaching the millisecond scale with specialist tools. The satellite tobacco mosaic virus, containing over 1 million atoms, was the first complete virus investigated via all-atom MD simulation.[14] Since then, the size and complexity of the systems MD simulations can be used to study has only grown, with simulations being carried out on the HIV-1 mature capsid[15] and, more recently, a microsecond simulation of a complete influenza A virion.[16] MD simulations of larger systems benefit greatly from the parallisabilty of many MD codes, allowing the work to be split over more cores to speed up simulation time. The Folding@home project utilised this to investigate the Chicken Villin Headpiece using 20,000 CPU's from participating home computers.[17] GPU acceleration of MD simulations[18] is likely to increase

further the size of systems studied and the length of simulations carried out.

One of biggest barriers in chemistry (and related disciplines) is the size of systems being studied, since we cannot directly observe the behaviour of a protein in water for example. Many methods in spectroscopy provide a way of observing behaviour of molecules, but this is indirect, based on the interaction of electromagnetic radiation with the molecule in question. MD simulations give us a way to directly observe the behaviour of molecules—with the caveat that the observations are only as accurate as the methods used to model the molecules—and can help explain experimental spectroscopic observations by linking observed simulated dynamics to calculated spectra which can then be compared with experiment.

MD simulations have had a large impact in many fields, having paved the way for the vast field of computational protein folding[6,19]. Early MD simulations led to the current understanding of proteins as dynamical entities rather than static.[20] Investigations of the dynamics of proteins and related systems, whether on an all-atom or more coarse-grained scale, have provided essential insights into the behaviour of these systems.[6,15,16] Combination of MD simulations with calculation of spectra have led to a greater understanding of a number of important protein related systems.[21–25] The drive to obtain MD simulations of ever larger systems at longer time scales has led to developments in computer hardware, software and various improved algorithms.[17,18]

1.2 Vibrational Spectroscopy of Proteins

As stated above, whilst X-ray and NMR methods have been indispensable in developing our understanding of protein structure, these techniques are not necessarily suitable for use on all proteins; for example, X-ray crystallography requires the growing of crystal from the target molecule, and not all proteins will crystallise. Therefore, though IR spectroscopy probes only the secondary structure of a protein, it is still a useful tool in their study, particularly as the

existence of solution phase and solid state methods permits direct investigation of different phases. Indeed, in order to fully understand the structure and behaviour of a protein requires a range of tools and approaches. IR spectroscopy (and its two-dimensional analogue) are the tools which are the focus of this thesis.

1.2.1 Introduction to Vibrational Spectroscopy

Above 0 K all atoms in a given molecule are in motion, though even at absolute zero the energy of a molecule is still not zero. For a system treated as a harmonic oscillator, the ground state energy of will be $^1/_2\hbar\omega$, where $\hbar$ is the reduced Planck's constant and ω is the frequency of oscillation. In a three dimensional space the molecule will have a total of $3N$ degrees of freedom, where N is the number of atoms in the molecule. In the case of a non-linear molecule, three of these degrees of freedom correspond to translational motion and three to the rotational degrees of freedom. The remaining $3N-6$ degrees of freedom correspond to the vibrational degrees of freedom. For a linear molecule (such as CO_2) there is no rotational degree of freedom along the axis of the bonds and so linear molecules have $3N-5$ vibrational degrees of freedom.

The infrared region of the electromagnetic spectrum lies between wavelengths of 1 mm–750 nm. These wavelengths are too long to directly probe molecular structure as in X-ray crystallography, and still too long to induce electronic transitions such as in UV spectroscopy, but the mid-IR region (2 to 25 μm or 4000 to 400 cm^{-1}) can be used to study molecular vibrations. IR spectroscopy involves the vibrational transitions of a molecule and hence is also known as vibrational spectroscopy. There are other methods that fall under the umbrella term of vibrational spectroscopy, such as Raman spectroscopy, but IR is the technique covered within this thesis. A non-linear molecule of three atoms will have three IR active modes: two stretching modes and one bending mode, each with a symmetric and asymmetric form. These modes are given in

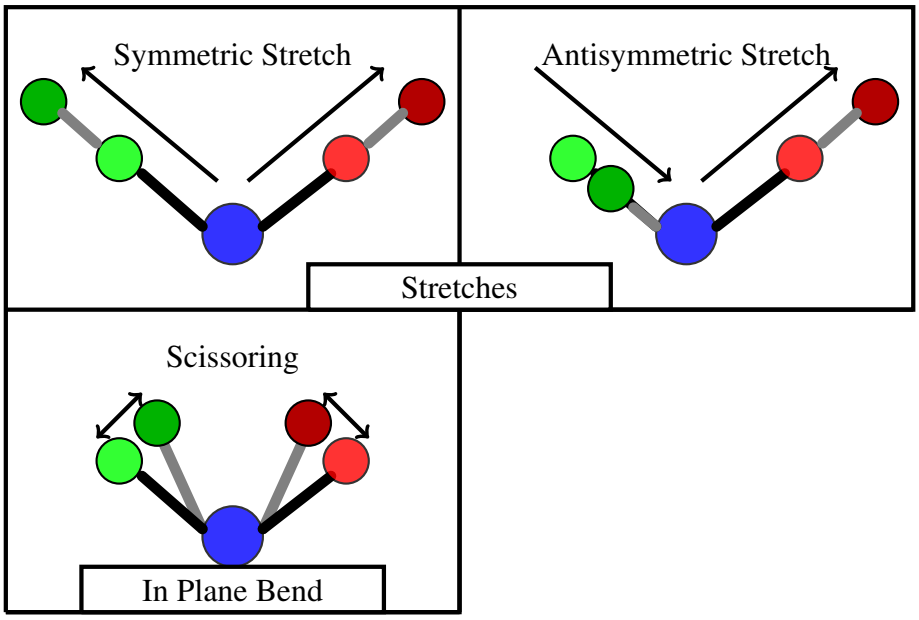


Figure 1.7: The various bond stretching and bending vibrations possible for a system of three atoms with non-linear geometry, e.g. H_2O.

figure 1.7.

Theoretically a bond between two atoms can be treated as a pair of balls on springs, where the balls represent the atoms and and spring represents the bond between them. Taking this approach the vibration of the bond in a diatomic molecule can be described by Hooke's law:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \tag{1.1}$$

where ν is the frequency of the vibration, k is a spring constant that determines the strength of the bond and μ is the reduced mass of the molecule. The larger the value of k, the stronger the bond and the higher the frequency of the bond's vibration. This implies that molecular bonds behave as simple harmonic oscillators, though in a real molecule no bonds behave purely harmonically, requiring the introduction of anharmonicity.

The selection rule for IR spectroscopy states that a vibrational mode will only be IR active (i.e. absorb IR radiation) if the vibration results in a change in dipole moment. The dipole moment for a bond is defined as $\vec{\mu} = Qr$, where Q is the magnitude of the charge at either end of the dipole and r is the distance

between them. This means that homonuclear diatomics are IR inactive, since the vibration of the bond is not associated with any changes in dipole moment. Carbon monoxide, however, is IR active, since bond vibration causes a change in the molecule's dipole moment. A bond involving atoms with different charges or electronegativities will generally be IR active.

Absorption of the incident IR photon results in the bond absorbing radiation with the same frequency as the vibration of the bond. This is known as the fundamental frequency. Since different bonds vibrate at different frequencies, they will absorb IR radiation at different wavelengths. The vibration of a particular bond (e.g. C=O) will be similar regardless of the molecule the bond is present in, hence IR spectra give information on the nature of the chemical bonds within a molecule. Some typical IR absorption bands are given in table 1.1.

Functional Group		Band Position (cm^{-1})
Alkane	C−H	2850–3000
	C−C	800–1000
Aromatic	C−H	3000–3100
	C=C	1450–1600
Alkene	C−H	3080–3140
	C=C	1630–1670
Alkyne	C−H	3300–3320
	C≡C	2100–2140
Alcohol	O−H	3400–3600
	C−O	1050–1200
Ether	C−O	1070–1150
Aldehyde	C=O	1720–1740
	C−H	2700 & 2900
Carboxylic Acid	C=O	1700–1725
	O−H	2500–3000
Ester	C=O	1735–1750
	C−O	1000–1300
Ketone	C=O	1700–1780
Amine	N−H	3100–3500

Table 1.1: Typical IR assignments of organic functional groups. Note: the Ester C−O has two bands.

Today, the most common method of obtaining IR spectra is with a Fourier

Transform IR (FTIR) Spectrometer. In an FTIR spectrometer, a polychromatic beam is produced by an interferometer, and this is what interacts with the sample. Absorption by the sample will be detected in the output interferogram. The resultant interferogram is a time domain signal and this can be converted into the frequency domain i.e. the IR spectrum through Fourier Transform. The nature of FTIR spectrometers also allows accuracy of 0.01 cm^{-1} or better.

1.2.2 Application to Studies of Protein Structure

There are three major bands of interest when considering IR spectroscopy of proteins and polypeptides, designated the amide I, II and III bands.[26] Whilst there are others, e.g., the amide A band, the amide I-III bands are used most. The amide I band is strongly correlated with protein secondary structure, with α-helical and β-sheet structures giving rise to distinct spectral signatures. The amide I band for an α-helix is generally a single peak at around 1650 cm^{-1}. The β-sheet signal appears as two peaks at around 1630 and 1690 cm^{-1}. The amide I mode is primarily a C=O stretching mode, with some contribution from an out-of-phase C−N stretch mode and a smaller C−C−N deformation component. It appears in the 1600-1700 cm^{-1} region of the spectrum and, in the case of FTIR, is generally very broad, congested and difficult to interpret, mainly due to the large number of vibrations in this part of the spectrum. This large number of vibrations is due to the Amide I modes of other amino acids in the protein.

The amide II band (1510-1580 cm^{-1}) consists mainly of an in-plane N−H bend and a C−N stretch. The contributions to the amide II band consist of an out of phase combination with smaller contributions from a C=O in-plane bend, a C−C stretch and an N−C stretch. The amide II band is sensitive to protein structure, but is not as widely used as the amide I band, as the relationship between this band and secondary structure is not as well understood. That said, this band is primarily known for its sensitivity to peptide

unit protonation state[27] and has been used in 2DIR experiments.[28]

Finally, the amide III band (1200-1350 cm^{-1}) arises from a combination of N−H in-plane bends and C−N stretches. It also has contributions from a C−C stretch and a C=O in-plane bend. This mode produces only very weak bands in IR spectra and is not as useful for secondary structure analysis as the amide I band. The amide III band exists in a very complex spectral region and is too low in frequency to probe effectively using ultrafast laser equipment. The amide A band appears around 3500 cm^{-1} and is an N−H stretch located completely within that group. This band is not sensitive to backbone configuration and secondary structure, but it is sensitive to hydrogen bonding strength.[27]

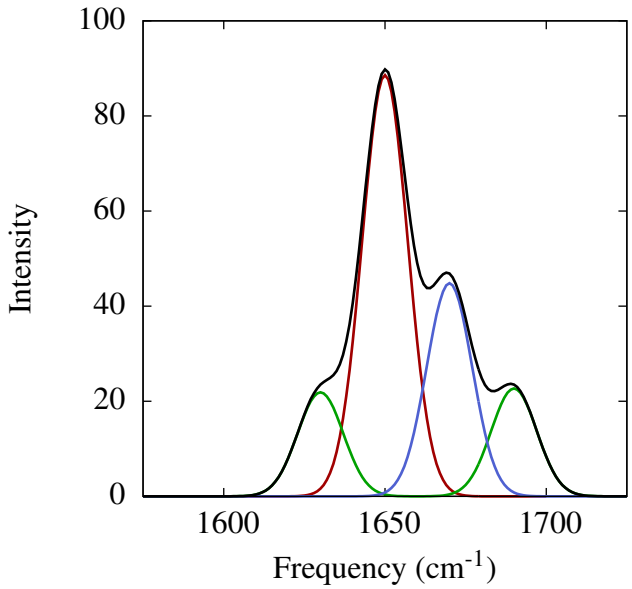


Figure 1.8: Diagrammatic representation of a protein IR spectrum showing the amide I bands of various secondary structure elements with arbitrary, scaled intensities plotted as a Lorentzian with a bandwidth of 10 $cm^{\cdot}$ Red line is α helix, green line β sheet, blue line random coil and the black line is the resultant band profile from summing the components.

Despite its wide use in investigations of protein structure, traditional FTIR is quite limited in the information it can provide. The large number of vibrations in the Amide I region—one for every amide bond present in the protein—means that IR spectra of proteins tend to be very broad and congested. This makes accurate assignment very difficult, and is why assignment is largely restricted to secondary structure elements rather than individual residues. The congestion of

the spectrum is demonstrated in figure 1.8, which gives a diagram of an example spectrum of various secondary structure elements. When multiple secondary structure elements are present, the total band-shape becomes broad and requires deconvolution in order for any assignments to be made. The band-shape used in figure 1.8 is relatively narrow at 10 cm^{-1}; the bandwidth for experimental spectra may be much broader and therefore more difficult to deconvolute.

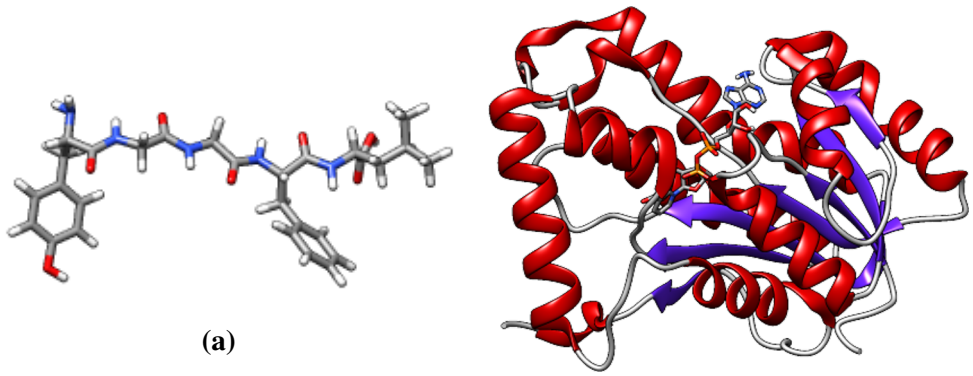
This is one area in which 2DIR may have an advantage over FTIR in the study of protein structure; since the spectrum is spread over two frequency axes, 2DIR produces spectra with more unique spectral lines and this may help improve band assignment. A 2DIR spectrum can interpreted as a 2D map linking a set of excitation frequencies (ω_1) to a set of detection frequencies (ω_3).[29] Features on the diagonal correspond to excitation and detection of the same frequencies, whereas off-diagonal peaks correspond to the excitation of one particular peak and detected emission from a separate peak. This excitation of one peak and detection of another is a reflection of energy transfer, or coupling, between the two modes. Due to vibrational anharmonicity the peaks in the spectrum appear as positive/negative pairs. The peaks along the diagonal roughly correspond to the linear absorption (usually FTIR) spectrum, and the presence of anharmonicity and the cross-peaks in 2DIR gives additional detail to that found in the linear spectrum. Much like in the FTIR, different protein secondary structures tend to give rise to different band shapes in the 2DIR; for example, anti-parallel β-sheets have been shown to give rise to a characteristic 'Z' pattern in the 2D spectrum.[30]

1.3 Outline of Thesis

MD simulations are used, in conjunction with exciton based calculations of 2DIR spectra in order to probe the structure and dynamics of two peptide/protein systems. The two systems chosen are a single polypeptide and a protein-ligand system. The work in this thesis aims to show how 2DIR

calculations—in conjunction with analysis of atomic level MD simulations—can be applied to these different systems and used to gain new insight into the relationship between their structure, dynamics and function.

1.3.1 Molecules Studied



(a) **(b)**

Figure 1.9: Molecules of interest in this thesis: a) Leu-enkephalin pentapeptide; b) Wild-type InhA protein bound to NADH (from PDB 2AQ8).

Two different protein systems are studied in this thesis, representing different levels of peptide/protein function within biological systems. The first system is the polypeptide Leu-enkephalin, which is involved in pain mediation in humans. The other system of interest is the enoyl-acyl carrier protein reductase that forms part of the mycolic acid biosynthesis pathway in *Mycobacterium tuberculosis*. Four variants of this protein are studied in order to model the effects of point mutation and the binding of endogenous and exogenous ligands on the dynamics and structure of the protein.

1.3.1.1 Leu-Enkephalin

Leu-enkephalin is one of a pair of related structures that are endogenous ligands for the opioid receptors. The two structures are pentapeptides that differ only in the nature of the terminal residue which in Leu-Enkephalin is leucine. The enkephalins have a role in nociception within the human body, and they are the smallest known molecules with pain killing or opiate activity. Enkephalins

are found in the spinal cord and central nervous system and have a number of functions besides mediating pain.

Leu-enkephalin binds primarily to the delta-opioid receptor, where it acts as an agonist similar to that of opioid alkaloids like morphine. The structure of the enkephalins, however, is highly flexible resulting in them binding more strongly to the opioid receptors than the opioid alkaloids.[31] This paradox has led to suggestions that the more conformationally dynamic enkaphalins are able to penetrate membranes more readily than exogenous ligands such as morphine, which is rigid and highly inflexible.[32]

1.3.1.2 *Mycobacterium tuberculosis* enoyl-ACP reductase InhA

Tuberculosis is a serious, widespread and often fatal infectious disease caused by *Mycobacterium tuberculosis*. Tuberculosis has long been treated using the drug isoniazid and it has been found that the target for the drug is an NADH-specific enoyl-acyl carrier protein (ACP) reductase coded for by the *inhA* gene. The exact mechanism of action for isoniazid is still unknown and remains fertile ground for research.

The prevalence and mortality of tuberculosis in the world today, coupled with strains that are increasingly resistant to front-line antibiotic treatments, the isolation of the mechanism and suggestion of other targets for tuberculosis treatments is a high priority. A naturally occurring mutation at residue S94 has been identified, which engenders resistance to isoniazid. The exact nature of the changes this mutation causes in the active site is unknown, and identification of these changes could lead to fresh research into new drugs for treating tuberculosis, or help identify a cofactor to increase the effectiveness of isoniazid.

The molecules studied in this thesis are shown in figure 1.9. These two systems play a role in either human pain pathways or human pathogenic bacterial activity. Further understanding of the relationship between the

structure and dynamics of these systems has the potential to provide targets for pharmaceutical research. This is especially important in the case of the bacterial protein; in an age where antibiotic resistance is becoming more serious and widespread, new front-line drugs are needed in order to prevent the deaths of millions of people. Novel targets for research have a role to play in the search for more effective antibiotics, and understanding the structure and dynamics of these proteins—and how they are related—is an important step towards this.

1.3.2 Summary of Chapters

Chapter 2 of this thesis introduces two-dimensional infrared spectroscopy, and its potential advantages over traditional FTIR in the study of protein structure and dynamics. One of the potential experimental set ups is described, along with the processes involved that give rise to the observed 2D spectrum. Some recent literature on 2DIR of protein systems is then reviewed, to demonstrate the applicability of the technique and insights already gained as a result.

Chapter 3 outlines the theoretical methods used throughout this thesis, focusing on force field methods, MD simulations and related concepts. The theoretical basis of 2DIR is described in some detail as a prelude to exciton theory and how it can be used to calculate 2DIR spectra from MD simulations.

Work on Leu-Enkaphalin forms the core focus of chapter 4. MD simulations were carried out on ten different low energy conformations of the pentapeptide in three different media in order to investigate the effect and importance of solvent on the 2DIR spectra of these small molecules.

Chapter 5 presents investigations on the InhA protein using MD simulations and network models derived from the one exciton Hamiltonian matrices of these structures. 2DIR calculations are still largely intractable on proteins of this size, so the exciton Hamiltonians were computed in order to investigate what can be learned from this relatively simple calculation and how

this related to observations in the experimental 2D spectra.

Chapter 6 gives some concluding remarks on the work in this thesis, the impact of 2DIR calculations and presents ideas on further development of this work.

Chapter 2

Two-Dimensional Infrared Spectroscopy

2.1 Introduction

Two-dimensional infrared spectroscopy (2DIR) is a rapidly developing technique[34–36] that can furnish greater structural detail than traditional FTIR. This is done by spreading the transitions over a second frequency domain, exposing cross peaks—off-diagonal peaks which quantify the coupling between molecular vibrations—as well as other vibrational correlations. These correlations contain information on molecular structure, but perhaps more importantly, information about dynamics—that is, the forces and motions characterising the molecule—and solvent effects. The presence of cross peaks helps untangle the information hidden in congested bands and even has the potential to elucidate information on three-dimensional structure.

2DIR is a form of nonlinear optical spectroscopy[37,38] and utilises a series of infrared (IR) pulses, which interact with the vibrational transitions of the sample molecule(s). This is analogous to two dimensional NMR methods, where multiple radio-frequency signals interact with the nuclear spin transitions. 2DNMR provides detailed spectroscopic information, but usually

This chapter is based on work that has been published previously.[33]

offers a time-averaged picture of the molecule due to the intrinsic time-scales. 2DNMR methods are capable of accessing time scales down to the picosecond level, but do so indirectly which can lead to oversimplification of complex models. 2DIR can access time scales from picoseconds to seconds, though it does so at the expense of the detail provided by 2DNMR methods. Hence, while 2DNMR is a powerful tool for studying the structure of many systems, 2DIR offers the capability to study the dynamics in greater detail.[39]

2.1.1 Advantages of 2DIR

2DIR offers many advantages over traditional linear spectroscopies such as FTIR. The spreading of the vibrational transitions over a second frequency dimension gives access to direct measurement of the coupling between two peaks, and hence information on the structure of the system. It enables the effects of homogeneous and inhomogeneous broadening to be separated from each other. Homogeneous broadening of the spectral linewidth is generated by mechanisms that affect the line shape in the same way for each atom in the system. An example of this would be "natural" broadening due to spontaneous emission. Homogeneous broadening is given by a Lorentzian. Inhomogeneous broadening results from effects that broaden the spectral line differently for each atom, such as broadening due to impurities in the sample. Inhomogeneous broadening gives a Gaussian line shape. A 2DIR spectrum will contain information on both structure (in the diagonal peaks and cross peaks which arise from the coupling between sites) and dynamics (which is obtained from looking at the diagonal line shapes). This makes 2DIR a powerful tool for studying molecular structures, environmental dynamics, and structural kinetics.

Figure 2.1 gives an example 2D spectrum showing both the $v = 0 \rightarrow 1$ and $v = 1 \rightarrow 2$ transitions. 2D spectra measure "overtone" and combination bands alongside the fundamental transitions which gives information on the anharmonicity of the vibrational modes. The off diagonal is often referred to as

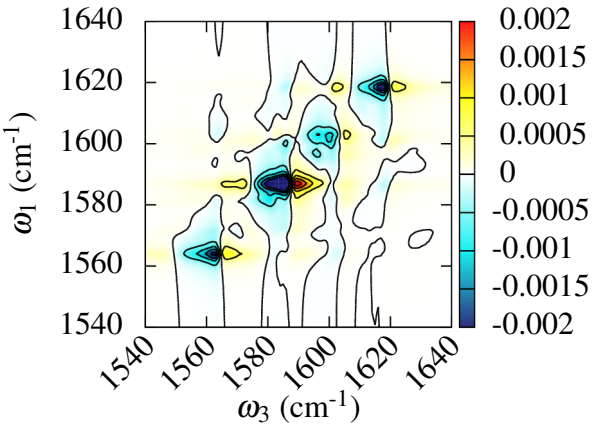


Figure 2.1: 2DIR spectrum of tryptophan zipper trpzip2 (PDB code 1LE1) showing the $v = 0 \rightarrow 1$ (fundamental) transitions on the diagonal (shown in blue). The $v = 1 \rightarrow 2$ "overtone" transitions are the off diagonal peaks and are shown in red. From chapter 10 of *Concepts and Methods of 2D Infrared Spectroscopy* by Hamm and Zanni.[38]

the overtone in 2DIR literature even though this is technically incorrect. The overtone would be the $v = 0 \rightarrow 2$ transition. The transition actually present in the spectrum is the $v = 1 \rightarrow 2$, making it a sequence band. The shape of the 2D peak also gives information on the frequency fluctuations of the vibrational modes, which are related to the dynamics of the environment.

The extra detail offered by 2DIR is especially useful when studying proteins. FTIR is a well established technique for studying protein structure since the Amide I band is primarily a C=O stretch of the amide carbonyl and appears between 1600 cm^{-1} and 1700 cm^{-1}. It is sensitive to the secondary structure conformation of a protein. For example, an α-helix will tend to give rise to a single peak at around 1650 cm^{-1} while a β sheet will produce two peaks; the strongest at around 1620 cm^{-1} and a weaker band at around 1690 cm^{-1}. This secondary structure sensitivity is is an invaluable tool when investigating proteins and peptide structures. However, the Amide I band region of proteins is very congested, often appearing broad and featureless due to the large number of amide units present in a large protein. 2DIR offers a method for extracting some of the information that is under the broad amide I band in FTIR.

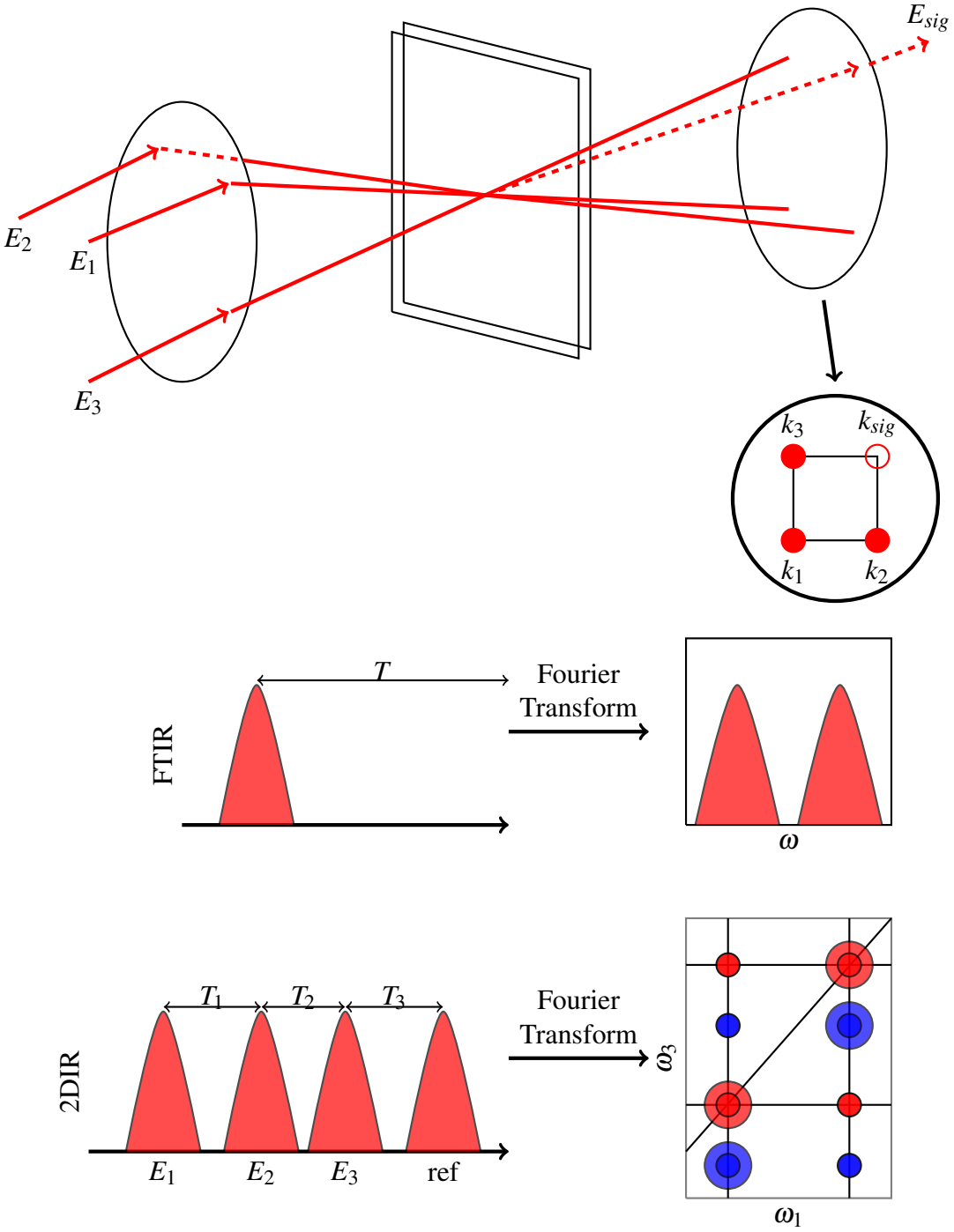


Figure 2.2: Diagrammatic representation of the experimental set up for a photon echo experiment with example pulse sequences for both 2D photon echo and FTIR experiments.

2.1.2 Experimental Methods

Two methods have developed for acquiring 2DIR data experimentally; the first is frequency domain double resonance 2DIR[34,36] based on a standard pump-probe experiment, and the second is a time domain, Fourier transform method based on vibrational photon echoes.[35] The techniques both probe the same third order response function by means of a four wave mixing approach; three IR pulses interact with the sample, producing the desired non-linear signal field. The information obtained from both frequency and time domain methods is identical and related by the Fourier transform relationship. The two methods differ only at the technological level. A pump-probe set-up for picoseconds work is only a two laser experiment.

The most widely used photon echo method involves four beams; the incident laser is split into three beams arranged at three corners of a square focused into the sample. The fourth beam—the signal or echo pulse—is emitted towards the fourth corner of the square. The three laser pulses in a 2DIR experiment interact with the sample at three separate instances. The first pulse interacts with the sample to produce a coherent superposition or mixed state of the ground and first excited state. The interval following the first pulse is known as the coherence evolution period (t_1); initially the excited molecules oscillate in phase, but slight variations in the frequency of molecular vibrations cause the initial phase coherence to be lost, a process known as dephasing. The second pulse converts the superposition into a population state either of the ground or first excited state, depending on the processes involved. The period after the second pulse is the waiting time or mixing period (t_2). It is during this period that vibrational dynamics are observed and the length of t_2 can be varied. The third pulse reproduces the superposition state, the make-up of which depends on the nature of the population state occupied during t_2. The sample does not immediately regain coherence, instead it undergoes a rephasing process in which, after a time comparable to t_2, the coherence is recovered and the

polarization associated with this results in the emission of the echo pulse. The general experimental set-up and pulse sequence is shown in figure 2.2.

The major benefit of the photon echo method over the double resonance is that the former retains the time resolution that is lost in the latter, due to the reduced pump pulse bandwidth.[39] This arises from the ability to separate the time orderings of the laser pulses and probe more completely the way in which the laser and the molecule intersect. While the spectral peaks appear the same for both methods, the photon echo spectrum is free of the distortions which appear in the pump pulse experiment as a result of the finite bandwidth, i.e., the finite range of frequencies in the wave. The vibrational lifetimes of the key vibrational modes in proteins (i.e. the amide I band) is less than 2 ps.

2.1.3 Origin of Peaks in 2DIR Spectrum

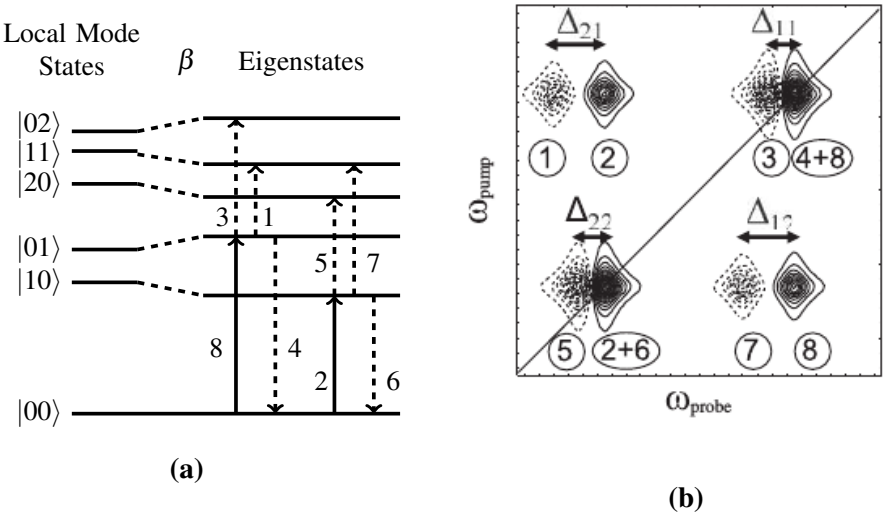


Figure 2.3: a) Level scheme of two coupled oscillators before coupling (local modes) and after coupling (eigenstates) with dipole-allowed transitions depicted (pump processes are given by the solid lines and probe processes by the dashed lines) and b) the resulting 2DIR spectrum. Labels relate the peaks in the spectrum to the transitions in the energy level diagram. Dashed lines are a positive response and solid lines are a negative response. Figure adapted from *Concepts and Methods of 2D Infrared Spectroscopy* by Hamm and Zanni.[38]

Figure 2.3 gives the energy level diagram for two coupled oscillators and an example spectrum. The labelling nomenclature is such that $|00\rangle$ is the

ground state, $|10\rangle$ and $|01\rangle$ are the singly excited states and $|20\rangle$, $|11\rangle$ and $|02\rangle$ are the doubly excited states. The nomenclature for this diagram is the same as in *Concepts and Methods of 2D Infrared Spectroscopy* by Hamm and Zanni[38] from which the figure has been adapted. Assuming that the anharmonicity is small, the selection rules for a harmonic oscillator will apply and the only allowed transitions will be the ones where the state of one oscillator changes by one quantum at a time; $|10\rangle \rightarrow |20\rangle$ is allowed while $|10\rangle \rightarrow |02\rangle$ is not. The arrows in figure 2.3 show all the allowed transitions for a system consisting of two coupled oscillators. The energy level diagram in figure 2.3a allows for the construction of a 2DIR spectrum (such as in figure **??**) for a pump-probe experiment. 2DIR from photon echoes will be considered separately.

During a pump-probe experiment, if the pump frequency comes into resonance with the higher frequency oscillator, the $|01\rangle$ state will be excited as shown by transition 8 in figure 2.3a. The probe pulse that follows will then have three possible transitions corresponding to those labelled 1, 3, and 4 in the figure. In addition to these transitions, since there are now fewer molecules in the ground state able to make transitions 8 and 2, a bleach in both oscillators will be observed. For the probe pulse, transitions 8, 4 and 3 are the bleach, stimulated emission and excited state absorption respectively. If the pump pulse is in resonance with the lower frequency oscillator then the $|10\rangle$ state will be excited (transition 2) and the probe pulse will have possible transitions corresponding to those labelled 5, 6 and 7.

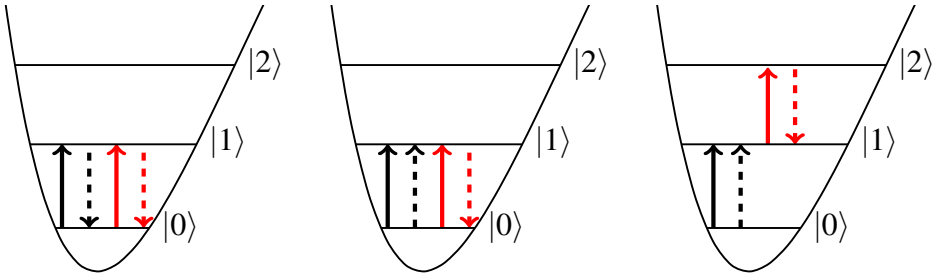
Transitions 2 + 6 and 4 + 8 give rise to the diagonal peaks, which are negative since the probe pulse gives rise to a bleach as discussed in the previous paragraph. The off-diagonal portion of the doublets correspond to the $|01\rangle \rightarrow |02\rangle$ and $|10\rangle \rightarrow |20\rangle$ transitions, labelled 3 and 5. This is sometimes referred to in the literature as the overtone peak (as mentioned in section 2.1.1) though this is technically incorrect. This off diagonal peak appears when the frequency of the pump pulse (ω_{pump}) is resonant with the $|01\rangle$ transition (8) and

the probe pulse (ω_{probe}) is resonant with the $|02\rangle$ transition (3), giving rise to the positive peak for the higher frequency oscillator. The diagonal peaks will be separated by the anharmonicity of the $|01\rangle$ transition.[39] The same is also true for the lower frequency oscillator.

The coupled nature of the oscillators is what results in off-diagonal cross-peaks. Since these peaks are absent if the oscillators are uncoupled, these peaks reveal additional information on the structure of the system under study. Cross-peaks arise in the spectrum as a result of the following: in the case that the higher energy oscillator is already excited to the first excited state then transition 1 will also excite the lower energy oscillator to the first excited state leading to the following transition: $|01\rangle \rightarrow |11\rangle$. If the two oscillators were not coupled then the excitation frequency of the second oscillator would not depend on the state of the first oscillator resulting in transitions $|00\rangle \rightarrow |10\rangle$ and $|01\rangle \rightarrow |11\rangle$ having the same frequency. Since the two transitions are equal in strength but opposite in sign, peaks 1 and 2 would cancel each other and a cross-peak would not be observed there. In addition, for a cross-peak to be observed the off-diagonal anharmonicity Δ_{21} must be non-zero, otherwise they will again cancel. The two peaks are separated by the anharmonicity of the off-diagonal and can be directly read from the spectrum.[39] The same process applies if the lower energy oscillator is in the first excited state, leading to cross-peaks at 7 and 8.

2.1.4 Processes Involved in the Photon Echo experiment

To outline the processes involved in generating a 2D spectrum, let us take the example of a single vibrational mode, such as the carbonyl stretch of an acetone molecule. In order to generate the 2DIR spectrum (with only diagonal peaks; cross peaks will be dealt with later) all that is required are the eigenstates and the transition dipoles of the vibrational modes of interest. If the potential energy curve is modelled as a Morse oscillator then the vibrational energy levels



(a) **(b)** **(c)**

Figure 2.4: Energy level diagram of an anharmonic oscillator showing the dipole allowed transitions incited by a photon echo experiment for a) the ground state bleach, b) the stimulated emission and c) the excited state absorption. For each signal the pulses are represented as follows: first pulse (solid black line); second pulse (dashed black line); third pulse (solid red line); signal (red dashed line).

for this oscillator are given in figure 2.4.

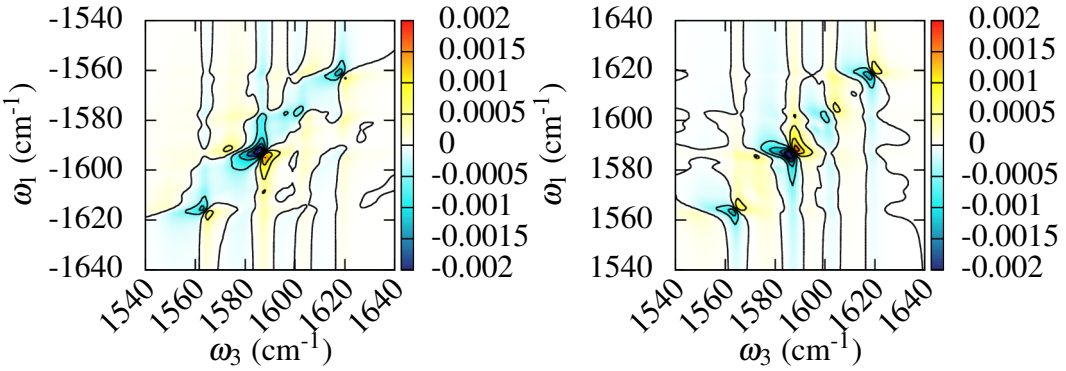
Whenever the laser frequency is in resonance with a dipole-allowed $0 \rightarrow 1$ transition, a fraction of the molecules in the system will be excited from the ground state $|0\rangle$ to the first vibrationally excited state $|1\rangle$. In the photon echo experiment (shown in figure 2.2) it is the first laser in the sequence (E_1) that generates this superposition of the ground and first excited states. When the second pulse (E_2) arrives, one of two things can occur: in the first situation, the pulse will return the system to the ground state, $|0\rangle$, leaving the third pulse (E_3) to regenerate the superposition after waiting time t_2. The system will relax back down to the ground state by the emission of the signal. This is known as the ground state bleach.

In the second case, the second pulse (E_2) in the train will excite the molecules still in the ground state to the first excited state. In this case, the system will either exist in a single excited state or as a superposition of two singly excited states. The third pulse (E_3) causes the system to relax back down to the ground state via the emission of the signal. This is known as the stimulated emission. In both of these cases, the second laser pulse will be less strongly absorbed than the first since there are now fewer molecules in the ground state.

In the third type of signal, the first laser pulse (E_1) excites some of the molecules to the first excited state as before. The second laser pulse (E_2) in the train excites the remaining molecules into the first excited state, either as a single state or a superposition of two, similar to the stimulated emission pathway. The third pulse (E_3), however, excites from the first excited state to create a superposition of the first and second excited states. The signal is emitted via relaxation from the second excited state back to the first excited state. This signal is known as the excited state absorption. Since doubly excited states are involved, this signal is sensitive to anharmonicity; if these are not present then the total signal vanishes.[40]

A 2DIR spectrum is made up of six types of contributions, as each of these signals have a rephasing and non-rephasing pathway. The rephasing signal is given as $k_{sig} = -k_1 + k_2 + k_3$ and the non-rephasing signal as $k_{sig} = +k_1 - k_2 - k_3$, where k_1, k_2 and k_3 are the incident photons at the sample (as shown in figure 2.2) and k_{sig} is the emitted signal. These signals arise as a result of slight differences in pulse sequence for the photon echo experiment. The pulse order for the rephasing pathways is k_1, k_3, k_2 as shown in figure 2.2. The sequence for the non-rephasing pathways is k_3, k_1, k_2. The vibrational phase of the rephasing pathways is given as $e^{-i\Omega t_1}e^{+\Omega t_3}$, where Ω is the frequency, t_1 is the time period after the first pulse and t_3 is the time period after the third pulse in a photon-echo experiment. The vibrational phase of the non-rephasing signal is $e^{+i\Omega t_1}e^{+\Omega t_3}$. A consequence of this different in phase is that the signals appear in difference quadrants of the spectrum, meaning that the ω_1 must be multiplied by -1 before addition to the non-rephasing signal in order to obtain the purely absorptive signal.

Generally speaking the rephasing and non-rephasing signals can be distinguished from each other but cannot be disentangled further. For example, the rephasing form of the ground state bleach, stimulated emission and excited state absorption cannot ordinarily be distinguished from each other. The



(a) Rephasing **(b)** Non-rephasing

Figure 2.5: Rephasing and non-rephasing signal of a tryptophan zipper (PDB 1LE1). Note that the signals appear in different quadrants and have opposite phase twists.

combination of all six signal pathways gives rise to the observed, purely absorptive, 2DIR spectrum.

2.1.5 Obtaining Information from 2DIR Spectra

Since 2DIR has the potential to provide a wide range of information on the target being studied, careful consideration must be given to the goal of each experiment as this will inform the design of the 2DIR experiment and hence the type of analysis that is done. Hydrogen bonding and other short range structure and dynamics can be investigated using 2D line shape analysis[41] on localized vibration. On the other hand, the global structure of a protein system would be better investigated via delocalized amide I spectra (with the potential for using isotope labelling of specific residues of interest).[42] Non-equilibrium phenomena can be investigated with triggered experiments such as temperature jumps.

The diagonal peaks in the spectrum reveal the vibrational frequencies of the sample under study in much the same way as traditional FTIR does. In fact, along the diagonal of the spectrum (where $\nu_{pump} = \nu_{probe}$) should provide the same information as the one-dimensional experiment. In 2DIR, however, the diagonal peaks appear as a doublet, and the off-diagonal portion of the doublet

provides information on the transition between the first excited state and the second excited state $|01\rangle \rightarrow |02\rangle$.

Coupling of two (or more) molecules leads not only to cross-peaks in the spectrum but also to changes in the diagonal peaks. Coupling creates a multidimensional potential energy surface that is different to the potential energy surface for each isolated molecule. This leads to changes in the diagonal peak frequencies, intensities and anharmonic shifts. Coupling strength and orientation of molecules can be extracted from this information.

2DIR spectra of an ensemble of molecules will give rise to peaks that are broadened—either homogeneously or inhomogeneously. For example, within a protein molecule the vibrational frequency of the C=O of each individual peptide unit will be affected by the surrounding environment; secondary structure, hydrogen bonding status (and whether the hydrogen bond is with solvent or another peptide unit), electrostatic interactions with neighbouring units will all have an effect. The result of this is that the spectrum will show a distribution of frequencies along the diagonal at any given instant—this is *inhomogeneous* broadening. Each molecule also has a *homogeneous* linewidth that is dependent on the vibrational lifetime of the molecule. The inhomogeneous broadening of the line shape gives information on the conformational diversity of the sample; a signal that is inhomogeneously broad indicates an ensemble with greater conformational diversity than a signal that is less inhomoegeously broadened.

For large polypeptides and proteins the 2D spectrum can still be congested due to the large number of individual peptide vibrations in the amide I region of the spectrum leading to complex line shapes. Analysis of delocalized amide I spectra of proteins can reveal information on the secondary structure content of the protein in much the same way as FTIR spectra do. Like the one dimensional experiment, 2DIR of proteins with different secondary structure elements show distinctive spectral features that can be interpreted qualitatively. Deeper

analysis can be obtained using simulation and theoretical models. Studies of 2DIR spectra for proteins with different secondary structure content show that proteins with a high α-helix content (such as myoglobin) will tend to show a round diagonal peak in the region of 1650 cm^{-1}. On the other hand, proteins with high β-sheet content (like concanavalin A) will tend to show more elongated diagonal peaks, often forming a "z" shape in combination with the off diagonal peaks.

In the case of protein systems (which is the focus of this thesis) 2DIR spectra can still appear congested due to the large number of vibrations in the amide I region. The detail contained within the 2D spectrum can be extracted via isotope labelling of specific residues of interests with a $^{13}C{=}^{18}O$ moiety which redshifts the frequency of the isotope labelled residue by approximately 60 cm^{-1}. Since a 2DIR spectrum is essentially the sum of the 2D response for each residue, isotope labelling has the effect of moving the peak for that particular residue out from under the main peak so the coupling and line shape for that residue can be examined in greater detail.

2.2 Studies of Proteins

Since the early 2DIR experiments[34] 15 years ago, the technique has developed into a powerful tool; the first commercial 2DIR systems are becoming available and the full range of potential applications is starting to be realised. A comprehensive review of all 2DIR investigations of biomacromolecules is beyond the scope of this chapter. Therefore, the focus is on papers published from 2009 onwards, giving an overview of more recent applications and developments in the field.

Flexibility and dynamics play an important role in protein (thermodynamic) stability, structure and function. One current challenge in biophysics is the accurate characterisation of protein dynamics on very fast time scales. Direct measurement of the exact relationship between dynamics,

structure and function of a protein remains a challenge, in part due to the fact that protein behaviour occurs across numerous time scales, from the picosecond to the second. It is difficult to carry out a single experiment that probes all these time scales simultaneously. However, much work has been conducted to flesh out the picture of protein dynamics on a picosecond time scale. 2DIR is capable of picosecond time resolution and may allow the exact nature of this relationship to be discovered.

2DIR is well placed to investigate the structure of proteins and peptides, because of the information available within the amide I band. These vibrations are sensitive to local variation of structure, leading to inhomogeneous broadening of the line shape, due to the static distribution of different frequencies arising from these local variations. The exact nature of the relationship between frequency and amplitude of IR bands and the structures responsible for these vibrations is not entirely understood. However, 2DIR spreads the spectra over a second frequency axis, which eliminates some of this congestion and allows extra structural and dynamic information to be obtained. In addition, there is a need to understand solution phase structure of proteins, since all proteins carry out their function in the solution phase. With its structural sensitivity and high time resolution 2DIR is a useful technique for this type of investigation.

2.2.1 Secondary Structure

A more complete understanding of how secondary structure elements give rise to observed spectral indicators would have profound implications not only for understanding folding pathways, but also for ligand binding and drug development. Techniques for probing the structure of proteins include cryo-electron microscopy and X-ray crystallography and NMR. X-ray crystallography has allowed the structures of numerous proteins to be elucidated. However, many proteins are difficult to crystallise, and their

structure cannot be solved using this method, furthermore there is a real need to understand protein structure in real biological environments, i.e. in solution where interactions with surrounding water molecules is known to play an important role in determining both protein structure and function. This latter point being a driver for the development of 2DIR. NMR is often used to solve protein structures and is sensitive to the structure, heterogeneity or conformational dynamics of proteins. Despite the numerous tools available, several classes of proteins have proven difficult to characterise, including fibrous proteins, intrinsically disordered proteins, gels, amyloids, and aggregates. Consequently, a method of quantifying the fraction of amino acids in α helices or β sheets would be particularly useful for these proteins. Amyloid fibrils are discussed in more detail in section 2.2.4.

One of the major drives for the development of 2DIR in the last decade is the scope the technique has for using 2DIR to measure the structure of proteins in solution, relying on the specific signatures of secondary structure elements.[42] The method assumes that there is a unique spectrum associated with each individual secondary structure and that the spectrum for the whole protein is the sum of these individual contributions weighted by content. 2DIR is a more accurate predictor of structure, since it provides more spectral information than FTIR.

Testing of the model on 16 commercially available globular proteins with well-defined structures, including myoglobin, insulin and ubiquitin,[42] showed that 2DIR is capable of quantitatively determining the secondary structure content of stock protein without need for isotopic labelling. The errors associated with structure determination using 2DIR were comparable with those associated with electronic circular dichroism (CD), a standard method of measuring protein conformation in solution. CD experiments are easier to carry out than 2DIR, but the extra, off-diagonal information available through the latter may be useful. Exploiting its ultrafast time resolution and the ability to

isotope label individual residues to increase structural resolution, 2DIR has been combined with a laser-induced temperature-jump[43] to follow protein (un)folding in solution.

A theoretical study investigated the suitability of isotope-labelled 2DIR for obtaining site-specific structural information on trpzip2.[44] Trpzip2 is an artificial β-hairpin peptide specifically designed as a model system for studying protein folding in β-sheets. It contains four fluorescent tryptophans and its amino acid sequence is given in table 2.1 Four folding mechanisms have been suggested: zip-out, hydrophobic collapse, reptation and the hybrid zipper model. The zip-out model has the turn forming first, followed by the inter-strand hydrogen bonds forming successively towards both termini and the hydrophobic core forming at the end. The hydrophobic collapse model has the hydrophobic core forming first and followed by the inter-strand hydrogen bonds forming in both directions. In the reptation mechanism there is a sliding motion of the β-sheets, in which non-native hydrogen bonds are formed initially with the native state achieved at the end of the slide. The hybrid zipper model has the turn forming first followed by a collapse of non-polar residues into a hydrophobic core, with terminal salt bridges stabilising the core. The cross-strand hydrogen bonds form last.

Structure	Sequence
trpzip1	SWTW**EGNK**WTWK
trpzip2	SWTW**ENGK**WTWK
trpzip3	SWTW**EpNK**WTWK

Table 2.1: Sequences of the three tryptophan zippers[45] with turn sequence highlighted in bold. p=D proline.

Trpzip1 and trpzip3 are related zippers, which differ from each other only in turn structure and yet have different folding rates. This suggests that the formation of the turn is a key step in the folding process. There have been several studies on isotope labelling various residues of the peptide.[30,46,47] The vibrational frequency of the labelled carbonyl is red shifted, due to its heavier

mass, and this forms isolated peaks in the spectra. However, temperature-jump experiments on just these few labelled sites would not be sufficient to differentiate between the proposed folding mechanisms. Lessing and co-workers, therefore, focussed on all of the amide I sites in the peptide.[48] The proteins were simulated using MD and at every snapshot (saved every 10 fs) in the trajectory and the exciton Hamiltonian calculated. Exciton theory is covered in more detail in section 3.2.2. The site frequencies were found using a Stark effect based method and corrected using a nearest neighbour frequency shift map, rather than an electrostatic map. The 2DIR spectra were calculated from the time-dependent Hamiltonian using the NISE (Non-linear Integration of the Schrödinger Equation) method.[49] This method is discussed further in section 3.2.3.4.

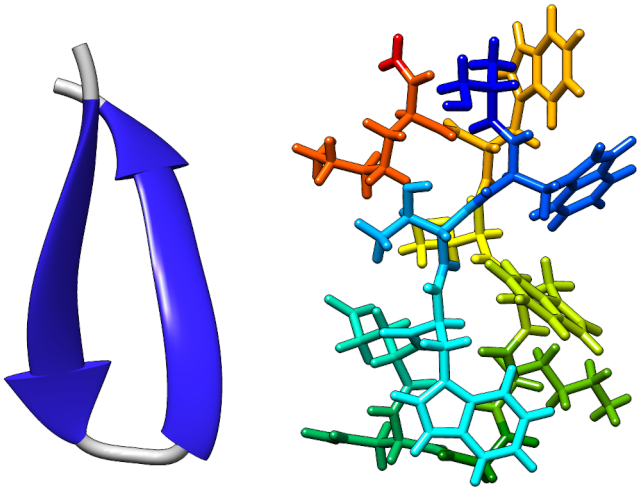


Figure 2.6: Structure of tryptophan zipper trpzip2 showing orientation of the β-strands. All atom model coloured by residue. Rendered with Chimera.[50]

The variations of several spectral line shape parameters were investigated, such as the diagonal and anti-diagonal line width of the 2DIR peaks, in order to examine their utility in classifying the amide I groups of trpzip2. The amide I groups were classified according to their location in the hairpin: terminal, internal, external and turn and could be recognised by the standard deviation of their amide I frequency fluctuations. This fluctuation is related to the diagonal

and anti-diagonal peak widths in an isolated site within the 2DIR spectrum. The anti-diagonal widths were broader for the water exposed sites and narrower for those sites with carbonyl groups buried in the centre of the protein. The broadening of the line shape is mainly due to the electric field generated by the water acting on the amide group. This quantification of the line shapes gives a method of differentiating between the internal sites and the solvent exposed sites. The anti-diagonal line width of peaks in an isotope-labelled 2DIR spectrum should prove useful for studying the folding dynamics in both trpzip2 and other peptides. During unfolding, carbonyl groups that were previously in the interior of the peptide will become exposed to water, resulting in the broadening of the spectral line. By labelling the internal carbonyl sites and performing temperature-jump experiments, it should be possible to distinguish between the proposed zip-out, hybrid zipper, hydrophobic collapse and reptation folding mechanisms.

Theoretical models for coupling between residue pairs in parallel β-sheets were experimentally tested and environmental effects on coupling were investigated.[51] A synthetic peptide macrocycle was used, in order to avoid the issues with using natural β-sheets, i.e., the comparative rarity of parallel β-sheets and the fact small peptides do not readily form the structure. The macrocycle (structure shown in figure 2.7 comprised two strand-forming segments, each consisting of eight L-α amino acids. The strands are linked at the N-termini with a glycyl-succinyl unit and at the C-termini by a D-prolyl-1,2,-diamino-1,1-dimethylethyl unit. Inter-strand β-sheet interactions are promoted by these terminal linkers. Isotope labels were applied to pairs of residues to sample the interactions between the residues believed to give rise to the most significant elements in the exciton Hamiltonian matrix. The understanding gained from both experiment and simulation will allow a fuller interpretation of IR spectra of parallel-β-sheets and this study will help improve the design of isotope labels for testing structural models.

It is important to obtain a precise understanding of parallel-β-sheet IR spectra, since vibrational spectroscopy provides a useful tool for investigating the structures of amyloid fibrils, which have proven difficult to characterize using other methods such as X-ray crystallography and are mainly comprised of β-sheet structures. Amyloids are discussed in further detail in section 2.2.4. There are already many spectra, both simulated and experimental, for antiparallel β-sheets, but this study by Woys et al. is one of the most comprehensive on parallel β-sheets. Their results assess the coupling models utilised over the past five decades to interpret IR spectra of β-sheets in proteins. The influence of the environmental leads to variations which are comparable to or larger than the inherent couplings within secondary structure. Because of this, the frequency distribution of the secondary structure is not substantially larger than that of random coils. The primary effect of coupling is the redistribution of the oscillator strength, as even a very small coupling significantly changes the IR band intensities. Previous work on nitrile groups demonstrated how small couplings affected the intensities but not the frequencies.[52] This benchmark study indicates that refining the orientation of the transition dipole with respect to oxygen, carbon and nitrogen may allow more accurate calculation of intensities.[53] This study strengthens confidence in the accuracy of calculations of solvent-dependent spectral quantities, such as line widths and frequency shifts, and fully demonstrates how FTIR and 2DIR

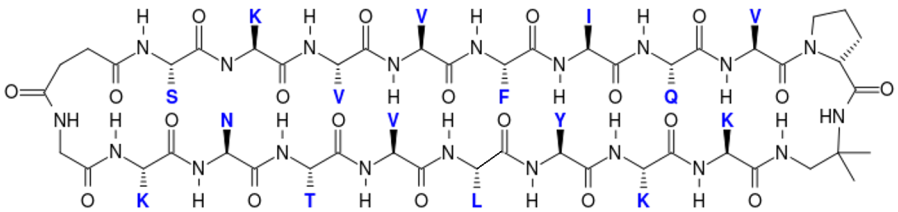


Figure 2.7: Structure of the synthetic peptide macrocycle model of a β-sheet with N-terminus glycyl-succinyl unit linker and C-terminus D-prolyl-1,2,-diamino-1,1-dimethylethyl linker. Side chains indicated by the 1 letter amino acid code shown in blue.

are becoming more quantitative tools for solution phase protein structure elucidation.

2.2.2 Peptides and Small Protein Domains

Even polypeptides as short as trimers adopt only a limited number of conformations in aqueous solution.[54,55] Understanding the conformations of these short peptides may shed light on protein folding pathways. One of the early steps in a folding pathway could well be the formation of local secondary structure, which might serve as a nucleation site for the folding of the whole protein. The location of such nucleation sites might be predictable, if the conformational preferences of short peptides and their dependence on environmental factors, such as pH or temperature, could be investigated systematically. 2DIR offers a powerful tool with which to conduct this investigation.

The structures of tripeptides AHA and AKA have been probed using 2DIR and an excitonic model for the coupling of the amide I' modes.[21] The prime indicates that measurements were taken in D_2O. The protonation of side chains or C-terminal groups does not significantly affect the backbone dihedral angles ϕ and ψ. This indicates that low pH might not be a barrier to secondary structure formation for these peptides. The couplings between the amide I' modes and the angle between the transition dipole vectors were used to analyse the data. The confidence limits of these variables were used to ascertain the upper and lower boundaries for the dihedral angles. From these limits, confidence regions were determined for the tripeptide conformations in the (ϕ, ψ) map. A confidence region is a generalisation of confidence intervals applied to several variables (in this case ϕ and ψ) simultaneously. There was a highly non-linear relationship between the observed couplings and transition dipoles and the dihedral angles (ϕ, ψ), leading to a complicated confidence region for these. This may have an impact on future work, as any method which utilises

intramolecular coupling (such as NMR[22]) will be subject to analogous considerations.

MD simulations and theoretical 2DIR calculations have been carried out on a *de novo* 17-residue α-helix (Protein Databank Code 2I9M).[56] It is a recently designed peptide that folds autonomously into an α-helix without the need for disulfide bond formation. Its sequence is SAAEAYAKRIAEAMAKG. Zhang and colleagues started with an extended structure of the peptide and followed a single trajectory at room temperature and in explicit solvent to capture the folding event. The trajectory was produced with two different force fields: a standard, non-polarizable force field and a force field in which the atomic charges were replaced with the adaptive hydrogen bond specific charges (HBC) scheme. The HBC scheme treated the formation and breaking of hydrogen bonds by including their polarisation into the folding process; residues not involved in the formation or breaking of hydrogen bonds were not included in the scheme. Folding occurred within 10 ns using the polarizable force field with the HBC scheme. The standard, non-polarisable force field did not show a folding event due to weak hydrogen bonding, though partially folded structures were observed.

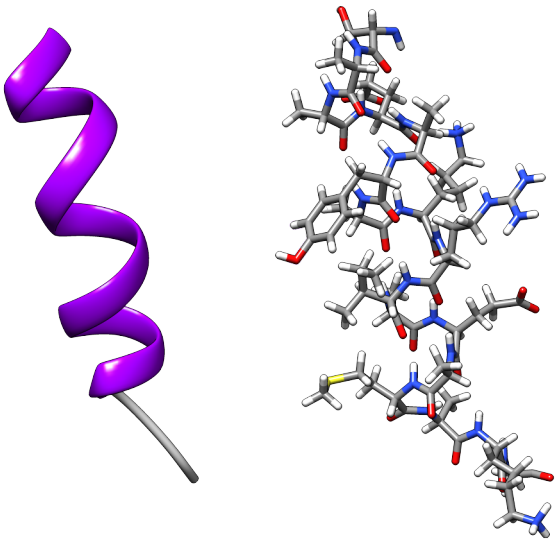


Figure 2.8: Structure of 17 residue *de novo* α-helix PDB code 2I9M. Rendered with Chimera.[50]

The 2DIR spectra were calculated from the MD simulations and averaged over 1000 snapshots of the trajectory. The linear absorption is calculated from the frequency of the vibrational excited state, the homogeneous dephasing rate and the vibrational transition dipole from the ground state to the excited state. The frequency of the vibrational excited state and the vibrational transition dipoles were both obtained using an electrostatic map.[57] The optical response used in the 2DIR calculation was generated by solving the non-linear exciton equations[58]).

The simulated folding of this peptide was in accordance with the funnel free-energy landscape theory for protein folding. The system folded autonomously following downhill folding kinetics. The 2DIR simulations provided information that had not been previously accessible with other experimental methods. Prior to this study, both experimental and theoretical 2DIR studies had focused on either native protein or denatured protein structures. In combination with isotope labelling the 2DIR simulations provided residue specific information. The formation of backbone hydrogen bonds, and the resulting formation of the α-helix, manifest as changes in the line shape and position of the diagonal and cross peaks in the 2DIR spectra. This study represents the first time the complete folding process of a helix has been characterised and its corresponding spectral changes mapped. The experimental 2DIR spectra have yet to be obtained, but as 2DIR advances it should become possible to follow the folding process experimentally. Most current 2DIR studies on protein folding have focused on the unfolding process using a temperature-jump approach[43] or on photoswitchable proteins.[59] Conclusions can be drawn from the structural and dynamical information provided by the MD simulations, but calculating the 2DIR spectra allows the accuracy of the simulations to be verified.

The chicken villin headpiece 35 (HP35) is a 35 amino acid residue peptide. It is a sub-domain of the F-actin binding protein and has been a model for many

computational and experimental studies.[60,61] Despite its small size, this protein has many characteristics that are usually associated with larger proteins, such as autonomous folding, substantial resistance to various point mutations and a hydrophobic core that stabilises the tertiary structure. As one of the fastest folding peptides, it has been an obvious candidate for characterisation by MD simulations and experimental techniques such as fluorescence resonance energy transfer (FRET). Recently, 2DIR has been used to examine HP35, using nitrile (CN) groups attached to the peptide, to act as a vibrational dynamics label.[62–64] Variants of HP35 with one (HP35-P) and two (HP35-P2) cyanophenylalanines in the hydrophobic core were studied. Within experimental error, the absorption spectra and the dynamics of the two variants measured by 2DIR were equivalent. That is, the addition of the CN label via substitution of phenylalanine with cyanophenylalanine only mildly perturbs the structure and dynamics of the system and should be anticipated to be the case for larger proteins also. The nitrile label was used to measure the dynamics of both HP35 and its fast-folding mutant HP35 NleNle. HP35-P NleNle has two charged amino acids in its hydrophobic core replaced by non-polar amino acids which resulted in substantially slower folding. The mutation gave a more stable peptide, as measured by temperature-dependent CD for the CN-labelled peptides, in accord with work on unlabelled peptides.[60,65] The increased stability and faster folding time of the mutant, thus, may be a consequence of the change in structure of the folded peptide. Both HP35-P and HP35-P NleNle would be suitable candidates for further MD simulations.

Intrinsically disordered proteins have been relatively recently recognised as an important class of proteins. The lack of information about intrinsically disordered proteins is a direct result of a dearth of suitable structural tools; they cannot be crystallised and X-ray diffraction and scattering experiments give incomplete information. NMR is a promising tool, but can only probe long-lived conformational states (generally considered to be in the microsecond

region, but this is not definitive) and cannot fully resolve the heterogeneity of these systems. The sub-picosecond time resolution of 2DIR is well suited for studying disordered systems and characterising distinct conformers and structural variations.[66–68]

2DIR and MD simulations were used to study the various turn conformations that exist in the GVGXLPGVG family of peptides,[48] where XL can be a number of different amino acids. Snapshots were taken from the MD trajectories and used to calculate the vibrational Hamiltonian, with a Stark effect based method used to calculate the amide I site frequencies. The spectra were calculated using the NISE method.[49] The role of the amino acid preceding the proline-glycine turn in determining the turn structure was investigated. This XPG sequence is frequently found in both fibrous and elastomeric proteins and the nature of X has an influence on the structural and mechanical properties. Examples of proteins which contain this turn include collagen (X = proline), elastin (X = valine), and wheat glutanin (X = glutamine). In order to understand how the structure and folding of proline-glycine is modulated by the XPG turn, the effect of the size of the side-chain X was investigated. In this study, X was glycine, alanine or valine.

Varying the size of the side chain had a profound effect on the conformation of the peptide. Proline peak shifts and intensity changes were compared to a structure-based spectroscopic model. The tertiary structure of proline shifts the amide-I vibration of the preceding peptide bond to a lower frequency. The shift effectively decouples the proline vibration from those of its neighbours, moving it to a relatively sparse region of the spectrum. This means that proline amide-I spectroscopy is a useful probe of local secondary structure. Frequency shifts in the proline vibration is primarily the due to changes in the hydrogen bonding of the amide or other changes in the immediate electrostatic environment. The populations of type-II β turns, bulged turns, and irregular β turns for each peptide were assigned from the simulated spectra. The turn

commonly occurring in elastin (X = valine) was found to contain a 25% population of irregular β turns with two hydrogen bonds to the proline carbonyl.

The X point mutation allows one to tune the stability of these sorts of turns in engineered proteins and biopolymers. The study demonstrates that it is possible to measure the number of hydrogen bonds to proline using the frequency shift of the lowest-frequency amide-I vibration in the spectrum. This provides a method of probing the structure of samples rich in proline, such as elastin, which are not easily investigated by other methods. It was also established that it is possible to assign the composition of the hydrogen bonds of these turns through the use of 2DIR. As modelling methods improve, it should become possible to quantitatively assign the population of different conformations in the ensemble. This work provides a new approach for characterising the conformational states which exist in intrinsically disordered proteins and could be used to assess the quality of force fields used to model the structural and dynamical properties of small peptides.

2DIR spectroscopy was used to study the dynamics of hen egg white lysozyme (HEWL) that had been labelled using a ruthenium dicarbonyl.[69] The probe consisted of a ruthenium centre with two carbonyl ligands and two waters, and was coordinated to the His15 residue of HEWL. The labelled protein was investigated in both D_2O and a D_2O/glycerol mixture. The protein folds correctly and functions even in nearly anhydrous solutions of glycerol.[70] This allows for systematic changes to the bulk solvent viscosity while the protein remains structurally stable, which makes it a useful model system for studying site-specific protein-solvent coupling. The observed hydration dynamics around the hydrophobic metal-carbonyl resembles bulk-like dynamics. Spectral diffusion is a dynamic fluctuation of frequency, and occurs on time scales of around 1.5 ps. It is brought about by each molecule in a system having a different “instantaneous frequency” which evolves through time and leads to broadening of the spectral line. The surface labelling of the protein shows a

modest slowdown of approximately a factor of two between the bulk D_2O and the hydrating D_2O. This quantitatively agrees with predictions taken from MD simulations, and is attributed to the solvation dynamics decelerating due to the hydrophobic surface of the protein impeding the switching of hydrogen-bonds. As glycerol is added to the system, viscosity increases by a factor of 100, but this causes only a three-fold slowing of the hydration dynamics. Accompanying this is a complementary slowdown of the protein dynamics. These results show that the coupling between hydration water and bulk solvent is weak, but that the coupling between protein and water dynamics is strong.

The region at the interface of two helical domains in HEWL is susceptible to solvent slaving. A slaved process is tightly coupled to the solvent, with their rates showing approximately the same temperature dependence as the rate of solvent fluctuation though they are smaller, whereas the rate of a non-slaved process will be determined by protein conformation and vibrational dynamics. The water in the hydration shell is also slaved to the bulk solution. This combination of surface-labelling with 2DIR, especially in conjunction with site-directed mutagenesis, provides a means to investigate in detail the dynamics of protein-water interfaces. Clearly, analysis will have to account for any structural changes caused by the mutations and surface labelling procedure. This procedure is suitable for describing a number of fundamental interactions including protein-protein, protein-water and protein-water-lipid, due to the sensitivity of vibrational chromophores to solvent environments.

In summary, 2DIR is a useful tool for investigating peptides and small proteins. These systems are often used for benchmarking purposes, testing new methods and as models for larger systems. 2DIR, in combination with other methods, such as MD simulation and isotope or vibrational labelling, has produced work of potentially enormous relevance to current problems in the biological sciences. It has been used to test nitrile vibrational probes and confirmed that the substitution of phenylalanine for cyanophenylalanine should

have only a minor impact on the structure of large proteins. It played a role in the development of a new method for probing hydrogen bonding in proline-rich proteins. 2DIR has also been used to follow the changing structure in a complete folding event of an α-helix.

2.2.3 Membrane Proteins

Transport proteins move ions and small molecules across the membrane and there are a large number of sub-categories for this class of membrane protein. Proton transport is important in a diverse range of biological processes, including photosynthesis, enzyme catalysis and acid-base neutralization. Proton conduction may occur via the formation of hydronium ions. This allows an 'excess' proton to utilise the hydrogen bond network of water molecules within the channel to traverse the cellular membrane. Transfer of the proton via formation of a hydronium ion is possible due to the constricted conditions within the protein hydrophilic cavity, which transforms the water hydrogen bond network to a wire-like structure, as opposed to the structure seen in bulk water. The underlying assumption in the study of proton transport by 2DIR is that the dynamics of the protein channel results in fluctuations of the amide I vibrational frequency, and are also affected by the transfer of the proton.[71,72]

Jansen and colleagues investigated what determines the line shape of the amide I band, proton transport channels and the extent to which the proton transport itself can be probed by 2DIR.[23] They studied the gramicidin A peptide, a well characterised prototypical proton channel. Gramicidin A has antibiotic activity, which arises from the increased permeability of inorganic monovalent cations through bacterial cell membranes. This destroys the proton/ion gradients between the intracellular and extracellular environments. The transport of protons in the channel may happen via one of two mechanisms: proton hopping or water reorientation. MD simulations and 2DIR calculations were used to study proton transport through the channel. The OLPS-AA force

field was used along with a Stark effect method to calculate the amide I site frequencies. The 2DIR spectra were calculated from the Hamiltonian using the NISE scheme.[49] Rates in the centre of the channel were approximately five times higher than near the channel entrance. The work revealed that proton transport does modify the dynamics of the system, which can be followed by monitoring the time evolution of the anti-diagonal line width and slope of the spectrum. In the measured isotope labelled 2DIR spectrum the position dependent proton transport rates are smeared out due to the mechanisms at work, including channel water rotation and proton hopping.

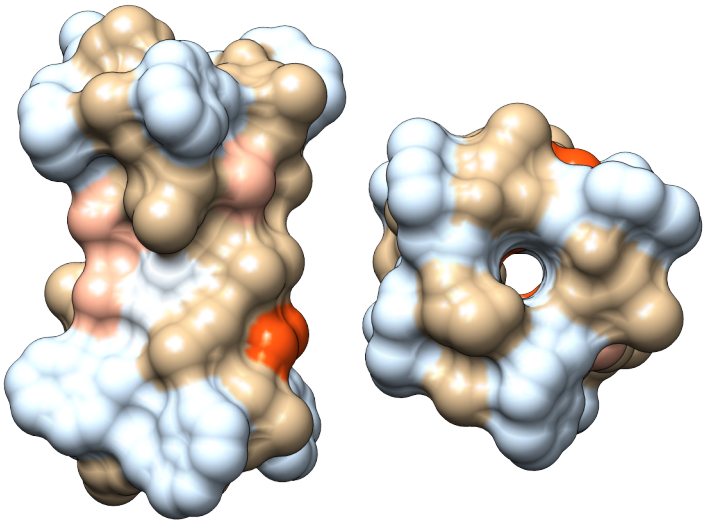


Figure 2.9: Hydrophobicity surface of proton channel Gramicidin A (PDB code 1GRM), showing the side of the protein and the channel itself. Rendered with Chimera.[50]

The local amide I site frequencies are sensitive to proton hopping within about 10 Å. The simulations indicated that the rate limiting step appears to be a proton entering the channel. The proton easily escapes from the channel back to bulk, but has low probability of entering the channel from the bulk region. Thus, only at low pH will a proton remain inside the channel for a significant amount of time. Therefore, when probing proton transport in membrane proteins with 2DIR will be sensitive only at low pH, i.e. when the concentration of protons is high. Proton entry into the channel appears to be a significant bottleneck in proton transport, and is an interesting area for future research. Other systems

involving proton transport, such as enzyme families, condensed phase crystals, carbon nanotubes, and fuel cell electrolyte membranes, are also being studied. The combination of MD simulations and 2DIR spectrum modelling presents a promising tool for studying proton transport in these systems.

Work by Knoester and colleagues showed that measuring the line width of the spectrum as a function of waiting time could be used as an experimental tool to determine the intermolecular proton transport rate in biological channels.[72] They found that the rate of the time evolution of the anti-diagonal line width is not equal to the proton hopping rate. Instead it is determined by a complex interplay between proton hopping rate, the water reorientation rate and the water diffusion rate. Comparison with theory is, therefore, required to extract the local proton hopping rate. Their model showed a relationship between spectral behaviour and the actual proton transport rate, which should be verifiable by more advanced proton transport model based methods. Their method only strictly applies to residue four of gramicidin A, for which it was parametrised, though it is expected that the major contributions to frequency fluctuations arise from electric fields generated by the proton and the water and contributions from the specific site should be negligible. The method should be applicable to other proteins involved in proton transport.

Tokmakoff and colleagues have looked at membrane proteins that are involved in the transport of ions instead of protons.[24] Recent experiments on photosynthetic centres[73] show non-classical phenomena, such as coherent excitation transfer, may occur in protein complexes on certain temporal, spatial and energetic scales. Although a great deal of theoretical and experimental work has been carried out, especially on the photosynthetic system, there remain questions as to whether the observed phenomena occur in other systems. Specifically, is quantum coherence or tunnelling limited to processes involving electrons or are they also involved in much larger particles such as ions? The observation of quantum effects could shed new light on the fundamental

principles governing the function of these systems.

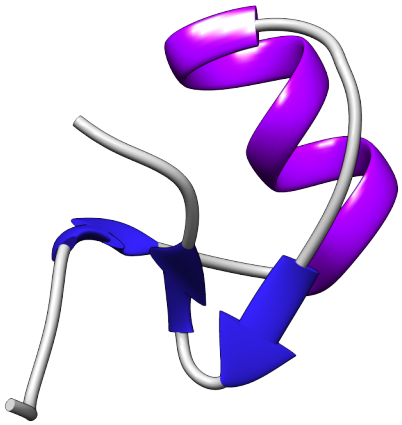


Figure 2.10: Structure of the potassium ion channel (PDB code 1PVZ). Rendered with Chimera.[50]

The potassium ion channel is one of the best studied classes of ion channels. Crystallography and molecular modelling have posited several models for the selectivity and high transport rate. The currently favoured model of ion selectivity is the 'snug-fit' model.[74] MD simulations were carried out on the channel protein and the resulting trajectories used to calculate the 2DIR spectra.[71] At the time, this was the largest protein for which simulated 2DIR spectra had been calculated. Preliminary work on nonactin and valinomycin in conjunction with potassium showed spectral line narrowing due to increased rigidity from ion coordination. This same line narrowing is observed for the ion channel itself when the channel contains potassium. The intensity changes may be sufficient to differentiate potassium binding. Alongside the rigidity induced by potassium binding was a shift in the frequencies of oscillators coordinating potassium. In the native channel protein, this shift is difficult to detect. Isotope labelling can highlight these changes and produce difference signals to identify the switching. This work concluded that the electrostatic changes are the primary cause of the observed fluctuations in the vibrational spectrum. Local changes such as carbonyl group rotations contribute less to the observed

differences in the spectrum than the electrostatic changes caused by the movement of the potassium ion.

2.2.4 Amyloid Fibrils

Amyloid fibrils are a group of insoluble fibrous aggregates that share similar structural traits. They result from misfolding events in a particular protein's folding pathway. The deposition of these insoluble aggregates as plaques is thought to lead to cell death and tissue degeneration. Islet amyloid polypeptide (IAPP) is a 37 residue peptide, whose misfolding results in fibrils associated with the dysfunction of β cells found in type 2 diabetes. Fibrils formed from other proteins are associated with neurodegenerative disorders, including Alzheimer's disease, Parkinson's, Huntington's disease and the transmissible spongiform encephalopathies. The fibrils form a cross β structure (figure 2.11), where the β-strands are perpendicular to the fibril axis with hydrogen bonds between strands in the axis of the fibril.[75] The fibrils can form different morphologies, depending on experimental conditions, which mostly differ in their overall symmetry and quaternary structure. The application of 2DIR may help elucidate structures along the aggregation pathways. Understanding these folding pathways is essential in order to develop therapeutics targeting the early stages of the diseases resulting from the formation of amyloid fibrils. 2DIR can also give an insight into the mechanism of existing fibril inhibitors, allowing for refinement and improvement for future design studies.

Work has been carried out on a two-fold symmetry structure associated with untwisted "striated ribbon" type configuration, though a "twisted pair" structure with three-fold symmetry also exists. Mukamel et al. built two systems based on the two-fold symmetry configuration.[25] The first (S1) incorporated protonated aspartic and glutamic acid residues. The second (S2) was neutral and proved to be stable over 25 ns in MD simulations. The

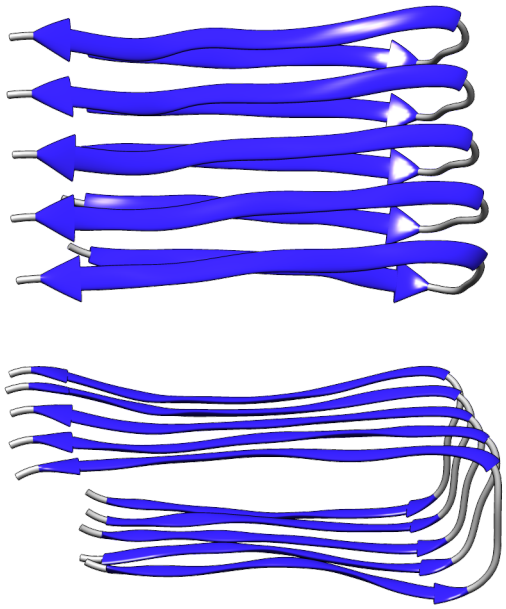


Figure 2.11: Cross-β structure of the Aβ (1-42) fibril (PDB code 2BEG), which is involved in Alzheimers disease. Rendered with Chimera.[50]

Hamiltonian was computed from these trajectories using the Cho frequency map[76,77] and the transition dipole coupling model.[78] The signal was calculated using the non-linear exciton propagation formalisation.[58] The frequency distributions were decomposed into the three major contributions, which were from water, the backbone and side chains. The observed vibrational frequencies are dominated by side chain contributions and these play a major role in the inhomogeneity of the spectral line width. A set of experimentally measured 2DIR semi-diagonal traces for 18 isotopomers were compared to the simulated frequency distributions. The simulations for S2 showed good agreement with experimental data for several residues.

The Aβ monomer is the peptide building block of the fibrils that occur in Alzheimer's disease. Work has looked at the Aβ42 monomer, which is a mixture of conformationally ordered and disordered species. These can be classified based on a characteristic β-hairpin in a particular region. MD simulations in conjunction with simulation of 2DIR spectra have shed light on

the structure and kinetics of these species.[25] A new method[79] has been developed with the addition of chirality sensitive techniques, which shows much promise in the structure elucidation of these monomers, since conventional order parameters (e.g. backbone dihedrals) cannot discriminate between the two conformers of Aβ42. The chirality induced 2DIR method showed a well-resolved cross peak associated with the dominant conformer of Aβ42, indicating potential for the identification of early misfolding events of the conformer.

Currently, most inhibitors of protein aggregation are small molecules or peptides. The two main features of these inhibitors are high sequence similarity with one region of the target protein, which enhances binding, and secondly a mutation to prevent or destabilise the fibril formation. Mutations may be naturally occurring amino acids such as proline which disrupts the formation of β-sheets[80] or unnatural amino acids. Rat amylin has features in common with known inhibitors of amyloid fibril formation. It does not form fibrils and rats do not develop type-2 diabetes; yet rat amylin is only a moderately effective inhibitor of fibril formation in humans. The lack of detailed structural information of inhibitors, like rat amylin (rIAPP), means the variable efficacy in humans is not fully understood. Middleton and co-workers[81] used 2DIR to investigate the behaviour of rat amylin in complex with human amylin. By labelling backbone carbonyls with $^{13}C^{18}O$ isotope, they created a linear column of vibrationally coupled residues within the fibril. The frequency of such a column of coupled oscillators is shifted from the fundamental vibrational frequency by twice the coupling strength.[82,83] Based on the structure of rat amylin, it might be anticipated that fibril inhibition occurs as a result of blocking the formation of the C-terminus β-sheet. However, it was observed that eight hours after mixing human and rat amylin, the N-terminus β-sheet is prevented from forming. 24 hours after mixing neither C- nor N- terminus β-sheets had formed, the rat amylin instead forming its own β-sheet. This

difference in structure of the rat amylin after 24 hours could explain why it is only a moderate inhibitor. The proline residues present in the rat amylin do not prevent the human peptide from forming β-sheets. In fact, the three prolines prevent formation of strong interactions at the C-terminus. This might be why rat amylin does not inhibit fibril formation as effectively as related peptides with a single proline.

There is increasing evidence that intermediate structures might be the most cytotoxic entities in the folding pathway, though the precise nature of these species is contentious. 2DIR has furnished new insights into the aggregation progress both in the presence and absence of lipid membranes. MD simulations were used in conjunction with experimental spectroscopy to characterise the amyloid fibrils.[84] The MD simulations were carried out on two two-fold symmetry models of an amyloid fibril. The protofilament of these models comprises four parallel β-strand layers with two symmetric columns of hIAPP. The primary difference between the two models is in the side chain arrangement: some residues, which are buried in one model, are on the surface in the other model. The simulations indicated that model I was more stable than model II. The 2DIR spectra were calculated from the MD trajectory using the frequency map developed by Wang et al.[84] and a nearest neighbour frequency shift to account for the through-bond effects. Tycko and colleagues[85] constructed models for hIAPP fibrils using data from electron microscopy, atomic force microscopy and solid-state NMR. These models were refined using work by Zanni and co-workers on the frequency maps for protein backbones and side chains.[84] The spectra calculated from the MD trajectories showed good agreement with experimental spectra and validates both of these models.

The position of the ω_3 probe frequency axis of the fundamental frequency and the diagonal line width were used to quantify the spectra for each individual chromophore. Plotting the diagonal line width against residue number gives a W pattern, with an inhomogeneous distribution of frequencies. Large values for

the diagonal line width indicate large structural fluctuations and this is seen at the termini and the turn regions. Conversely, small values for the diagonal line width indicate more stable and rigid structures and this is reflected in the data for the β-sheet regions and is consistent with previous structural analysis. The stability of the β-sheet indicates a potential target in designing drugs to inhibit aggregation of the fibrils. Model I was used to construct two additional morphologies, which were based on the structures of Aβ fibril. The spectra for all three models show a W pattern (figure 2.12), in accord with experimental findings, and perhaps indicative that different morphologies of fibrils may show similar structural behaviour, especially with respect to the N-terminus and the turn region. The C-terminal region is the main site of polymorphism in this case.

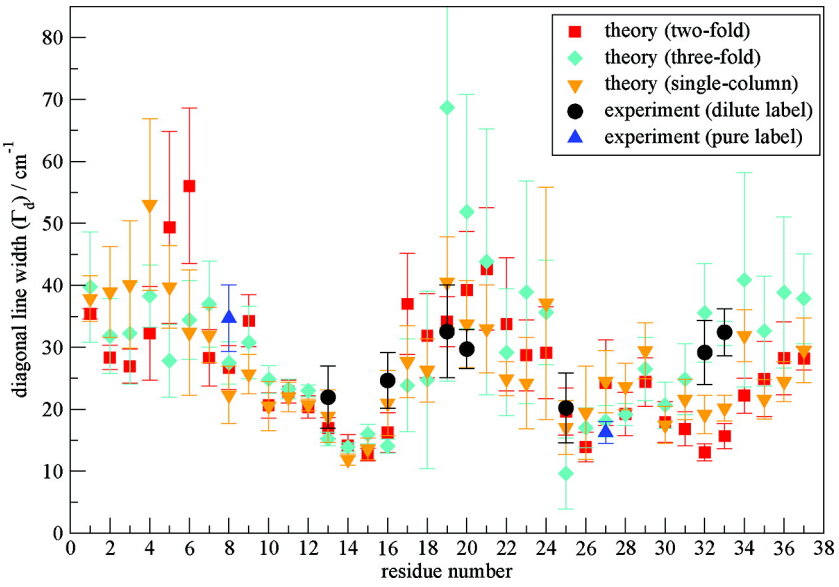


Figure 2.12: Plot showing diagonal line width versus residue number, illustrating the "w" pattern. Reprinted with permission from Wang *et al.* 2011.[86]

Zanni and colleagues utilised 2DIR and several other techniques (including isotope labelling, enzymatic digestion and mass spectrometry) to analyse the structure and identify the sequence of the core region of human γD-crystallin amyloid fibril.[87] The aggregation of lens crystallin in the eye is the cause of cataracts, which affect more than half of the population over 55. Currently, the aggregation process in cataracts is not well understood. Each of the domains of

γD-crystallin was ^{13}C labelled using expressed protein ligation in order to resolve them in the 2DIR spectra. From this it was possible to probe independently their structural features in the both fibril and the soluble native state. Amyloid β-sheet structure formed in the C-terminus domain, alongside some disordered structures, while the N-terminus domain became entirely disordered. The β-sheet region is smaller than that predicted by mass spectrometry, which suggested that the N-terminus domain contributed to core instead of being disordered, as was seen in the 2DIR. It is likely that the core forms from a β-superpleated structure. Also indicated were loop and helical structures that may be useful in the design of aggregation inhibitors. The results do not define the secondary structure of individual residues, but could guide further labelling schemes that could identify details such as the number and length of β-strands, α-helices and loops in the fibrils. It could be possible to label short segments or individual residues in order to identify the locations of loops and short helices by taking advantage of the extended vibrational coupling frequency and line shapes in 2DIR. A distinct advantage of this method lies in the fact that detailed structural information about aggregates is available with samples as small as 1 μg. This means that samples that are hard to prepare could be studied using this method under many different experimental conditions.

2DIR has great potential in the study of amyloid fibrils. It can clarify information on the folding pathways of fibrils, by elucidating the structure of intermediate species. 2DIR has provided insight, not only into how these fibrils are formed, but also into which structures are most toxic and under what conditions. Elucidation of these folding pathways can provide potential targets for developing drugs which prevent the formation or toxicity of these fibrils. Also, 2DIR can shed light on currently available therapeutics, including why some drugs may not be as effective as expected. This knowledge will inform future drug design and allow researchers to develop more effective therapeutics

in the future.

2.3 Concluding Remarks

2DIR has wide ranging uses and implications, particularly in the field of biomacromolecules. It builds upon the established IR methods, but provides more detail about the three-dimensional structure and dynamics of molecules. 2DIR spectra can be obtained theoretically and experimentally, and both approaches are essential to maximise the information available from this spectroscopy. Experimentally, 2DIR can be combined with isotope or chemical labelling to probe specific sites and yield a great deal of structural information.

2DIR is particularly useful in the study of peptides and proteins, where the congested amide I band of linear vibrational spectroscopies makes it difficult to make accurate assignments. By expanding the spectra across another frequency dimension, cross peaks become visible and these yield more detailed information about the structure and dynamics of the system. Theoretical calculations are an important tool that can complement experimental 2DIR investigations by aiding the interpretation of the spectrum to support structure determination and dynamic effects. In conjunction with MD simulation, 2DIR can give access to systems that are difficult to study experimentally, but also allows the refinement of the models used to assign peaks. Experimental equipment such as lasers and detectors are getting simpler, and several manufacturers are beginning to offer dedicated stand alone systems for 2DIR, which will increase the number of studies carried out. As a result of this, the continued development of modelling tools is paramount to ensure that theory continues to support the interpretation of data obtained experimentally.

2DIR has been in wide use for little over a decade and it has already made an impact on some fundamental questions in the biological sciences, as well as being applicable elsewhere. It complements the linear spectroscopies that are well established and given that it provides more unique spectra, and greater

spectroscopic detail, multidimensional spectroscopies like 2DIR are indispensable in the study of systems where traditional spectroscopies are limited or cannot be used. Theoretical studies and computed spectra greatly enhance the interpretation of 2DIR spectra resulting in a richer understanding of the ultrafast dynamics of proteins, particularly in solution phase.

Chapter 3

Theoretical Methods

3.1 Molecular Dynamics Simulations

Molecular dynamics simulations (MD) are a method used to propagate a system of particles—in this thesis, atoms or molecules—through time via the numerical integration of Newton's equations of motion. The results of this propagation, a trajectory, represents a sample of the conformational ensemble, which can be used to calculate numerous properties and give insight into the behaviour of the system.

3.1.1 Force Field Methods

In force field or molecular mechanics methods such as (classical) MD, the potential energy surface is calculated as independent of the nuclear coordinates. In the simulation, the energy is calculated using an empirical force field, which has been fitted to experimental data. In the case of *ab initio* force fields the energy is fitted to high level computational data. In force field methods bonds are often treated harmonically, i.e. using a ball on a spring model. The movement of the electrons is assumed to be a function of the nuclear motion (Born-Oppenheimer approximation). Since solving the electronic Schrödinger equation is bypassed using force field methods, the bonding information must

be given explicitly.

Molecular mechanics methods work because of the observation that different groups behave similarly regardless of the molecule they are part of. For example, a C=O bond will have a bond length of approximately 1.22 Å and will vibrate with a frequency in the region of 1700 cm^{-1}. There will be small differences in bond length and frequency between a C=O bond in a ketone and C=O bond in an amide bond, but C=O bonds in amide groups will always be similar. The differences between C=O bonds in different environments are small enough that molecules of great complexity can be built from these functional group building blocks.

The force field energy is given as a sum of terms, each term being the energy required to distort the molecule in a specific fashion.

$$E_{FF} = E_{str} + E_{bend} + E_{tors} + E_{vdw} + E_{el} + E_{cross} \tag{3.1}$$

where E_{str} is the energy required to stretch a bond between two atoms, E_{bend} is the energy required to bend an angle, E_{tors} is the torsional energy of rotation around a bond. E_{vdw} and E_{el} are non-bonded interactions and describe the van der Waals and electrostatic energy respectively. E_{cross} describes the coupling between the bonded interactions.

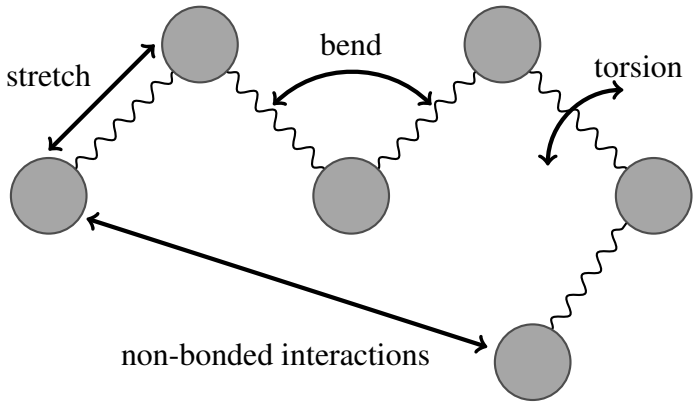


Figure 3.1: Various terms included in a force field.

E_{str} is the energy required to stretch or compress a bond from its "natural" or equilibrium bond length. The simplest method of calculating bond stretch

energy is a Taylor expansion around a "natural" bond length that is terminated at the second order, giving a harmonic approximation.

$$E(r) = \frac{1}{2}k(r - r_0)^2 \tag{3.2}$$

where r_0 is the natural or equilibrium bond length, r is the distance between the atoms and k is the force constant. The accuracy of this harmonic approach can be increased by using higher order terms but this has the disadvantage of being more computationally expensive. The harmonic potential is quite accurate close to the equilibrium bond length but increasingly inaccurate further away from this point. Hence it is sometimes more appropriate to use the Morse potential to calculate the bond stretch energy.

$$\begin{aligned} E(r) &= D_e(1 - e^{-\alpha(r-r_0)})^2 \\ \alpha &= \sqrt{\frac{k_e}{2D_e}} \end{aligned} \tag{3.3}$$

where D_e is the well depth, r is the distance between the atoms and r_0 is the equilibrium bond length. α is related to k_e which is the force constant at the minimum of the potential well. The Morse potential is compared to the harmonic potential in figure 3.2.

The Morse potential is more accurate than the harmonic potential but it is not without its limitations. For long bond lengths the restoring force is small, which results in distorted structures displaying slow convergence towards equilibrium bond length.

E_{bend} is the energy required for bending an angle formed by three atoms A-B-C in which A and B and B and C share a bond. Like E_{str}, E_{bend} can be found using a harmonic approximation.

$$E(\theta) = \frac{1}{2}k(\theta - \theta_0)^2 \tag{3.4}$$

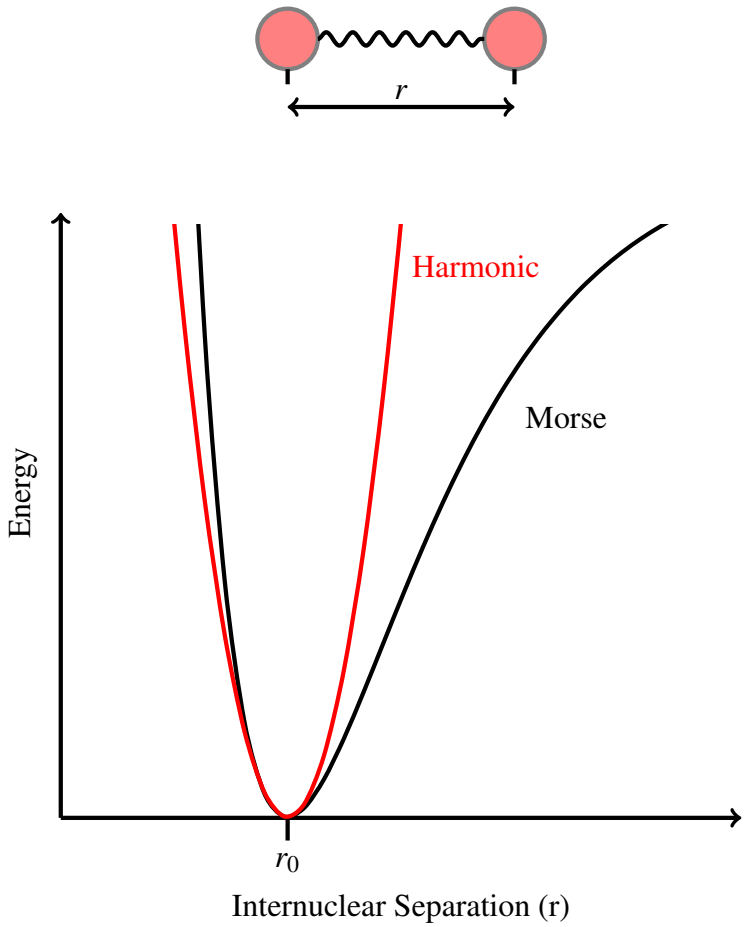


Figure 3.2: Illustration of the harmonic and Morse potentials used for bond energies.

E_{tors} is the energy change associated with the rotation around a B-C bond in a four atom sequence A-B-C-D where A is bonded to B, B to C and C to D. If one looks down the length of the B-C bond then a torsional angle is defined as the angle (ω) between the A-B and the C-D bonds as shown in figure 3.3. The angle ω is often expressed as a range from -180° to 180°, encompassing a full rotation of the bond.

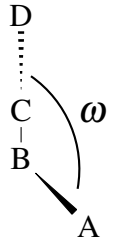


Figure 3.3: Definition of torsion angle.

The torsion energy differs from the stretch and bend energies in a number of ways. The energetic penalty for distorting a molecule away from minimum structure is quite low for bond rotations, meaning that large deviations from the minimum energy structure may occur. The barrier for bond rotations has contributions from non-bonded interactions e.g. van der Waals, hence the torsion energy is also coupled to the non-bonded parameters. The torsion energy must also be periodic i.e. if a bond is rotated 360° the energy will return to the same value. Because of this periodicity, E_{tors} is often expressed as a Fourier series:

$$E_{tors}(\omega) = \sum_{(n=1)} V_n \cos(n\omega) \tag{3.5}$$

where V_n determines the size of the rotational barrier around B-C. The $n = 1$ term describes a rotation about a bond that is periodic by 360°, $n = 2$ a rotation that is periodic by 180°, $n = 3$ by 120° etc. n controls the number of minima in the rotational energy.

The non-bonded interactions required for calculating E_{FF} are the van der Waals and electrostatic energies. The van der Waals term includes the energy between two permanent dipoles, the force between an induced dipole and a permanent dipole and the interaction between two induced dipoles. E_{vdw} describes the repulsion or attractions between atoms that are not directly bonded and not due to the electrostatics. E_{el} describes the energy of interaction between fixed partial charges within the molecule.

E_{vdw} is zero for large inter-atomic distances, rapidly becoming very repulsive for short distances. This repulsion can be attributed to the overlap of the electron clouds of the two atoms which occurs at short internuclear distances. The van der Waals interaction is very slightly attractive at intermediate distances due to induced dipole-dipole interactions.

A popular method for incorporating the van der Waals energy is by means of a Lennard-Jones potential which can be given as:

$$E_{LJ}(R) = \varepsilon \left[\left(\frac{R_0}{R} \right)^{12} - 2 \left(\frac{R_0}{R} \right)^{6} \right] \tag{3.6}$$

where R_0 is the minimum energy distance, R is the internuclear separation and ε is the depth of the minimum.

The electrostatic energy is given by the Coulomb potential:

$$E_{el}(r_{ij}) = \frac{Q_i Q_j}{\varepsilon r_{ij}} \tag{3.7}$$

where ε is a dielectric constant, Q is the charge of atom A or B and R_{AB} is the distance between them. The Coulomb potential can be expanded to take into account dipoles, multipoles and polarizabilities.

The final term in the force field energy is a cross term that takes into account the coupling between the other, fundamental terms. These cross terms are usually written as a Taylor expansion of the individual coordinates, but are not present in every force field. There are a number of packages available for carrying out MD simulations including, but not limited to, CHARMM,[88,89] NAMD[90] and GROMACS[91,92] and many of these have their own forcefields, for example, in the case of the CHARMM forcefield[93,94] and the GROMOS forcefield.[95]

The combination of these energy terms allows for the construction of the Potential Energy Surface which effectively gives the energy of a molecule as a function of its geometry. As an example, the CHARMM forcefield[89,93,94] is given as:

$$
\begin{aligned}
E = & \sum_{bonds} K_b(b-b_0)^2 + \sum_{UB} K_{UB}(S-S_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\
& + \sum_{dihedrals} K_\chi(1+\cos(n\chi-\delta)) + \sum_{impropers} K_\phi(\phi-\phi_0)^2 \\
& + \sum_{nonbond} \left\{ \varepsilon_{ij} \left[\left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}} \right\} \\
& + \sum_{residues} V_{cmap}(\phi, \psi)
\end{aligned}
\tag{3.8}
$$

where the bond stretch and bending terms are given harmonically, similar to equations 3.2 and 3.4. K_b and K_θ are the bond force constant and bond angle constant respectively and are taken from the forcefield. The improper angle term is also given harmonically with ω being the improper angle and K_ω being the force constant of the improper. The non-bonded terms are given via a Lennard-Jones and Coulombic potentials as in equations 3.6 and 3.7. The dihedral term is sinusoidal term in which n is the periodicity of the dihedral angle and δ is the phase shift. The Urey-Bradley (UB) term defines interactions between 1,3 pairs and is a quadratic function of the distance, S, between atoms one and three. It is only very rarely used within the CHARMM forcefield. The CMAP term[96,97] is a backbone correction term for proteins based on *ab initio* quantum mechanical (QM) calculations. The CMAP term greatly improves structural and dynamical results obtained from crystalline or solution phase simulations.

3.1.2 Energy Minimisation

Stable geometries correspond to minima on the potential energy surface and can be identified by minimising the force field energy E_{FF} as a function of

the nuclear coordinates. Energy minimisation thus becomes a problem of finding minima on the potential energy surface i.e. finding stationary points (where the first derivative—the gradient—is zero) of the energy where the second derivatives have positive values. There are a number of different methods for energy minimisation, each with their attendant advantages and disadvantages.

The steepest descent method utilises a series of function evaluations in the direction where the gradient, $\vec{g}$, is negative, $\vec{d} = -\vec{g}$. This means that steepest descent will always decrease the value of the gradient and is guaranteed to approach a minimum. When the gradient begins to increase again the minimum can be approximated by interpolating between the two calculated points. Once a minimum is located a new gradient is calculated perpendicular to the previous one and a new line search is carried out. Since the new search gradient is necessarily perpendicular to the previous one, gradient components in that direction will be missed. Steepest descents therefore has a tendency to spoil the previous search with each new iteration. Additionally this algorithm will never locate a true minimum, but simply oscillate around it at an ever decreasing rate.

The Conjugate Gradient method attempts to improve on the steepest descent method, specifically the partial undoing of the previous search step. It does this by performing each new line search not along the current gradient, nor perpendicular to it, but the line is constructed such that it is "conjugate" to it. The first step is equivalent to the steepest descent method but each successive step is performed along a line that is a mixture of the current negative gradient and the previous search direction. The conjugate gradient algorithm is given via:

$$\vec{d}_i = -\vec{g}_i + \beta_i \vec{d}_{i-1} \tag{3.9}$$

where $-\vec{g}$ is the negative gradient, $\vec{d}_{i-1}$ is the previous search direction and β is the conjugate property. There are a number of different methods for setting the β value. For example, in the Fletcher-Reeves method β is given by:

$$\beta_i^{FR} = \frac{\vec{g}_i^t \vec{g}_i}{\vec{g}_{i-1}^t \vec{g}_{i-1}} \tag{3.10}$$

It is not uncommon for the conjugate gradients algorithm to need to be reset i.e. β set to 0 during the course of an energy minimisation. Multiple methods of minimisation may be used together. This can reduce the computational cost of the minimisation step by using a more coarse grain method to get close to the minimum then a more expensive method to find the minimum itself.

The Newton-Raphson method is a second order method where the true function is expanded around the current point. This is given by

$$f(\vec{x}) = f(\vec{x}_0) + \vec{g}^t(\vec{x} - \vec{x}_0) + \frac{1}{2}(\vec{x} - \vec{x}_0)^t \vec{H}(\vec{x} - \vec{x}_0) \tag{3.11}$$

where $\vec{x}_0$ is the current point, $\vec{g}$ is the gradient and $\vec{H}$ is the Hessian matrix. In order for the energy surface to be at a minimum, the the second derivative of the gradient must be zero. This leads to the following step:

$$(\vec{x} - \vec{x}_0) = -\vec{H}^{-1}\vec{g} \tag{3.12}$$

where $\vec{H}^{-1}$ is the inverse of the Hessian matrix. In the case that the Hessian is a diagonal matrix, $\vec{x}$ is denoted $\vec{x}'$ and the Newton-Raphson step can be written as

$$\begin{aligned} \Delta\vec{x}' &= (\Delta x'_1, \Delta x'_2, \Delta x'_3, \ldots, \Delta x'_N)^t \\ \Delta x'_i &= -\frac{f_i}{\varepsilon_i} \end{aligned} \tag{3.13}$$

where f_i is the projection of the gradient along the Hessian eigenvector. It's eigenvalue is ε_i, the gradient component pointing in the direction of the ith eigenvector.

In general, the Newton-Raphson method will attempt to converge on the nearest stationary point on the potential energy surface, regardless of whether that is a minimum, a maximum or a saddle point. Using a straightforward

Newton-Raphson method can be computationally demanding owing the the need to calculate and manipulate the Hessian matrix. This can be alleviated somewhat by mixing Newton-Raphson methods with other minimization algorithms.

The Adopted Basis Newton Raphson (ABNR) method is a widely used algorithm that was first implemented in CHARMM.[88,89] It combines the steepest descent method with the Newton-Raphson scheme. In ABNR, the Newton-Raphson method is applied only to a small subspace of the molecule undergoing minimization. The overall displacement of geometry is therefore mostly steepest descent (SD) with a small contribution from Newton-Raphson (NR):

$$\Delta q_k = \Delta q_k^{SD} + \Delta q_k^{NR} \tag{3.14}$$

where Δq_k is the displacement of the geometry at iteration k. At the beginning of the minimization only the steepest descent part is carried out, meaning that initially $\Delta q_k^{NR} = 0$. After several steepest descent steps, the last m geometry displacements can be used as a basis to obtain the Newton-Raphson step.

3.1.3 Dynamics Propagation and Newton's Equations of Motion

Once the initial positions and velocities of the particles in the system are known—or can be assigned—the dynamics of the system can be propagated. Dynamics are usually propagated once the system has reached an energetic minimum, though this is not strictly necessary. Dynamics are carried out using classical mechanics i.e. Newton's laws of motion, specifically Newton's second equation $\vec{F} = m\vec{a}$ which in its differential form is

$$-\frac{\mathrm{d}E}{\mathrm{d}\vec{r}} = \vec{m}\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2} \tag{3.15}$$

where E is the potential energy at position $\vec{r}$. The vector quantity $\vec{r}$ contains the coordinates for all the particles in the system. The left hand side of the equation is the negative of the energy gradient or the force on the particles such that

$$F = -\frac{\mathrm{d}E}{\mathrm{d}\vec{r}} \tag{3.16}$$

The position of the atoms at time, Δt, after the current time, $t = 0$ can be found via the Taylor expansion

$$\begin{aligned} \vec{r}_{i+1} &= \vec{r}_i + \frac{\mathrm{d}\vec{r}}{\mathrm{d}t}(\Delta t) + \frac{1}{2}\frac{\mathrm{d}^2\vec{r}}{\mathrm{d}t^2}(\Delta t)^2 + \frac{1}{6}\frac{\mathrm{d}^3\vec{r}}{\mathrm{d}t^3}(\Delta t)^3 + \ldots \\ \vec{r}_{i+1} &= \vec{r}_i + \vec{v}_i(\Delta t) + \frac{1}{2}\vec{a}_i(\Delta t)^2 + \frac{1}{6}\vec{b}_i(\Delta t)^3 + \ldots \end{aligned} \tag{3.17}$$

where the velocities, $\vec{v}$ are the first derivatives of the positions with respect to time $(\mathrm{d}\vec{r}/\,\mathrm{d}t)$ at time t_i, the accelerations, $\vec{a}$, are the second derivatives $(\mathrm{d}^2\vec{r}/\,\mathrm{d}t^2)$ at time t_i, the hyper-accelerations, $\vec{b}_i$ are the third derivatives $(\mathrm{d}^3\vec{r}/\,\mathrm{d}t^3)$ etc. The positions of the particles at previous time step Δt_{i-1} can be found via

$$\vec{r}_{i-1} = \vec{r}_i - \vec{v}_i(\Delta t) + \frac{1}{2}\vec{a}_i(\Delta t)^2 - \frac{1}{6}\vec{b}_i(\Delta t)^3 + \ldots \tag{3.18}$$

The positions of the particles at time step later, Δt_{i+1}, can be found from the current positions, the previous positions and the current acceleration. This gives the Verlet algorithm[98] for solving Newton's second equation numerically. The Verlet algorithm can be expressed as

$$\begin{aligned} \vec{r}_{i+1} &= (2\vec{r}_i - \vec{r}_{i-1}) + \vec{a}_i(\Delta t)^2 + \ldots \\ \vec{a}_i &= \frac{\vec{F}_i}{\vec{m}_i} = -\frac{1}{\vec{m}_i}\frac{\mathrm{d}E}{\mathrm{d}\vec{r}_i} \end{aligned} \tag{3.19}$$

At the starting point the positions at the previous time step are not available but can be approximated as

$$\vec{r}_{-1} = \vec{r}_0 - \vec{v}_0 \Delta t \tag{3.20}$$

In the Verlet algorithm the acceleration must be re-evaluated from the forces at each time step. One disadvantage of this algorithm is that the velocities do not appear explicitly, which can cause problems in generating ensembles with constant temperature as velocity is related to temperature via the kinetic energy of the system

$$E_{kinetic} = \frac{3}{2} kT = \sum_i \frac{1}{2} m \vec{v}^2 \tag{3.21}$$

where T and $\vec{v}$ are the temperature and velocity respectively, m is the mass of a given particle and k is the Boltzmann constant.

The problem of the velocities not appearing explicitly is solved in the leap-frog algorithm in which the positions at time, Δt_{i+1} can be found by

$$\vec{r}_{i+1} = \vec{r}_{i+\frac{1}{2}} \Delta t \tag{3.22}$$

with the velocities being obtained via

$$\vec{v}_{i+1} = \vec{v}_{i-\frac{1}{2}} + \vec{a}_i \Delta t \tag{3.23}$$

Since the velocities do appear directly in this algorithm the system can be coupled to an external heat bath for constant temperature simulation. However there is the disadvantage that since the positions and velocities are out of phase by half a step, they cannot be known at the same time. The velocity Verlet algorithm solves both issues, in that the velocities appear directly and can be known at the same time as the positions. The velocity Verlet algorithm[99] is given by

$$\begin{aligned} \vec{r}_{i+1} &= \vec{r}_i + \vec{v}_i \Delta t + \frac{1}{2}\vec{a}_i \Delta t^2 \\ \vec{v}_{i+1} &= \vec{v}_i + \frac{1}{2}\{\vec{a}_i + \vec{a}_{i+1}\}\Delta t \end{aligned} \tag{3.24}$$

Both the Verlet and leap-frog type algorithms are fully time-reversible i.e. it is not only possible to find the position of the atoms at time Δt_{i+1} but also at time Δt_{i-1} as well.

An important parameter to consider is the time step, Δt, taken in the simulation. The maximum time step that can be taken is determined by the fastest processes that occur in the system. For peptide and water systems this is typically the heavy atom-hydrogen bond motions which occur on the order of 10^{-11}–10^{-14} seconds. The time step taken is typically an order of magnitude smaller than the fastest motions which requires Δt to be on the order of a femtosecond (1×10^{-15} s). This is a significant problem, particularly in the study of proteins as most interesting behaviour occurs on the scale of nanoseconds or longer, requiring tens of millions of time steps. Even for relatively small systems this is a significant computational effort.

Fortunately, the hydrogen stretching vibrations have little to no effect on many of the macro properties of interest and can be frozen out. This allows for longer time steps (typically 2 fs with no other measures included) which halves the computational cost for a given simulation length. The hydrogen-heavy atom bond lengths can be constrained by a number of different algorithms, typically SHAKE[100] in CHARMM and NAMD and LINCS[101,102] in GROMACS.

3.1.3.1 Statistical Ensembles

A typical MD simulation will generate a system within the microcanonical (NVE) ensemble i.e. a system in which the number of particles, the volume and the energy of the system remain constant. The total energy, which remains constant, is the sum of the potential and kinetic energies. It can be calculated

from the positions and velocities of the particles in a similar manner to equation 3.21.

It is possible to generate a canonical (NVT) or isothermal-isobaric (NPT) ensemble by scaling the positions or velocities of the particles after a certain number of time steps. As equation 3.21 shows, the temperature is proportional to the average kinetic energy. If the temperature is different from that desired then the velocities of the particles can be scaled in order to reach the desired temperature. If this is done every step, a so-called "instant" correction, then the dynamics of the system are affected such that the simulation does not correspond to a true NVT ensemble.

In the place of the instant correction, a thermostat method[103] can be used to couple the system to a heat bath which will gradually add or remove energy to the system of a suitable amount over time. The kinetic energy is still modified by scaling the velocities but the rate of heat transfer is controlled by a coupling parameter (τ).

$$\begin{aligned} \frac{\mathrm{d}T}{\mathrm{d}t} &= \frac{1}{\tau} \quad (T_{desired} - T_{actual}) \\ &\Updownarrow \\ \text{velocity scale factor} &= \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{desired}}{T_{actual}} - 1\right)} \end{aligned} \tag{3.25}$$

The downside to thermostat methods is that they still do not produce a correct canonical ensemble; correct averages are produced but these methods give incorrect fluctuations of properties. In order to give a correct ensemble the heat bath can be considered an integral part of the system as in *Nosé-Hoover*[104,105] methods. In these methods the heat bath is assigned fictious dynamic variables which are evolved along with the other variables in the system and this allows the production of a true canonical ensemble.

The pressure of the system can also be held constant by scaling the positions of the atoms in a similar fashion to equation 3.25, only the system is coupled to a "pressure bath". This also does not produce a correct NPT

ensemble which can be addressed by applying the Nosé-Hoover approach.

3.1.4 Solvent Models

3.1.4.1 Explicit solvent and Periodic Boundary Conditions

When carrying out MD simulations it is important to consider any effect the surrounding environment may have. It is therefore important to include solvent effects on the protein in the simulation if any accuracy is to be achieved, especially in the case of water which can form hydrogen bonds with the solute protein *in vivo*. This is most readily done by the inclusion of explicit water molecules, and a number of models exist to achieve this. Within the CHARMM forcefield the TIP3P[106] water model is the most commonly used.

In the TIP3P model water is modelled as three point charges which correspond to the charges on the oxygen and two hydrogens. The sum of the positive charges on the two hydrogens is equal (but opposite) to the negative charge on the oxygen. Other three site models include the Simple Point Charge (SPC) model and the Extended Simple Point Charge (SPC/E) model. There are also four and five point models such as TIP4P[107] and TIP5P[108] which use dummy atoms to model the lone pairs on the oxygen, but these are more computationally expensive.

	SPC	**SPC/E**	**TIP3P**
***r*(OH) (Å)**	1.0	1.0	0.9572
∠HOH (°)	109.47	109.47	104.52
***q*(O) (C)**	-0.82	-0.8476	-0.834
***q*(H) (C)**	+0.41	+0.4238	+0.417

Table 3.1: Comparison of bond lengths, angles and charges in some three-site water models.

The main disadvantage of including explicit solvent molecules is that it increases the number of atoms in the system. Since it is only the solvent molecules that directly interact with the solute that are of interest, this means

computational time is taken up calculating the trajectories of molecules that will simply be discarded.

In order to simulate the effects of bulk solvent on a solute molecule periodic boundary conditions were utilised. In the simplest example, the molecules are placed within a cubic box which is duplicated in all directions. Figure 3.4 shows what this looks like in two dimensions with nine boxes, but the simulation box is also duplicated in front and behind the nine shown in the figure. Thus, the simulation box is surrounded by twenty six identical boxes which touch it on every edge and vertex. During an MD run, if a molecule leaves the box via one edge, it effectively reappears in the box at the opposite side of the box. This prevents edge effects and loss of solvent which can cause artefacts in the simulation. The dimensions of the simulation box should be chosen in conjunction with the electrostatic cut-off such that the solute molecule in one periodic image does not interact with the solute in the next periodic image, i.e. that the solute molecule does not interact with itself as this could cause artefacts.

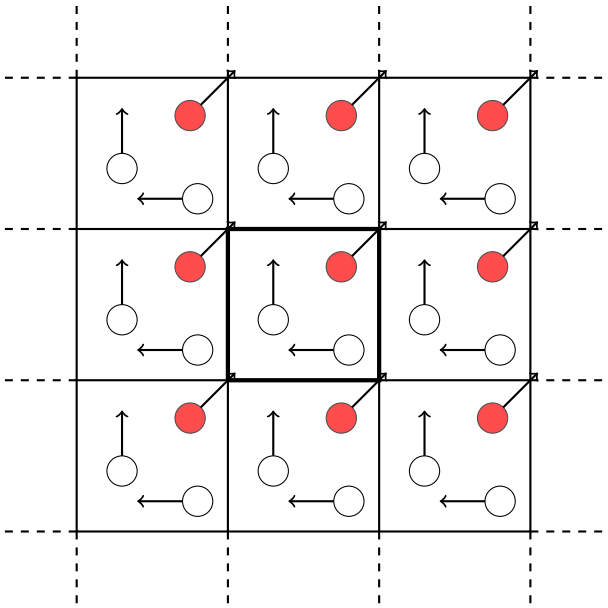


Figure 3.4: Periodic boundary conditions for a cubic box in two dimensions.

Figure 3.4 shows periodic boundary conditions with a cubic box, but other shapes that are congruent in three dimensions may be chosen as well. Other popular choices for the shape of the periodic cell include the truncated octahedron and the rhombic dodecahedron. These two are often chosen as they are close to being spherical and thus simulation of unnecessary solvent molecules is minimised. Choosing octahedral or rhombic periodic boxes reduces the computational cost for a given simulation length.

3.1.4.2 Implicit Solvent Models

In many cases it can be advantageous to use a system which includes the solvent implicitly rather than explicitly. Using an implicit solvent model can radically reduce the number of particles in the system and hence allow for much longer time-scales to be modelled. Implicit or continuum models[109] treat the solvent as a uniform polarisable medium with a set dielectric constant, ε. The solute molecule is placed in a cavity in the medium of suitable shape and size, as shown in figure 3.5.

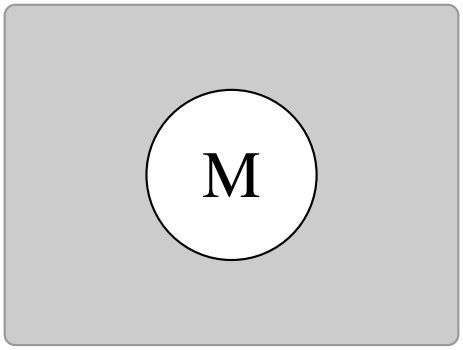


Figure 3.5: Implicit or continuum solvent model using a spherical cavity.

The creation of the cavity in the polarisable medium costs energy and the free energy required to move a solute molecule from vacuum to the cavity is given by:

$$\Delta G_{solvation} = \Delta G_{cavity} + \Delta G_{dispersion} + \Delta G_{elec} \tag{3.26}$$

The dispersion interactions between solvent and solute roughly correspond to the van der Waals energy and provide a stabilisation which counteracts the destabilisation caused by the creation of the cavity. These two terms are assumed to be proportional to the solvent accessible surface area giving

$$\Delta G_{cavity} + \Delta G_{dispersion} = \gamma SAS \tag{3.27}$$

where γ is an empirical surface tension coefficient[110] which relates the continuum solvent model to experimental solvation data. The electrostatic portion of equation 3.26 can be approximated at several different levels of theory.

The simplest continuum model is a solute molecule in a spherical cavity as shown in figure 3.5. In this case, only the lowest order of the electric moment needs to be accounted for. The Born model[110] details the difference in energy between a vacuum and a medium of dielectric constant, ε given a net charge of q in a cavity of radius a.

$$\Delta G_{elec}(q) = -\left(1 - \frac{1}{\varepsilon}\right)\frac{q^2}{2a} \tag{3.28}$$

GROMACS utilises the Generalised Born/Surface Area model,[111] which uses partial atomic charges in the place of q in equation 3.28. This can be combined with the Coulomb interaction to give

$$\begin{aligned} \Delta G_{elec}(Q_i, Q_j) &= -\left(1 - \frac{1}{\varepsilon}\right)\frac{Q_i Q_j}{f_{ij}} \\ f_{ij} &= \sqrt{r_{ij}^2 + a_{ij}^2 e^{-D}} \\ a_{ij}^2 &= a_i a_j \\ D &= \frac{r_{ij}^2}{4a_{ij}^2} \end{aligned} \tag{3.29}$$

3.2 Two-Dimensional Infrared Spectroscopy

In general, interpretation of experimental spectra is aided by theoretical calculations, and this especially applies to vibrational spectroscopy in biological systems. The field has a well-developed theoretical basis, meaning that experimental data can be compared with theoretical models, such as those based on MD simulations. These comparisons can help improve both the theoretical models and experimental methods. Experimentally, the information in a protein 2DIR spectrum comes from two main sources: local structure and global conformation. The local structure content includes, for example, hydrogen bonding and solvent exposure, which can be found via site-specific isotope labelling. Any theoretical model for protein 2DIR should take into account local effects, such as the influence of local electrostatic environment on vibrational frequency, as well as global effects, particularly delocalisation of amide excitons due to site-to-site coupling. A theoretical method for 2DIR should also describe the sensitivity of oscillator site energies, or the frequency of each individual oscillator, to their local environment.

3.2.1 Theoretical Framework

Perhaps the simplest place to begin a discussion of the theoretical framework of 2DIR is with a comparison with conventional FTIR. For a linear spectrum measured using weak infrared radiation one only needs to consider two vibrational levels and one Feynman diagram—assuming that all molecules in the sample are in the ground state prior to interaction with the laser pulse. Double sided Feynman diagrams are a useful way of describing the processes occurring in vibrational spectroscopy, summarising what is a complex pathway in an intuitive visual manner. The rules for the construction of these Feynman diagrams are described below.

The double sided Feynman diagrams describe the processes affecting the density matrix of the system under consideration. A density matrix describes a

quantum system in a mixed state or a statistical ensemble of several quantum states. The left and right vertical lines for each diagram represent the time evolution of the *ket* and the *bra* of the density matrix respectively, with the state at $t = 0$ at the bottom of the diagram and all times occurring after this above it[38]. *Braket* notation—or Dirac notation—is often used in physics to represent large or complex vectors. In the case of the double sided Feynman diagrams they represent the two halves of the density matrix and number in the *bra* or *ket* represents the current state of that part of the matrix. 0 is the ground state, 1 the first excited state and, in the case of the diagrams for non-linear spectroscopy, 2 is the doubly excited state. The arrows in the diagram represent the interactions with the light source with k being the incident photon. In the linear diagram there is only one pulse but for the non-linear diagrams there are three which are labelled accordingly[38]. The last interaction at the top of the diagram is the emission of the signal and by convention is often depicted with a different arrow type. Only diagrams with emission from the *ket* (left) are shown; the corresponding diagrams showing emission from the *bra* are simply the complex conjugate of the ones shown and carry no additional information.

Each diagram is given the sign $(-1)^n$, where n is the number of interactions on the right (*bra*) side of the diagram. Each time an interaction comes in from the right in the commutator it carries a minus sign. The exception to this is the last interaction, or the emission of the signal. Since it is not part of the commutator it carries no sign. An arrow pointing to the right represents an electric field with $e^{-i\omega t+i\vec{k}\cdot r+i\phi}$, while an arrow pointing left represents a field with $e^{+i\omega t-i\vec{k}\cdot r-i\phi}$, where ω is the carrier or central frequency, $\vec{k}$ is the wave vector, $\vec{r}$ is the position vector and ϕ is a phase function[38]. This represents the fact the real incident electric field can be separated into the complex electric field and its complex conjugate. The emitted signal has a frequency and wavevector which is the sum of all the input frequencies and wavevectors. Arrows pointing in towards the system represent up-climbing (or

excitation) of the *bra* or *ket* of the density matrix; arrows pointing away from the system represent de-excitations. The last interaction always points away from the system since it represents emission of light. The last interaction must result in the system being left in a population state. For linear spectroscopy this will always be the ground state $|0\rangle\langle 0|$ but in non-linear spectroscopy can be higher excited states $|n\rangle\langle n|$[38]. The double sided Feynman diagram describing the processes important in FTIR is shown in figure 3.6.

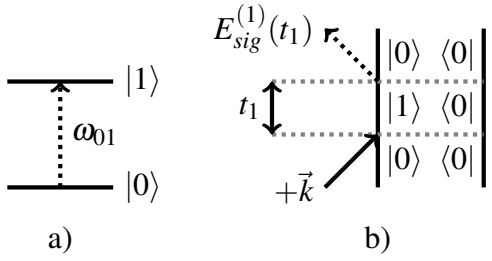


Figure 3.6: a) The $v = 0$ and 1 vibrational states under consideration and b) the only important Feynman diagram. Adapted from *Concepts and Methods of 2D Infrared Spectroscopy* by Hamm and Zanni.[38]

At times prior to interaction with the laser field the system is in the ground state, denoted by the density matrix $\rho = |0\rangle\langle 0|$. At time $t = 0$ the system is excited by the laser pulses and the off-diagonal density matrix element ρ_{10} is generated. A corresponding element is also generated at ρ_{01} but since this is redundant it will not be considered. The probability of this element being generated is proportional to the transition dipole moment μ_{10}

$$\rho_{10} \propto i\mu_{01} \tag{3.30}$$

This off-diagonal element (also known as the coherence) oscillates with frequency ω_{01} and decays with homogeneous lifetime T_2:

$$\rho_{10} \propto i\mu_{01} e^{-i\omega_{01}t_1} e^{-t_1/T_2} \tag{3.31}$$

At time t_1 the first order response of the system to the laser pulse, denoted as $R^{(I)}(t_1)$, becomes $\langle\mu\rangle = Tr[\mu_{01}\rho]$, which is

$$R^{(I)}(t_1) \propto i\mu_{01}^2 e^{-i\omega_{01}t_1} e^{-t_1/T_2} \tag{3.32}$$

The first order response function, $R^{(I)}(t_1)$, is the property of interest since it contains information about the molecule being studied. Unfortunately the molecular response is convoluted by the laser pulse itself, giving rise to the actual macroscopic polarization of the sample, $P^{(I)}$, which is proportional to the first order response.

$$P^{(I)}(t) \propto \int_0^\infty \mathrm{d}t_1 R^{(I)}(t_1) E(t-t_1) e^{-i\omega(t-t_1)+i\vec{k}_1\cdot\vec{r}+i\phi} \tag{3.33}$$

where $\vec{k}$ is the wave vector, $\vec{r}$ is the position vector and ϕ is a phase function

The macroscopic polarization gives rise to an emitted signal field with a 90°phase shift:

$$E_{sig}^{(1)}(t) \propto iP^{(1)}(t). \tag{3.34}$$

The lifetime of most vibrational modes is typically 1–5 ps, requiring the use of femtosecond pulses to measure them. Only with a femtosecond laser will the emitted signal truly reflect the first order response function rather than just the laser pulse. Once this is achieved the signal can be measured and Fourier transformed and the absorptive part of the band obtained. The width of the observed band is related to T_2 via

$$\Delta\nu = \frac{\Delta\omega}{2\pi} = \frac{1}{\pi T_2} \tag{3.35}$$

where $\Delta\nu$ is the full width half maximum of the band and $\Delta\omega$ is the frequency range. For a $T_2 = 1$ ps the full width half maximum of the band will be around 10 cm^{-1}.

In the case of 2DIR it is the third order response function that is probed by the incident laser. By necessity the system of interest must be anharmonic for

there to be a 2DIR spectrum and this is reflected in figure 3.7. Whereas for linear spectroscopy only one Feynman diagram was important, for 2DIR there are eight possible diagrams, of which six need to be considered. These are given in figure 3.7. The pathways outlined in the Feynman diagrams give the rephasing and non-rephasing contributions of the ground state bleach, stimulated emission and excited state absorption, as referenced in chapter 2.

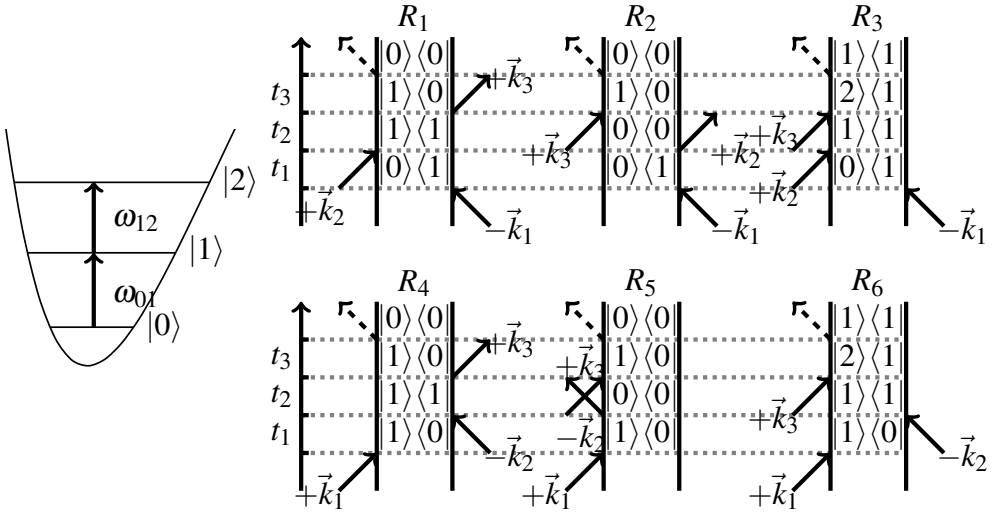


Figure 3.7: Potential energy curve for an anharmonic oscillator and six possible Feynman diagrams for third-order non-linear spectroscopy when the system starts in the ground state $\rho = |0\rangle\langle 0|$ (not shown). R_1 to R_3 are rephasing diagrams, R_4 to R_6 are non-rephasing diagrams. Adapted from *Concepts and Methods of 2D Infrared Spectroscopy* by Hamm and Zanni.[38]

As an example, the processes occurring in the R_4 non-rephasing pathway is illustrated below. This pathway is chosen because the interactions with the first laser pulse in this case is identical to the linear, first order response. Again we assume that the sample starts out in the ground state i.e. before time $t = 0$, $\rho = |0\rangle\langle 0|$. The ground state is not shown in figure 3.7. At time $t = 0$ the sample interacts with the field of the laser pulse and an off-diagonal matrix element ρ_{10} is generated as given in equation 3.30. The system undergoes dephasing during time t_1 as given in equation 3.31. After time t_1 the system undergoes interaction with the second laser pulse and is switched into a population state

$$\rho_{11} \propto i\mu_{01}^2 e^{-i\omega_{01}t_1} e^{-t_1/T_2} \tag{3.36}$$

After time t_2 the system is switched back into the coherence state ρ_{01} which undergoes dephasing during time t_3.

$$\rho_{10} \propto i\mu_{01}^3 e^{i\omega_{01}t_1} e^{t_1/T_2} e^{t_2/T_1} e^{i\omega_{01}t_3} e^{t_3/T_2} \tag{3.37}$$

After time t_3 the third order response function is emitted. This response function, responsible for the polarization $P^{(3)}(t)$ and the third order signal field $E_{sig}^{(3)}(t)$ is given by $\mathrm{Tr}[\mu_{01}\rho]$:

$$R_4(t_1,t_2,t_3) \propto i\mu_{01}^4 e^{i\omega_{01}t_1} e^{t_1/T_2} e^{t_2/T_1} e^{i\omega_{01}t_3} e^{t_3/T_2} \tag{3.38}$$

The response functions for the other pathways can be constructed in a similar fashion, the only differences between them are the signs of the coherences and the fact that pathways R_3 and R_6 involve doubly excited states. The response functions for the other five pathways are as follows

$$\begin{aligned}
R_1(t_1,t_2,t_3) &\propto i\mu_{01}^4 e^{+i\omega_{01}t_1} e^{t_1/T_2} e^{t_2/T_1} e^{i\omega_{01}t_3} e^{t_3/T_2} \\
R_2(t_1,t_2,t_3) &\propto i\mu_{01}^4 e^{+i\omega_{01}t_1} e^{t_1/T_2} e^{t_2/T_1} e^{i\omega_{01}t_3} e^{t_3/T_2} \\
R_3(t_3,t_2,t_1) &= i\mu_{01}^2\mu_{12}^2 e^{+i\omega_{01}t_1t_1/T_2^{(01)}} e^{t_2/T_1} e^{i\omega_{12}t_3t_3/T_2^{12}} \\
R_4(t_1,t_2,t_3) &\propto i\mu_{01}^4 e^{i\omega_{01}t_1} e^{t_1/T_2} e^{t_2/T_1} e^{i\omega_{01}t_3} e^{t_3/T_2} \\
R_5(t_1,t_2,t_3) &\propto i\mu_{01}^4 e^{i\omega_{01}t_1} e^{t_1/T_2} e^{t_2/T_1} e^{i\omega_{01}t_3} e^{t_3/T_2} \\
R_6(t_3,t_2,t_1) &= i\mu_{01}^2\mu_{12}^2 e^{-i\omega_{01}t_1t_1/T_2^{(01)}} e^{t_2/T_1} e^{i\omega_{12}t_3t_3/T_2^{12}}
\end{aligned} \tag{3.39}$$

where μ is the transition dipole moment (the electric dipole moment associated with the transition between the two states.).

Equation 3.39 gives the third order response functions as being identical for pathways R_1 and R_2, and for R_4 and R_5. This is only an approximation and

does not apply to the more sophisticated theories of pure dephasing. However, it is sufficient for our purposes.

Due to the fact that the signal in rephasing diagrams R_1, R_2 and R_3 are all emitted in the same direction (i.e. the $-\vec{k}_1+\vec{k}_2+\vec{k}_3$ direction) these signals cannot be further separated. The same is true for the non-rephasing diagrams R_4, R_5 and R_6 which are emitted in the $+\vec{k}_1-\vec{k}_2+\vec{k}_3$ direction. The sum of the rephasing and non-rephasing signals is therefore measured instead

$$\begin{aligned} R_{1,2,3}(t_3,t_2,t_1) &= \sum_{n=1}^{3} R_n(t_3,t_2,t_1) \\ R_{4,5,6}(t_3,t_2,t_1) &= \sum_{n=4}^{6} R_n(t_3,t_2,t_1) \end{aligned} \tag{3.40}$$

If one assumes that the homogeneous dephasing time T_2 is the same for the 1-2 transition as it is for the 0-1 transition, then the response functions can be combined as follows

$$\begin{aligned} R_{1,2,3} &= 2i\mu_{01}^4 (e^{-i\omega_{01}(t_3-t_1)} - e^{-i((\omega_{01}-\Delta)t_3-\omega_{01}t_1)}) e^{-(t_1+t_3)/T_2} \\ R_{4,5,6} &= 2i\mu_{01}^4 (e^{-i\omega_{01}(t_3+t_1)} - e^{-i((\omega_{01}-\Delta)t_3+\omega_{01}t_1)}) e^{-(t_1+t_3)/T_2} \end{aligned} \tag{3.41}$$

where $\Delta \equiv \omega_{01} - \omega_{12}$ is the anharmonic frequency shift. These third-order response functions are what any theoretical method of calculating 2DIR spectra must be able to model. One common method of modelling 2DIR spectra—particularly of peptides and proteins—is via the exciton model.

3.2.2 Exciton Theory

Vibrational spectroscopy generally deals with a *normal mode* view, which is unsuitable for use with 2DIR due to the harmonic nature of normal modes. Anharmonicity is required in order for a 2D signal to appear and accounting for anharmonicity with normal modes is complex. It is much better, therefore, to approach the modelling of 2DIR spectra through a *local mode* description,

which is useful for describing spectral signals of molecules made up of repeating units—such as proteins. Each repeating unit is treated as a local coordinate, which in a 3D structure will couple with other units, vibrating in unison and forming a delocalised state called a *vibrational exciton*. Due to the distance dependence of this coupling, the exciton bands provide information on the global structure content, including information on the secondary structure.

The exciton method was adapted for non-linear spectroscopies, such as 2DIR, by Hamm, Lim and Hochstrasser when they recast earlier matrix calculations into a quantum Hamiltonian which included doubly excited states.[34] The exciton states were expressed in terms of eigenstates of a Hamiltonian that described the interactions between N sites.

$$\begin{aligned} \hat{H} = & \sum_{n=1}^{N} \varepsilon_n |n\rangle\langle n| + \sum_{m,n=1}^{N} J_{mn} |m\rangle\langle n| \\ & + \sum_{m,n=1}^{N} (\varepsilon_m + \varepsilon_n - \Delta\delta_{mn}) |mn\rangle\langle mn| \\ & + \sum_{m,n=1}^{N} \sum_{\substack{j,k=1 \\ (m,n\neq 1)}}^{N} J_{mn,jk} |mn\rangle\langle jk| \end{aligned} \tag{3.42}$$

Equation 3.42 employs the *braket* notation introduced in section 3.2.1 and includes terms for the site energies (ε_n) and the coupling constants between the singly excited states (J_{mn}). The second set of terms for the doubly excited states corresponds to cases in which either a single oscillator is excited twice ($m = n$) or two different oscillators are excited once ($m \neq n$). The anharmonicity, Δ, is the difference between the absorption frequency of the fundamental ($0 \rightarrow 1$) transition and the overtone ($1 \rightarrow 2$) transition. The anharmonicity of the systems is critical in 2DIR, as there is no signal if the system is perfectly harmonic. δ is the Kronecker delta, which is 0 if $m \neq n$ and 1 if $m = n$. A schematic of the relevant Hamiltonian matrix elements and the energy levels for a two oscillator system is given in figure 3.8.

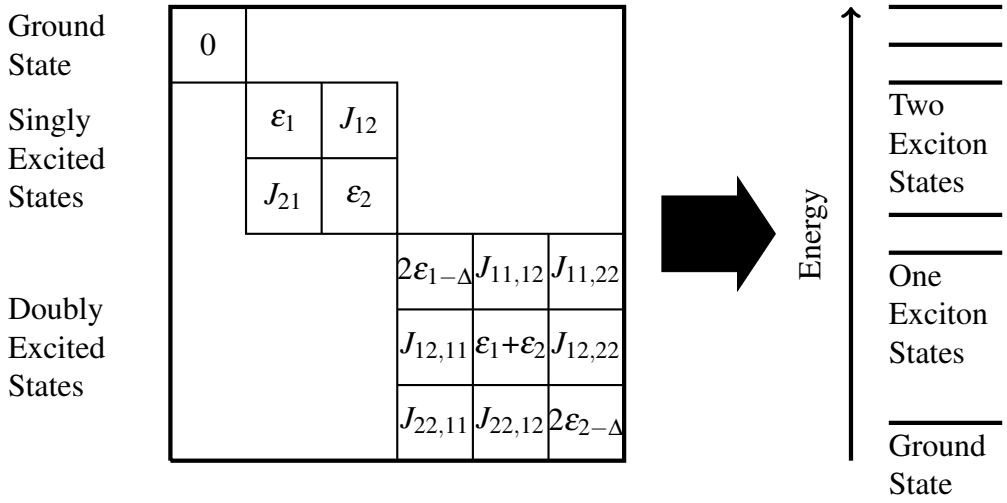


Figure 3.8: Schematic of a Two Exciton Hamiltonian Matrix.

For third-order non-linear spectroscopy such as 2DIR, it is sufficient to consider basis states only up to double excitations i.e ($|ij\rangle = |00\rangle, |10\rangle, |01\rangle, |20\rangle, |02\rangle, |11\rangle$), because the pulses in third-order non-linear spectroscopy do not probe higher states. Hence the schematic in figure 3.8 only shows the Hamiltonian matrix up to the doubly excited states. The double exciton states can be calculated from the single exciton Hamiltonian as shown in figure 3.8.

3.2.3 Calculating 2DIR Spectra from MD Simulations

3.2.3.1 Frequency Maps

Computational modelling of the amide I band of protein is of great importance in interpreting and understanding experimental spectra. These computed spectra allow the underlying behaviour of the protein to be linked with the observed experimental result. The shape of the amide I band of proteins is known to be sensitive to local environments and the couplings between amide units. This effect of environment on line shape is typically modelled using frequency “maps” which have parametrized the local-mode

frequencies as a function of the electrostatic potential field, and gradients involved in the amide I vibration.[57,76,77,112,113] Calculation of the electrostatic potential at several points on the potential energy surface may be sufficient to characterize the electrostatic field. Maps also exist for calculating nearest-neighbour frequency shifts, amide I transition dipoles, and vibrational couplings as a function of peptide geometry.[78,84] These frequency maps are much less expensive computationally than calculation of the band via *ab initio* methods and have been shown to give rise to relatively small errors (on the order of a few cm^{-1}).[114] There is some evidence to suggest that these frequency maps work better for residues in the interior of a protein and less well for the terminal residues, and better results may be obtained when using the map with the same forcefield used for the parametrization of the map.[114]

3.2.3.2 The Hamm and Zanni Code: A Simple Exciton Method

For this simple model,[38] only backbone atoms CO−NH−Cα are considered and the one-exciton Hamiltonian is calculated from these backbone atom coordinates. The nearest neighbour couplings are calculated using dihedral angles and a fully *ab initio* coupling map[115] while the non-nearest neighbour couplings are computed using transition charges taken from calculations on an isolated NMA molecule[116].

In this simple exciton model[38], there are only two amide bond environments considered: hydrogen bonded and non-hydrogen bonded. The frequency of a non-hydrogen bonded amide unit is set at 1660 cm^{-1}. When an amide C=O group is hydrogen-bonded to a N−H, the local mode frequency is down-shifted by 30 cm^{-1}.

$$\delta\omega = -\Delta\omega_{hbond}(2.6\text{Å} - r_{O\cdots H}) \tag{3.43}$$

The acceptance angle, $\angle$N−H−O, for this down-shift is less than 60 degrees[38].

Two types of broadening are important in 2DIR spectra: homogeneous

broadening and inhomogeneous broadening. Homogeneous broadening is the increase in the linewidth of an atomic transition caused by effects which affect different radiating or absorbing atoms in the same way. Inhomogeneous broadening is the increase in the linewidth of an atomic transition caused by effects which act differently on different radiating or absorbing atoms.

In this simple exciton method[38], inhomogeneous broadening arises as a result of the structural variation within the set of structures used, whether MD trajectory or several conformers included in a PDB file. To mimic broadening due to solvent effects, a random frequency shift is added to the diagonal elements of the one-exciton Hamiltonian. The practical consequence of this random frequency shift is that spectra will be unreliable when calculated for a single structure: spectra *must* be averaged over a number to structures in order to be even qualiatatively correct.

The two-exciton Hamiltonian is constructed from the one-exciton Hamiltonian. The one- and two-exciton Hamiltonians are diagonalized, and the transition dipoles μ_{0i} and μ_{ik} (the transition dipoles between the ground state to the one-exciton state and from the one-exciton states to the two-exciton states respectively) are calculated using the transition charge model[116]. Linear and non-linear response functions are calculated using:

$$
\begin{aligned}
R_1 = \;& i\langle(\hat{\mu}_{0i}\cdot\hat{Z})^2(\hat{\mu}_{0j}\cdot\hat{Z})^2\rangle e^{+i\omega_j t_1 + i(\omega_j-\omega_i)t_2 - i\omega_i t_3} e^{(-t_1+t_3)/T_2} \\
R_2 = \;& i\langle(\hat{\mu}_{0i}\cdot\hat{Z})^2(\hat{\mu}_{0j}\cdot\hat{Z})^2\rangle e^{+i\omega_j t_1 - i\omega_i t_3} e^{-(t_1+t_3)/T_2} \\
R_3 = \;& i\langle(\hat{\mu}_{0i}\cdot\hat{Z})(\hat{\mu}_{0j}\cdot\hat{Z})(\hat{\mu}_{ik}\cdot\hat{Z})(\hat{\mu}_{jk}\cdot\hat{Z})\rangle\cdot \\
& e^{+i\omega_j t_1 + i(\omega_j-\omega_i)t_2 - i(\omega_k-\omega_j)t_3} e^{-(t_1+t_3)/T_2} \\
R_4 = \;& i\langle(\hat{\mu}_{0i}\cdot\hat{Z})^2(\hat{\mu}_{0j}\cdot\hat{Z})^2\rangle e^{-i\omega_j t_1 - i(\omega_j-\omega_i)t_2 - i\omega_i t_3} e^{-(t_1+t_3)/T_2} \\
R_5 = \;& i\langle(\hat{\mu}_{0i}\cdot\hat{Z})^2(\hat{\mu}_{0j}\cdot\hat{Z})^2\rangle e^{-i\omega_j t_1 - i\omega_i t_3} e^{-(t_1+t_3)/T_2} \\
R_6 = \;& i\langle(\hat{\mu}_{0i}\cdot\hat{Z})(\hat{\mu}_{0j}\cdot\hat{Z})(\hat{\mu}_{ik}\cdot\hat{Z})(\hat{\mu}_{jk}\cdot\hat{Z})\rangle\cdot \\
& e^{-i\omega_j t_1 + i(\omega_j-\omega_i)t_2 - i(\omega_k-\omega_i)t_3} e^{-(t_1+t_3)/T_2}
\end{aligned}
\tag{3.44}
$$

for the $\langle ZZZZ\rangle$ polarization, where R_n is the response pathway (as given in 3.7) and $\hat{Z}$ is the polarization of the incident photon[38]. It is assumed that the homogeneous dephasing time, T_2, is 2 ps. The dipole terms are given as

$$
\langle(\mu_{0i}\cdot\hat{Z})^2(\hat{\mu}_{0j}\cdot\hat{Z})^2\rangle = \frac{1}{15}|\hat{\mu}_{0i}|^2|\hat{\mu}_{0j}|^2(1+\cos\theta_{0i,0j})
$$

$$
\begin{aligned}
\langle(\mu_{0i}\cdot\hat{Z})(\hat{\mu}_{0j}\cdot\hat{Z})(\mu_{ik}\cdot\hat{Z})(\hat{\mu}_{jk}\cdot\hat{Z})\rangle = \frac{1}{15}|\hat{\mu}_{0i}||\hat{\mu}_{0j}||\hat{\mu}_{ik}||\hat{\mu}_{jk}|\cdot(\cos\theta_{0i,0j}\cos\theta_{ik,jk} \\
+\cos\theta_{0i,ik}\cos\theta_{0j,jk}+\cos\theta_{0j,ik}\cos\theta_{0j,ik})
\end{aligned}
\tag{3.45}
$$

The response functions are averaged over all structures, and Fourier transformed to give linear and purely absorptive 2DIR spectra[38]. The diagonalization of the two-exciton Hamiltonian matrix is the most computationally demanding part of the routine, scaling with the sixth power of the number of amino acids. This scaling means that even for this relatively

simple exciton approach, 2DIR spectra calculations for larger proteins remains challenging and will often require further approximation to become tractable.

3.2.3.3 The SPECTRON Package: Non-linear Exciton Method

The SPECTRON code package also utilises an exciton based approach to calculating 2DIR spectra of peptide systems and does so by solving the non-linear exciton equations (NEE).[117] The SPECTRON code is capable of calculating the one-exciton Hamiltonian from an ensemble of system configurations (in this thesis the ensemble was generated by MD simulation). The method used by the package, however, does not contain a map that allows for the calculation of tertiary amide frequencies, meaning it is difficult to simulate the spectra of systems containing proline, such as the enoyl-ACP reductase in *M. tuberculosis*. The package can be used to calculate the non-linear response from a Hamiltonian matrix generated separately, which was the procedure used in this thesis.

Once the one-exciton Hamiltonian has been obtained, the fluctuating vibrational-exciton Hamiltonian, $\hat{H}_S(t)$, and transition dipole, $\mu(t)$, matrices can be calculated via the use of frequency maps. For the NEE approach only the one-exciton block of the matrix is required as the two-exciton block can be derived from the scattering of the one-exciton states. The products of four transition dipoles are obtained[37] and average and the response functions for each configuration obtained via the Green function expressions.[118] The response functions are averaged over all configurations.

It is the calculation of the scattering matrix that gives rise to the non-linear response and gives all necessary information regarding the two-exciton resonances while avoiding explicit calculation of the two-exciton eigenstates. The simplified response functions utilised do not include population relaxation and therefore are only suitable for use with short waiting times (t_2). The k_I signal($k_I = -k_1 + k_2 + k_3$) is given as

$$S^{k_I}_{\nu_4\nu_3\nu_2\nu_1}(\Omega_3, t_2 = 0, \Omega_1) = 2i\sum_{e_4\ldots e_1}\langle \mu^{\nu_4}_{e_4}\mu^{\nu_3}_{e_3}\mu^{\nu_2}_{e_2}\mu^{\nu_1}_{e_1}\rangle_o \times$$
$$I^*_{e_1}(-\Omega_1)I_{e_4}(\Omega_3)\Gamma_{e_4e_1e_3e_2}(\Omega_3+\varepsilon_{e_1}+\gamma_{e_1})\mathscr{T}_{e_3e_2}(\Omega_3+\varepsilon_1+i\gamma_{e_1}) \tag{3.46}$$

and the K_{III} signal ($k_{III} = +k_1 + k_2 - k_3$) is given by

$$S^{k_{III}}_{\nu_4\nu_3\nu_2\nu_1}(\Omega_3, \Omega_2, t_1 = 0) = 2i\sum_{e_4\ldots e_1}\langle \mu^{\nu_4}_{e_4}\mu^{\nu_3}_{e_3}\mu^{\nu_2}_{e_2}\mu^{\nu_1}_{e_1}\rangle_o \times$$
$$I_{e_4}(\Omega_3)I^*_{e_3}(\Omega_2-\Omega_3)[\Gamma_{e_4e_3e_2e_1}(\Omega_2)\mathscr{T}_{e_2e_1}(\Omega_2) -$$
$$\Gamma_{e_4e_3e_2e_1}(\Omega_3+\varepsilon_{e_3}+i\gamma_{e_3})\mathscr{T}_{e_2e_1}(\Omega_3+\varepsilon_{e_3}+i\gamma_{e_3})] \tag{3.47}$$

where $\Gamma_{e_4e_3e_2e_1}(\Omega)$ is the exciton scattering matrix which gives the two-exciton states in terms of one-exciton Green's functions.[117] The one-exciton Green's function, $I_e(\Omega)$, is given as

$$I_e(\Omega) = \frac{i}{\Omega - \varepsilon_e + i\gamma_e} \tag{3.48}$$

and the two-exciton Green's function, $\mathscr{T}_{ee'}(\omega)$, is given as

$$\mathscr{T}_{ee'}(\omega) = \frac{i}{\omega - \varepsilon_e - \varepsilon_{e'} + i(\gamma_e + \gamma_{e'})} \tag{3.49}$$

Green's functions are often used in theoretical physics as propagators in Feynman diagrams, such as in figure 3.7. In this case, the propagator serves to model the incident laser beam on the sample, and the frequency of the laser can be set in both equation 3.48 (Ω) and equation 3.49 (ω). γ_e is a broadening (of the laser pulse) that can be obtained from simulation or experiment and ε_e and $\varepsilon_{e'}$ are permitivity constants for the sample The one-exciton Green's function excites the $0 \rightarrow 1$ transitions present in all pathways shown in figure 3.7 and the two-exciton Green's function excites the $1 \rightarrow 2$ transition in pathways R_3 and R_6. The signals are calculated as a function of the four transition dipoles, $\mu^{\nu_4}_{e_4}\mu^{\nu_3}_{e_3}\mu^{\nu_2}_{e_2}\mu^{\nu_1}_{e_1}$, the one- and two-exciton Green's functions and convoluted

Lorentzians, $\Gamma_{e_i e_j e_k e_l}$, to simulate inhomogeneous broadening of the spectra. A non-fluctuating experimental anharmonicity of -16 cm^{-1} was used based on the results of earlier studies of amide I spectra,[119] which is incorporated into the Hamiltonian.

The signals for each snapshot of the system are calculated and averaged. Homogeneous broadening of the signal is obtained from this averaging. In the SPECTRON package, the rephasing and non-rephasing signals are calculated separately, necessitating some post processing in order to produce the final, purely absorptive, spectrum.

K_I is the rephasing signal incorporating pathways R_1–R_3 in figure 3.7 and K_{III} is the non-rephasing signal incorporating pathways R_4–R_6. These pathways cannot be further disentangled. Since in the rephasing and non-rephasing pathways the coherence is reversed during time t_1 ($|0\rangle\langle 1|$ for rephasing pathways, $|1\rangle\langle 0|$ for rephasing), the two signals have different phase twists and appear in different quadrants of the spectrum. Spectra of the non-rephasing pathways appear in the (ω_1, ω_3) = (+,+) quadrant while the rephasing pathways appear in the (ω_1, ω_3) = (−,+) quadrant as shown in figure 3.9.

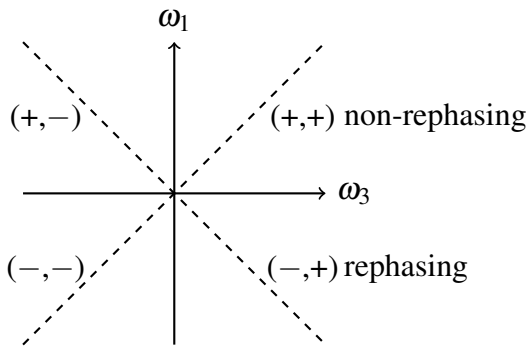


Figure 3.9: Schematic of the quadrants occupied by the rephasing (K_I) and non-rephasing (K_{III}) signals.

In order to obtain the purely absorptive spectrum, the ω_3 axis of the rephasing signal must be normalized i.e. multiplied by -1, in order to put the signals in the same quadrant of the spectrum. The real parts of both signals can then be added together, cancelling out the phase twist seen in figure 2.5. The

resulting spectrum is the purely absorptive signal for the system.

3.2.3.4 Numerical Integration of the Schrödinger Equation method

There are various approaches to calculating 2DIR spectra from MD trajectories. Many of the papers referenced in chapter 2 utilise the Numerical Integration of the Schrödinger Equation method (NISE)[120–122]. Therefore, this method is described here in more detail. In the NISE method, the sample is considered to comprise two parts: the system (the part that interacts with the applied laser field) and the bath (the part that does not). It is assumed that the system is affected by the applied fields, but that the state of the system has no effect on the bath. The system is treated quantum mechanically and is described by the time-dependent Schrödinger equation:

$$\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = -\frac{i}{\hbar}H(t)\phi(t) \tag{3.50}$$

where $\phi(t)$ is the wave function which describes the state of the quantum system Hamiltonian $H(t)$ which fluctuates, due to perturbations caused by the bath, which can be modelled by MD simulations. The time evolution of the bath is used to construct the time evolution of the Hamiltonian. The time-dependent Schrödinger equation cannot be directly solved in cases where the Hamiltonian is time-dependent. To get around this, the trajectory is split into short intervals; during these intervals, the Hamiltonian is assumed to be constant.

In the NISE method, the Hamiltonian is calculated in much the same manner as in the Hamm and Zanni method and the SPECTRON method, via the use of *ab initio* frequency maps, with longer range interactions coming from either transition dipole or transition charge coupling schemes. The simulation of the signal contributions outlined above involves a four point correlation function of transition dipoles, similar to what is seen in equation 3.46. Time evolution operators are inserted between the transition dipoles to account for the time evolution between interactions with the laser pulses.

The 2D spectrum itself is obtained by a Fourier transform of the signal with respect to the delay between the first and second pulse and a Fourier transform with respect to the delay between the final pulse and the signal emission. The frequency associated with the first Fourier transform is denoted ω_1; the second is denoted ω_3. The frequencies are taken for each time period, t_1, t_2 or t_3 , and are plotted as a function of ω_3 (probe or detection frequency) and ω_1 (pump or excitation frequency), which are the horizontal and vertical axes, respectively. The waiting time dynamics of the system can be revealed through analysing the t_2 dependence of the spectrum.

3.2.4 A Practical Outline of Exciton Calculations

The starting point for exciton calculations in this thesis was to carry out an MD simulation in order to obtain structural and dynamic distribution of the system. Once an MD calculation has been carried out it is then possible to extract a number of snapshots from the total trajectory. This is done to reduce the computational cost of the calculation; since the two-exciton Hamiltonian is found and diagonalised for each snapshot and an MD trajectory covering multiple nanoseconds will have several thousand snapshots even with infrequent saving to the trajectory, exciton calculations on the whole trajectory are computationally intractable. 200 structures sampled at intervals along the trajectory is sufficient to model the conformational diversity of the system.

The one-exciton Hamiltonian and dipoles are obtained from 200 snapshots. In this thesis this was done using the amideImaps program to calculate the Hamiltonian from GROMACS format trajectories using the Jansen scheme.[112,113,123] This program outputs the Hamiltonian and dipole trajectories in the format required for the NISE program.[120–122] Since the work in this thesis was carried out using SPECTRON[117] the trajectories needed to be reformatted. SPECTRON can be used to calculate the Hamiltonian and dipoles from NAMD or CHARMM format trajectories but the maps used in this scheme

do not take into account the secondary amide unit of proline residues. The decision was therefore made to use AmideIMaps (which has schemes which do include proline) in order to calculate the Hamiltonian for proline containing InhA. Required input for the AmideIMaps program is simply the trajectory in GROMACS xtc format and the GROMACS topology (tpr) file. Computation time for this portion of the calculation was generally less than an hour for each trajectory of the larger systems studied.

Modeling of the Amide I band of proteins is generally done via frequency maps which link the peptide environment to the frequency of that peptide. This approach is especially useful for larger proteins for which the frequency calculations are intractable by *ab initio* methods. The interaction between the Amide I modes of different residues is taken into account by maps that describe the nearest neighbour interactions and nearest neighbour coupling. Longer range interactions are taken into account by transition coupling models. The scheme also contains a dihedral map for proline residues.[113]

The AmideIMaps program was used in conjunction with the Jansen scheme to extract the Hamiltonian and dipole matrices from the MD trajectories. The Jansen scheme utilises Jansen's transferable electrostatic map for frequencies which was paramaterised using the electric field and electric field gradients at C, O, N, and D atoms of the NMA-D molecule.[123] Interactions between amide I modes is taken into account by combining the elctrostatic map with Jansen's nearest neighbour frequency shift map[112] (which is based on the Ramachandran angle of the nearest neighbour) and Jansen's nearest neighbour coupling scheme.[112] The scheme also uses Jansen's paramaterisation of the Transition Charge Coupling model for longer range interactions.[112]

In order to obtain the purely absorptive spectrum from a trajectory using the photon echo method, both the rephasing (KI) and non-rephasing (KII) signals must be calculated. Linear absorption spectra can be calculated separately. SPECTRON takes as input the correctly formatted one-exciton

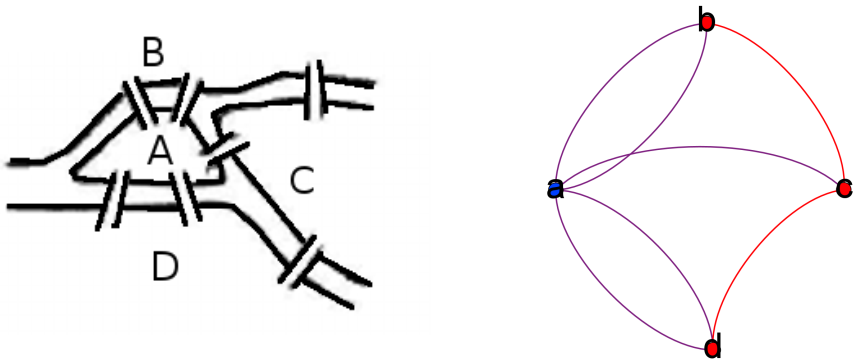
Hamiltonian and dipole trajectories and the rephasing and non-rephasing signals are calculated via the method given in given in section 3.2.3.3. The computational time for these each of the rephasing and non-rephasing calculations was found to be around 12 hours for 200 snapshots of a system the size of InhA (268 residues) with the calculations carried out in serial i.e. the purely absorptive spectrum for InhA can be obtained within approximately 24 hours.

In the case of the linear absorption signal the output from SPECTRON will be a file which lists frequencies in one column and intensities in the other that can be plotted without any further manipulation. In the case of both the rephasing and non-rephasing 2D signals, the output is more complicated. The output for these signals comes in six columns. The first two columns are indices of points in 2D space and can be ignored for our purposes. The first column – y axis, the second column – x axis. Columns three and four are the actual frequency (wavelength) values corresponding to these indices in the same order. The fifth and sixth columns are the real and imaginary parts of the signal at these frequencies.

In order to obtain the purely absorptive signal the real part of the rephasing and non-rephasing signal must be added together. For the 2DIR spectra presented in this thesis the purely absorptive data was extracted to a 200 by 200 matrix an d plotted using Gnuplot.

3.3 A Brief Introduction to Network Graphs

The Königsberg bridge problem, or more importantly Euler's approach to solving it, is widely considered to be the birth of graph theory. The city of Königsberg sat on a river which divide the city into four separate landmasses linked by seven bridges. The question underlying the problem is whether or not there is a path through the city which crosses each of the bridges exactly once. Euler simplified the problem by representing each of the landmasses of



(a) **(b)**

Figure 3.10: a) Diagram of the four landmasses of Königsberg showing the seven bridges connecting them and b) the problem represented as a graph, with the nodes being the landmasses and the edges being the bridges.

Königsberg as points and each of the bridges as lines joining each point. This abstract representation of the city was the first example of the kind of graph that graph theory is concerned with.

Many different problems can be represented using graphs, which can help simplify the problem or otherwise provide a different point of view of the problem. Graph theory then is a tool with which to approach the solving of a specific problem. One important thing to note is that the language describing graph theory is not unified or standardised and thus different authors and groups will use their own terminology. For the purpose of this thesis, the terms "graph", "network model", "network" and "network graph" may be used interchangeably. Further information on how a graph is defined is given below.

A graph,for our purposes, is a representation of a set of points and of how they are joined up.[124] It can be used to model the pairwise relationships between objects, or for other applications. The points on the graph (often referred to as *nodes* or *vertices*) can represent more or less anything: people, places or peptides. The lines between the points (or *edges*) represent the relationship between the nodes. For example the nodes in a particular graph could be people and the edges represent whether or not the pair are friends on Facebook; or the nodes could be cities around the world and the edges represent flight travelling between two cities. The graph displays the *relationship*

between the nodes and and any metrical properties (such as the distance between the cities) are not included.

All of the graphs in this thesis are what is called a *simple* graph, one in which each two nodes have a maximum of one edge connecting the,; there are no nodes with multiple edges between them nor any nodes with a loop connecting it to itself. A *weighted* graph,[125] which is sometimes referred to as a network,[126] is one in which the edges connecting two nodes are assigned a value, or *weight* which measures the strength of the relationship between the nodes. The nature of the edge weight will be dependant on the problem being investigated. In this thesis, the weight of the edge is the strength of the coupling between the nodes or peptide units. An undirected graph is one in which there is no distinction between the nodes associated with each edge, while in a directed graph the nodes are joined by edges that go in a specific direction, from one node to the other. In this thesis all network models are undirected.

Various algorithms exist for organising such graphs or models, and the resulting model can be used to visualise complex, often dynamic, data in a more intuitive manner than the raw data. The layout of a particular graph can have a critical effect on the graphs ability to convey information. Depending on the problem being investigated it might be useful to find out how specific sets of nodes cluster together, or have no edges; in other cases the layout may need to be done manually in order to maximise the usefulness of the graph. In this thesis the Force Atlas algorithm was used for the layout of the network models. This algorithm is described in more detail in section 5.2.3.

Graph theory is large area of study within mathematics and a only a small subsection is utilised for the purposes of this thesis. Network graphs of proteins are discussed further in section 5.2.3.

Chapter 4

Dynamics and Spectroscopy of Leu-enkephalin

4.1 Introduction

The enkephalins are a pair of related structures that are endogenous ligands for the opioid receptors.[127] The two structures are pentapeptides with consensus sequence Tyr-Gly-Gly-Phe-X. They differ only in the nature of the terminal residue X which in Met-enkephalin is methionine and in Leu-enkephalin is leucine. The enkephalins have a role in nociception—the encoding and processing of harmful stimuli in the nervous system—within the human body, specifically in pain regulation. Enkephalins are found in the spinal cord and central nervous system and have a number of functions including: gonadal function, pain perception, regulation of memory and emotional conditions, food and liquid consumption and regulation of immunological system, and some effects in the digestive system.

Leu-enkephalin binds primarily to the δ-opioid receptor, where it acts as an agonist similar to that of opioid alkaloids, such as morphine. Unlike morphine, both Met- and Leu-enkephalin have very flexible structures leading to debate about the conformational preferences of the enkephalins *in vivo*.

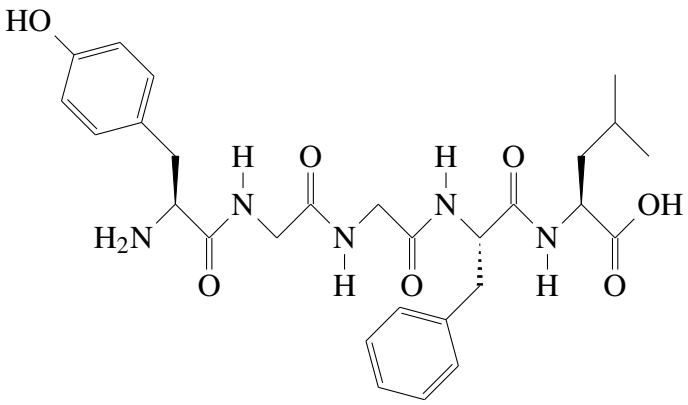


Figure 4.1: Structure of Leu-enkephalin.

Perhaps counter-intuitively, the enkephalins bind more strongly to the opioid receptors than the opioid alkaloids.[31] It has been suggested that conformationally dynamic species are able to penetrate membranes more readily than conformationally restricted species such as morphine, which is rigid and inflexible due to its cyclised nature.[32]

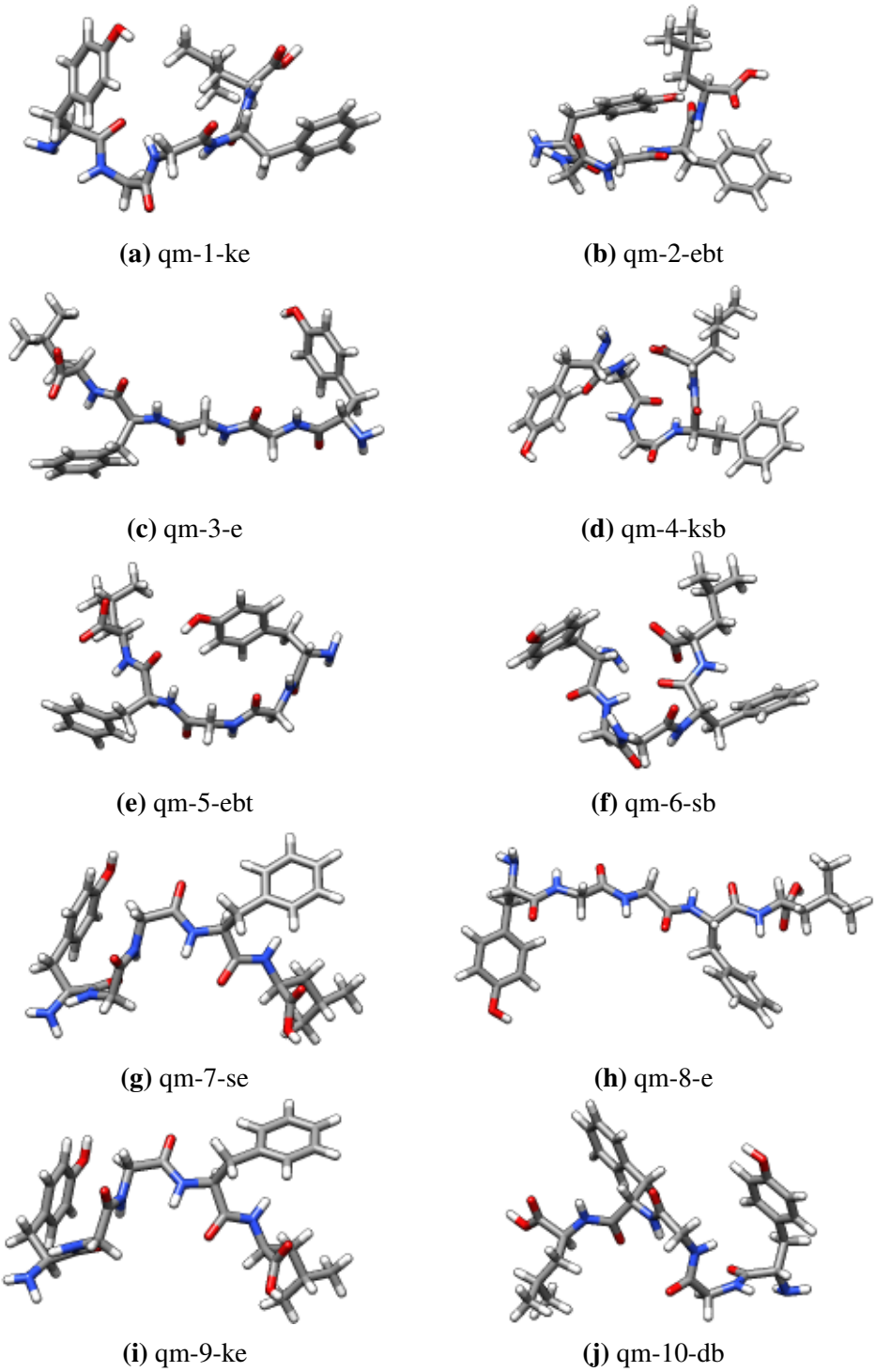
Previous work considered the *ab initio* amide I FTIR for 10 single low energy conformers calculated at the EDF1/6-31+G* level of theory and assignments of transitions were made.[128] This was compared with the FTIR for the same low energy conformers calculated via a floating oscillator model. The comparison between the *ab initio* amide I bands and model bands showed "mixed results". The findings were that the neglect of side chain interactions is likely to be important, especially when they are (e.g. for tyrosine) hydrogen bonded to the backbone.

Current work expands on this by considering the FTIR spectra of the low energy conformers incorporating dynamics of the systems in the calculation of IR spectra and investigating the effect of different solvent systems. Finally, we present computed 2DIR spectra for these systems and investigate whether the 2D spectrum provides additional information on the structure, dynamics of these systems and the effect of different solvent environments.

4.1.1 Low Energy Conformations of Leu-enkephalin

Ten low energy conformations of Leu-enkephalins were obtained from the literature.[128] These conformations were found via a molecular mechanics search in CHARMM followed by geometry optimisation of candidate structures at the EDF1/6-31+G* level of theory. EDF1 is an empirical functional that was shown to yield accurate vibrational frequencies in small molecules with low error compared to experiment.[129] 6-31+G* is a simple basis set which accommodates the main types of distortions that occur for atoms in molecules. The full procedure employed has been described previously[128] and the results of the conformational search are shown in figure 4.2.

The ten structures are denoted by a numerical code and a one to three letter suffix. This suffix is descriptive and indicates structural differences between the conformations. Structures labelled "e" (e.g. qm-3-e) exist in an extended conformation while those labelled "ke" (qm-1-ke) have a semi-extended conformation that is kinked. This is due to a seven-membered hydrogen bond ring that exists between adjacent residues. The "ebt" suffix refers to structures that are extended but have an arched or bent backbone, such as qm-2-ebt. This structure forms due to hydrogen bonding between the tyrosine phenoxy group and the end of the chain and resembles a large ring. Structures labelled "se" (qm-7-se) have an extended structure with several kinks. "sb" and "ksb" are two related labels where "sb" contains a single β-bend with hydrogen bonding patterns identical to the crystal structure; "ksb" is also a β-bend conformation but the hydrogen bonding patterns are different. The single structure labelled "db" is denoted as such because of the double bend structure it possesses. These ten conformations serve as the starting point for simulations in this chapter.



(a) qm-1-ke **(b)** qm-2-ebt

(c) qm-3-e **(d)** qm-4-ksb

(e) qm-5-ebt **(f)** qm-6-sb

(g) qm-7-se **(h)** qm-8-e

(i) qm-9-ke **(j)** qm-10-db

Figure 4.2: Ten low energy conformations of Leu-enkephalin, showing folded and extended conformations, rendered in Chimera.[50]

4.2 Modelling the Dynamics and Spectroscopy

4.2.1 MD Simulations

For each of the enkephalin simulations, the ten low energy conformers were used as starting structures. All MD simulations were carried out in GROMACS[91,92] unless otherwise stated. For the explicit water simulations, the structures were minimised using the steepest descent algorithm. The peptides were solvated in a truncated octahedron of TIP3P water[106] ensuring a minimum 10 Å of water between the peptide and the edge of the water box. 50 ps of dynamics was carried out with the positions of the protein atoms restrained in order to "soak" the peptide in the water molecules. 5 ns of equilibration was carried out under the NVT ensemble at 300 K using a leapfrog algorithm. Periodic boundary conditions and PME electrostatics[130] were used with a cut off of 10 Å to allow the Verlet cut off scheme[131] to be used. Heavy atom-hydrogen bonds were constrained using the LINCS[101,102] algorithm to allow a 2 fs time step. Production dynamics were carried out using the same parameters for 20 ns, with snapshots saved to the trajectory every 1 ps.

Vacuum MD simulations were carried out at 200 K to reduce the dynamical fluctuations of the peptide and investigate whether this would result in sharper, more distinct spectra. The leapfrog integrator was used, along with LINCS[101,102] to constrain the heavy-atom hydrogen bonds. 20 ns of dynamics were carried out with snapshots saved to the trajectory every 2 ps; the lower frequency was chosen due to minimise the broadening of spectral signals due to the higher conformational diversity in vacuum.

An implicit solvent simulation was carried out using the GBSA model[111] and the Still method of calculating the Born radii.[132] A dielectric constant of 37.5 was implemented to mimic acetonitrile. 20 ns of production dynamics were carried out at 300 K using LINCS.[101,102] Snapshots were saved to the trajectory every 2 ps.

4.2.2 Spectroscopy Calculations

The Hamiltonian matrices were obtained using CHARMM27 atomic partial charges to compute the shift of the diagonal elements of the exciton Hamiltonian and the Jansen scheme.[112, 113, 123]This scheme is applicable to proteins containing secondary (CONH)[112, 123] and tertiary ($CONC_{\delta}$)[113] amides. This scheme utilises Jansen's nearest neighbour frequency map; Jansen's electrostatic map for transition dipoles; Jansen's parametrization for transition charge coupling; and Jansen's nearest neighbour coupling map. The Jansen scheme overestimates the frequencies of a secondary amide unit (i.e. a non-proline unit) by 14 cm^{-1} but correctly reproduces the frequencies of a tertiary unit (i.e. the unit just prior to a proline). This can be corrected with utilities available in the NISE code,[123] but since SPECTRON was used here the spectra in this chapter were calculated using uncorrected frequencies. Linear absorption and 2DIR spectra were calculated from the Hamiltonian and dipole matrices using SPECTRON[117] and the parameters given in table 4.1.

4.3 Results

4.3.1 Characterisation of Initial Structures

The labelling system used for the enkephalin conformers gives useful structural details about them, but terms such as "kinked-extended" are not especially useful for interpreting spectral signatures and assigning them to specific structural variations. Therefore these conformations were characterised according to properties such as hydrogen bonding pattern, dihedral angles and carbonyl bond length to investigate whether these structural differences would be visible and assignable in the calculated spectra. Hydrogen bonding in particular contributes to a red-shift in peaks observed in IR spectra.

The hydrogen bonding data for the ten low energy conformers were calculated in GROMACS.[91, 92] OH and NH hydrogens were considered to be

Parameter	Setting
Linear absorption signal	
Lorentzian Linewidth (cm^{-1})	2
Initial Frequency (cm^{-1})	1550
Final Frequency (cm^{-1})	1750
Number of Frequencies	200
2DIR Calculation	
Method	Quasiparticle
Anharmonicity (cm^{-1})	16
Lorentzian (cm^{-1})	2
Number of Frequencies	200
kII signal	
Initial Frequency ω_1 (cm^{-1})	1550
Final Frequency ω_1 (cm^{-1})	1750
Initial Frequency ω_3 (cm^{-1})	1550
Final Frequency ω_3 (cm^{-1})	1750
Number of Frequencies	200
kI signal	
Initial Frequency ω_1 (cm^{-1})	-1550
Final Frequency ω_1 (cm^{-1})	-1750
Initial Frequency ω_3 (cm^{-1})	1550
Final Frequency ω_3 (cm^{-1})	1750
Number of Frequencies	200

Table 4.1: SPECTRON input parameters for calculating linear absorption and 2DIR spectra of Leu-enkephalin. The Lorentzian linewidth controls inhomogeneous broadening of the signal; the signal is computed in the range defined by the initial and final frequencies.

donor atoms and O and N were considered to be acceptors. A hydrogen bond is assumed to exist between an donor and an acceptor is the hydrogen–acceptor distance is less than 3 Å and the donor-hydrogen-acceptor angle is less than 120 $^{\circ}$.

The extended structures show fewer or no hydrogen bonds compared to the more folded structures. The "sb" and "ksb" structures are based on β-bend structures and show hydrogen bonding similar to that found in such structures. However, the double bend structure shows no hydrogen bonding at all.

Table 4.2 shows that the $n, n+2$ hydrogen bond interaction is the most common for this set of structures, which is to be expected since the majority of them are β-turn type structures and the number of residues is quite low. The

Structure	Hydrogen Bonding			Dihedral Angles		
	n,n+2	n,n+3	n,n+4	Residue	Phi (°)	Psi (°)
qm-1-ke	1			^{2}Gly	80	-66
				^{3}Gly	124	-24
				^{4}Phe	-142	129
qm-2-ebt		1	1	^{2}Gly	92	-128
				^{3}Gly	-133	25
				^{4}Phe	-87	87
qm-3-e				^{2}Gly	-179	170
				^{3}Gly	169	-174
				^{4}Phe	-150	168
qm-4-ksb	2	1		^{2}Gly	-83	63
				^{3}Gly	79	-64
				^{4}Phe	-105	-51
qm-5-ebt		1		^{2}Gly	-135	145
				^{3}Gly	134	-137
				^{4}Phe	-139	172
qm-6-sb	2			^{2}Gly	-170	50
				^{3}Gly	82	-62
				^{4}Phe	-74	171
qm-7-se	1		1	^{2}Gly	82	-71
				^{3}Gly	124	-25
				^{4}Phe	-156	170
qm-8-e				^{2}Gly	162	-159
				^{3}Gly	-176	175
				^{4}Phe	-154	172
qm-9-ke	2			^{2}Gly	81	-69
				^{3}Gly	119	-19
				^{4}Phe	-154	166
qm-10-db				^{2}Gly	-119	30
				^{3}Gly	-108	19
				^{4}Phe	-131	124

Table 4.2: Hydrogen bonding patterns and dihedral angles for the 10 low energy conformers.

structures that display only $n, n+2$ interactions are likely to give spectra with peaks in the β-sheet region of the spectrum. The $n, n+3$ and $n, n+4$ interactions are associated with α helices and are present in a number of structures. qm-2 contains only hydrogen bonds that are associated with α helical structures and it can be expected that the spectra will reflect this.

Linear absorption and 2DIR spectra were calculated for each of the ten low energy conformations of enkephalin in order to explore whether the small conformational difference in hydrogen bonding, dihedral angles etc. would be reflected in the 2D spectrum. The FTIR spectra given in figure 4.3 show distinct similarities between the qm-7 and qm-9 systems. These systems have the same total number of hydrogen bonds, though in different patterns, indicating that the total number of internal hydrogen bonds may influence the spectral signatures of the molecules. As shown in figure 4.2, qm-7 and qm-9 also have very similar structures which will also contribute to the similarity between the spectra. Conformation qm-6 also has the same number of hydrogen bonds and shows some similarities to both qm-7 and qm-9, though the splitting between the two groups of peaks is more pronounced in qm-6. Conformation qm-6 is similar to qm-7 and qm-9 but more compact, which may contribute to the different spectra. Conformations qm-3, qm-8 and qm-10 share the same number of hydrogen bonds (that is, none) but show no notable similarities in the IR spectrum. The same is true for structures qm-1 and qm-5, which both have a single hydrogen bond, though these are in different patterns.

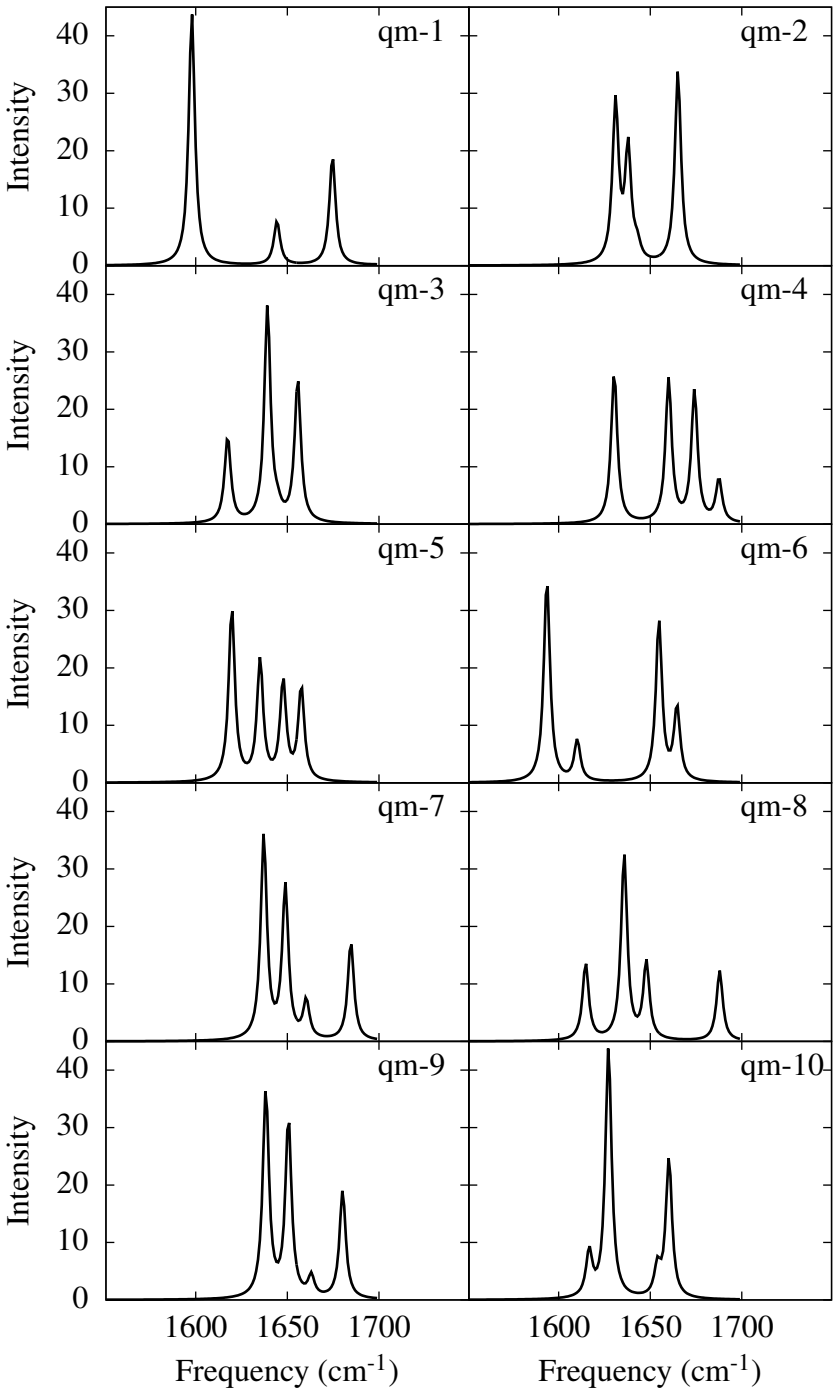


Figure 4.3: Linear absorption spectra calculated in SPECTRON for individual low energy conformations of enkephalin in the gas phase obtained at the EDF1/6-31G* level of theory.[129]

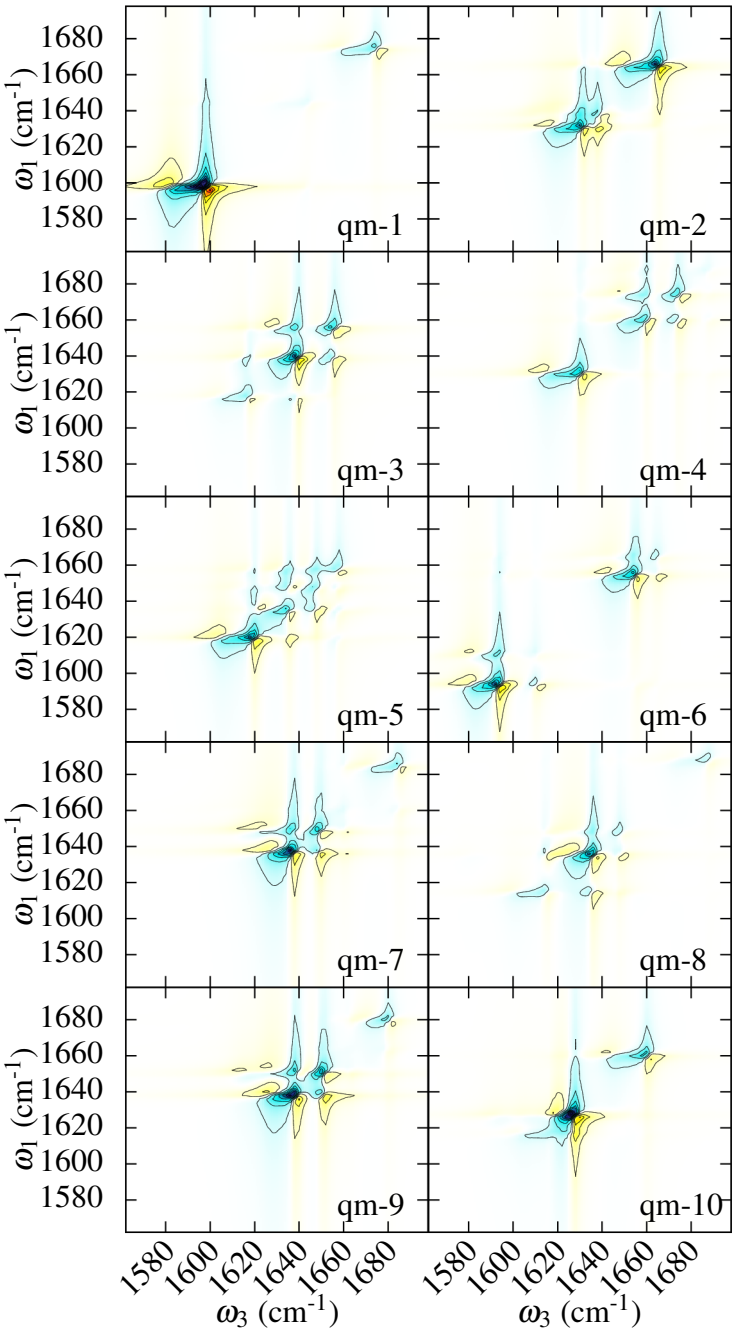


Figure 4.4: Purely absorptive 2D spectra calculated in SPECTRON for individual low energy conformations of enkephalin in the gas phase obtained at the EDF1/6-31G* level of theory.[129] Colour scale runs from blue (negative) to red (positive).

The 2DIR spectrum for qm-3 (shown in figure 4.4) is interesting because it shows strong coupling between the peak at 1640 cm^{-1} and the peak at 1660 cm^{-1} despite there being no hydrogen bond in table 4.2. The distance between one of the oxygen atoms on ^{5}Leu and the one of the hydrogens in the ^{4}Phe phenyl ring is 3.07 Å, which just lies outside of the acceptance parameters used to construct the hydrogen bond data in table 4.2, but it is possible there is a hydrogen bond there. Alternatively this coupling could be due to electrostatics or through bond interactions rather than hydrogen bonding. qm-8 shows a similar coupling pattern despite a lack of hydrogen bonds, though qm-10 lacks any coupling between the two major peaks in the 2DIR spectrum.

Conformations qm-7 and qm-9 are again very similar in the 2D spectrum, showing the same coupling pattern between the peaks at 1630 cm^{-1} and 1660 cm^{-1}. qm-6 is less similar in the 2D spectrum though it does show coupling between the peak at 1590 cm^{-1} and a peak at 1610 cm^{-1} that doesn't appear in the 2D spectrum. This peak may not appear due to its weak intensity in relation to the peak at 1590 cm^{-1}. This peak may be analogous to the main peak in the qm-7 and qm-9 spectra, just down-shifted by approximately 30 cm^{-1}.

Each of the structures qm-1 – qm-10 give rise to unique spectra for both the FTIR and 2DIR. The exception to this is structures qm-7 and qm-9, which are similar structures with almost identical spectra.

4.3.2 MD Simulation Analysis

Table 4.3 shows a comparison of the RMSD data for all ten structures in three different environments. It shows that, for the vacuum simulation, qm-4 and qm-6 have lowest mean RMSD—these structures are more compact conformations with higher numbers of hydrogen bonds. Both qm-4 and qm-6 are β-bent structures. Conformations qm-3 and qm-8 have the largest RMSDs as they are extended conformations with no internal hydrogen bonding. This means they have more flexibility in vacuum than structures with two or three

hydrogen bonds. Most of the other structures lie in between these two extremes with varying RMSDs between 2 and 3 Å.

Similar patterns are shown in the implicit solvent simulations, though the RMSD values for qm-4 and qm-6 are higher than in the vacuum simulation. In addition, the RMSD for qm-1 is very similar to that of qm-4. For the explicit water simulation all of the RMSDs are very similar, lying between 2 and 3 Å. In this simulation there is competition for the donor/acceptor sites and hydrogen bonding between the enkephalin and the surrounding water may disrupt intramolecular hydrogen bonds. In this simulation it is some of the more extended conformations that show the lowest RMSD (qm-8 and qm-9 and to lesser extent qm-1 and qm-7). Hydrogen bonding to the surrounding waters may be keeping these simulations in more rigid conformations than for the simulations in vacuum or implicit solvent. Again qm-4 and qm-6 show similar RMSD values.

For the single low energy conformation spectra, qm-7 and qm-9 showed very similar bands due to their similar structures. They also have two hydrogen bonds each which has the potential to keep the structures fairly close to the starting conformation during the simulation time. Looking at table 4.3, qm-7 and qm-9 show very similar mean RMSDs, even being identical in the implicit simulation. This suggests that the spectra that incorporate the effect of dynamics should show some similarities as well. The mean RMSD for qm-5 is the same across all three simulations which would suggest that the spectra calculated from each simulation may be similar.

	RMSD (Å)			
Structure	**Mean**	**Min**	**Max**	**SD**
	Vacuum			
qm-1	2.1	1.8	2.4	0.1
qm-2	2.8	2.3	3.1	0.1
qm-3	3.7	2.9	3.9	0.1
qm-4	0.8	0.2	1.3	0.2
qm-5	2.5	2.2	2.7	0.1
qm-6	0.8	0.2	1.5	0.3
qm-7	2.2	2.0	2.6	0.1
qm-8	3.8	3.3	4.0	0.1
qm-9	2.2	1.9	2.5	0.1
qm-10	3.0	2.5	3.2	0.1
	Explicit Water			
qm-1	1.6	0.5	2.3	0.3
qm-2	2.0	0.5	2.9	0.5
qm-3	2.9	1.1	3.9	0.6
qm-4	1.5	0.4	3.5	0.7
qm-5	1.9	0.4	2.7	0.4
qm-6	1.5	0.5	3.6	0.7
qm-7	2.0	0.8	2.8	0.3
qm-8	1.8	0.7	4.0	0.7
qm-9	1.8	0.5	2.6	0.3
qm-10	2.1	0.4	3.1	0.6
	Implicit Solvent			
qm-1	1.5	0.2	3.0	0.6
qm-2	1.7	0.2	3.1	0.6
qm-3	1.9	0.2	3.2	0.5
qm-4	2.1	0.2	3.2	0.6
qm-5	1.1	0.1	3.0	0.7
qm-6	1.5	0.2	3.3	0.6
qm-7	1.7	0.4	3.1	0.5
qm-8	1.5	0.2	2.5	0.4
qm-9	1.6	0.2	2.7	0.4
qm-10	1.7	0.1	2.5	0.5

Table 4.3: Mean, minimum and maximum backbone RMSD and standard deviation for MD simulations in vacuum, explicit water and acetonitrile.

4.3.3 Leu-enkephalin Spectra Calculated from Dynamics

FTIR and purely absorptive spectra calculated for all ten structures in three different environments are presented in order to investigate whether spectra generated from dynamics propagated from each low energy conformation resembles the spectra for that conformation. Essentially the question is this: do the conformations remain intact through the propagation of dynamics or not? We would expect the more extended conformations (e.g. qm-8) to be more flexible and therefore exhibit more conformational variation throughout the simulation, which should manifest in the spectrum. But for the more compact conformations with internal hydrogen bonding this might not be the case and we wanted to investigate whether the hydrogen bonds would be sufficient to maintain that conformation in vacuum or solvent simulations.

As might be expected for an extended conformation in vacuum, qm-8 shows a very broad and largely featureless signal in the FTIR spectrum (figure 4.5). This will be due to its flexibility and the large range of conformations it can adopt during the simulation. The structure labelled qm-3 is also an extended conformation but shows a markedly different spectrum to that of qm-8. Instead of a broad weak signal we see a single intense peak at around 1625 cm^{-1}. There appears to be some intensity underlying this at other frequencies, but the spectrum is dominated by this single peak.

Conformations qm-7 and qm-9 show similar FTIR spectrum in the vacuum, as they did in figure 4.3. This suggests that the hydrogen bonds survive in the vacuum simulation, keeping the structure close to the starting structures. Conformations qm-4 and qm-6, which show the lowest mean RMSD values for this simulation set, have markedly different spectra despite similar starting structures. That said, it is possible that the main peak at 1650 cm^{-1} in the qm-6 spectrum is also in the qm-4 spectrum, buried under the rest of the band.

The 2D spectrum for qm-1 is very inhomogeneously broadened (total width of the peak is round 100 cm^{-1} but two features can be distinguished under

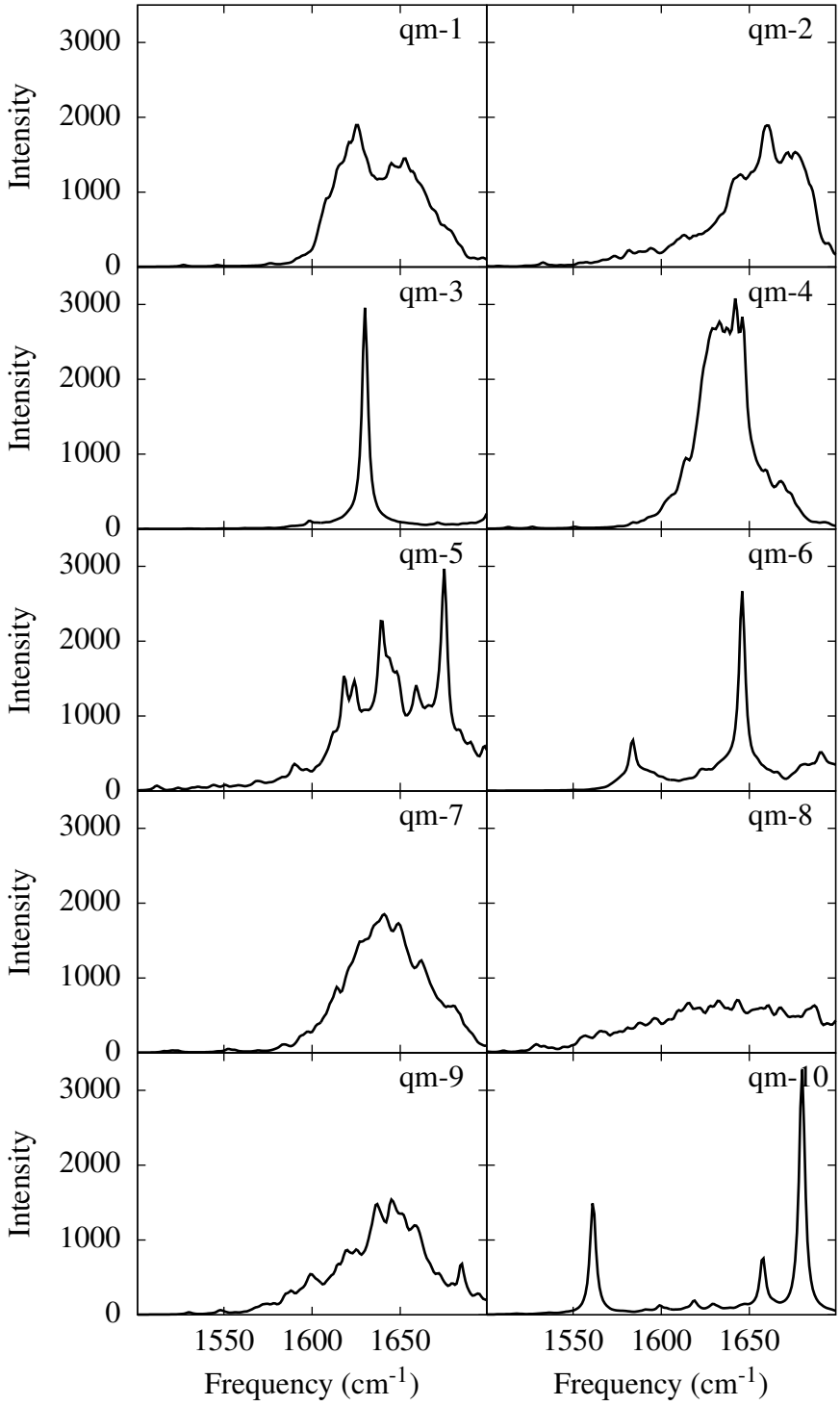


Figure 4.5: Linear absorption spectra calculated in SPECTRON from vacuum simulations of enkephalin.

the band, both with defined 1→2 off diagonal transitions. This is the same number of peaks seen in the single snapshot spectrum, though the peaks are much broader and appear closer together in the vacuum simulation. There may be some weak coupling between these two features.

qm-2 in the vacuum simulation shows a broad feature which appears to have only one peak, which contrasts with the two seen fr the static spectrum. Simulations for qm-3, qm-4, qm5, qm-6 and qm-10 show intense peaks with very little inhomogeneous broadening. The spectrum for qm-5, however, does resemble the spectrum for the static structure in number of diagonal peaks and in coupling pattern.

The 2D spectra for qm-4 and qm-6 (shown in figure 4.6) are much more similar than their FTIR counterparts, showing large intense peaks at 1640 cm^{-1}. There are some differences, with qm-6 showing a weak peak at 1580 cm^{-1} which may have a small amount of coupling to the main peak. There is also a small amount of coupling between the 1640 cm^{-1} peak and a peak at 1700 cm^{-1} that has no intensity in the 2DIR spectrum. The spectra for qm-3 and qm-10 also resemble this set of spectra, though with the former down-shifted slightly to 1620 cm^{-1} and the latter shifted up to 1680 cm^{-1}. The 1680 peak is the only one from the qm-10 FTIR spectrum that also appears in the 2D.

The spectrum for qm-5 is markedly different from the rest of the spectra for the vacuum simulations, showing three peaks at 1620, 1640 and 1680 cm^{-1}. There is strong coupling between the 1620 and 1640 cm^{-1} peaks and also between 1640 and 1680 cm^{-1}. This spectrum resembles that of the single conformations in figure 4.4, though the 1660 cm^{-1} in figure 4.4 seems to be shifted up to 1680 cm^{-1} in figure 4.6.

qm-7 is like the spectrum for qm-2 in that it appears to have only a single broad peak, whereas qm-9 has three diagonal peaks with some weak coupling between the major peak at 1680 cm^{-1} and the smaller peak at 1650.

The 2DIR spectrum for qm-8 is much like the FTIR spectrum in that it

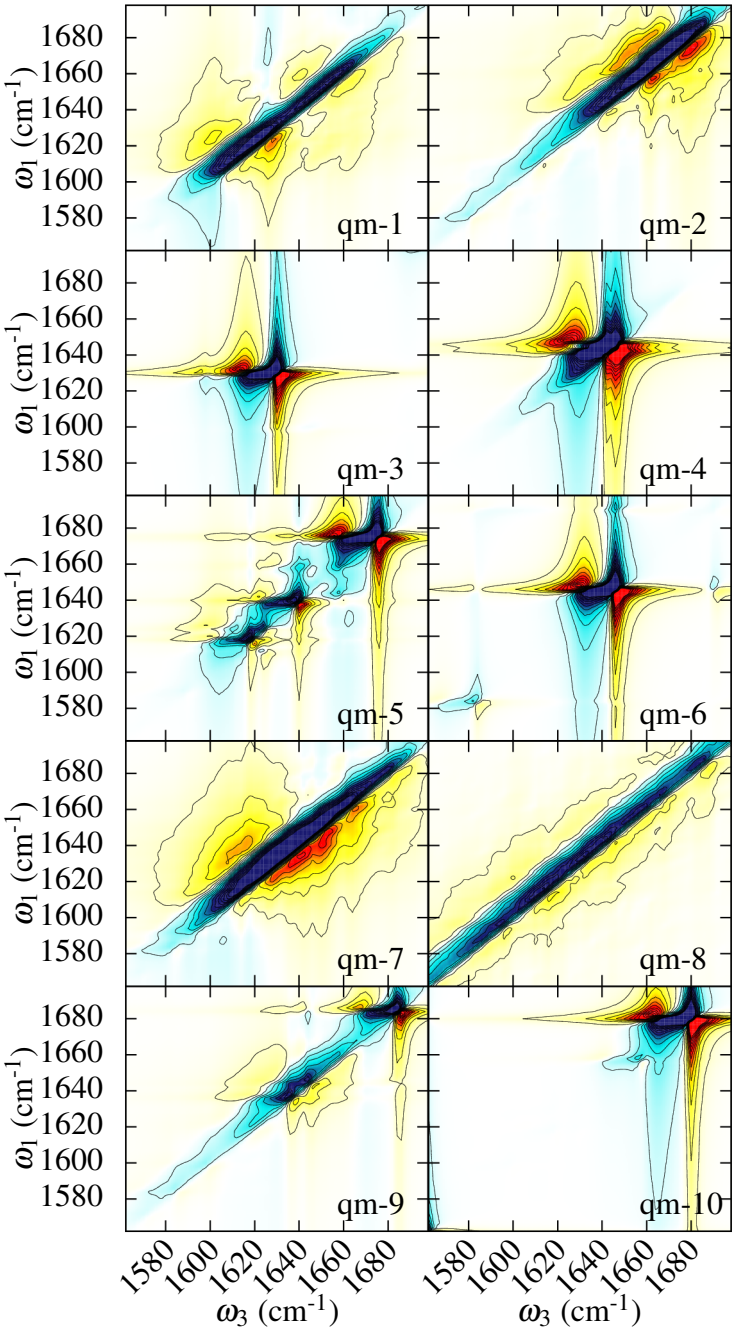


Figure 4.6: Purely absorptive 2D spectra calculated in SPECTRON from vacuum simulations of enkephalin. Colour scale runs from blue (negative) to red (positive).

shows a very broad, relatively weak band spanning around 200 cm^{-1}. This might be attributable to the conformational diversity of the extended confirmation in vacuum.

The rest of the spectra in this set are all broad bands rather than intense peaks, resembling their FTIR counterparts in figure 4.5. They are distinguishable in the 2DIR spectrum by their off diagonal components. qm-1 seems to display the two peaks as shown in figure 4.5. The same is true for qm-9, though the second peak at 1680 cm^{-1} is much more intense in the 2DIR spectrum than in the FTIR, distinguishing it from the spectrum for qm-7.

Figure 4.7 shows the FTIR spectra calculated from the simulations in explicit TIP3P water. There appear to be two broad categories of spectra in this set: those that can be said to be a single broad peak and those that show two clear peaks. In the former category are conformations qm-1, qm-6, qm-8 and qm-10, though the latter three could potentially be said to have two peaks.

Conformations qm-7 and qm-9 are differentiated in these spectra by the fact qm-9 shows two clear peaks between 1650 and 1700 cm^{-1}, but also a broad band of less intensity centred at around 1550 cm^{-1}. Interestingly, qm-5 shows a similar pattern to the qm-9 spectrum, though there is a much more intense peak at 1675 cm^{-1} and the lower frequency band is less broad and shifted up to centre on 1600 cm^{-1}. The spectrum for qm-5 only somewhat resembles that seen in figures 4.3 and 4.5, though the three peaks are just about visible.

qm-4 and qm-6 are somewhat similar for this set of simulations, with both displaying an intense peak, though for qm-4 this is at 1650 cm^{-1} and for qm-6 near 1675 cm^{-1}. The spectrum for qm-4 also shows another peak at almost 1700 cm^{-1} which may be there in the qm-6 spectrum but appears to be shifted outside the range of the calculation. qm-6 also has some intensity on the shoulder of the 1675 cm^{-1} peak, stretching down to 1600 cm^{-1}.

Again the extended conformation of qm-8 shows a relatively broad spectrum, though it can be divided into two major peaks, which are also visible

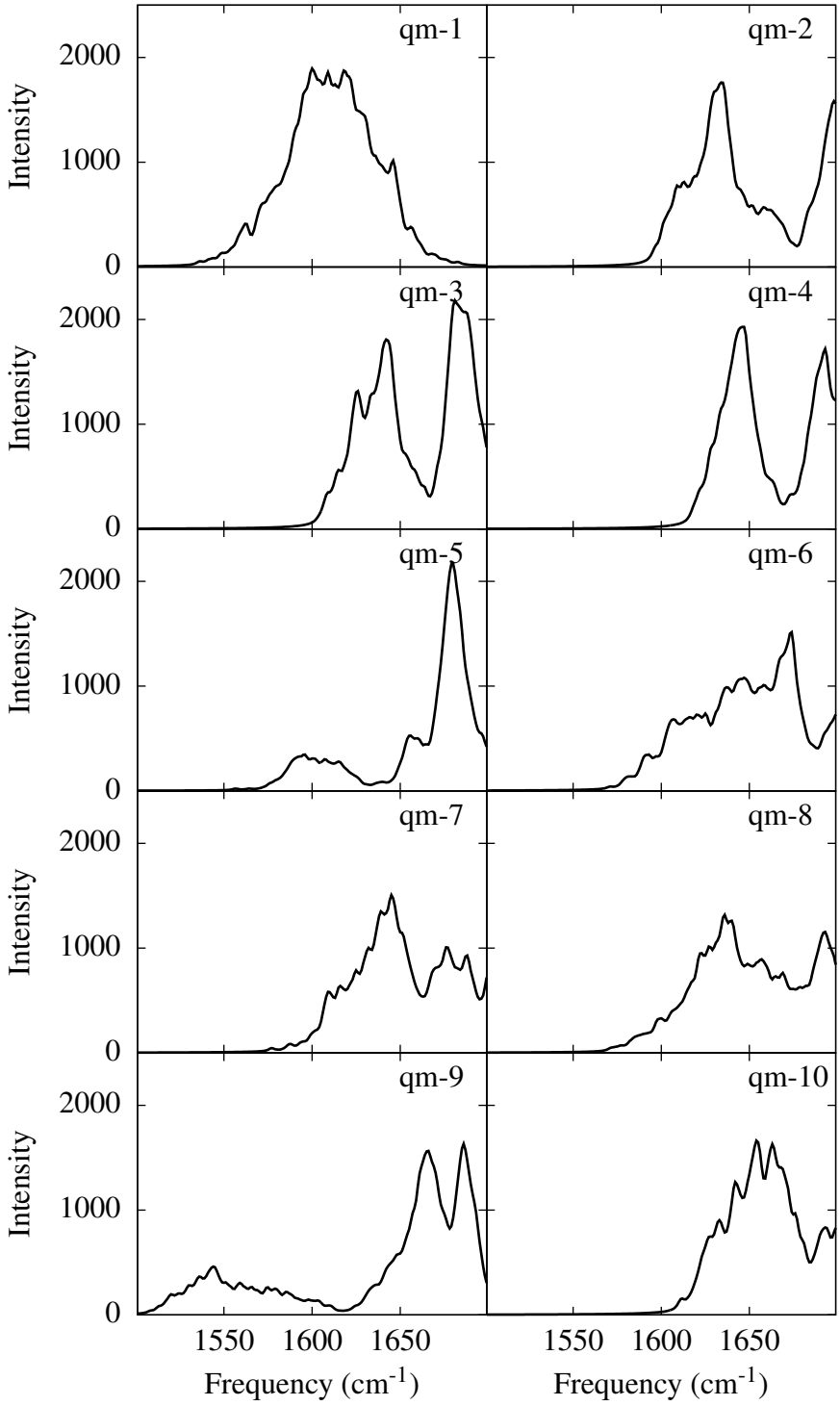


Figure 4.7: Linear absorption spectra calculated in SPECTRON from simulations of enkephalin in TIP3P water.[106]

in the qm-3 spectrum. The qm-3 spectrum is less broad and more intense than the qm-8 spectrum, and also has a third smaller peak on the shoulder of the peak at 1650 cm^{-1}. This suggests that in water these two extended conformations may have very different behaviour, potentially influence by the hydrogen bonds between the peptide and the surrounding waters.

The 2DIR spectra shown in figure 4.8 are much more similar for this set of simulations than in the FTIR spectrum. These spectra can again be grouped into those with single peaks and those with two definite peaks. Conformations qm-1, qm-3 and qm-10 can definitively be said to have a single peak in these spectra, with the rest having two clearly visible peaks. There doesn't appear to be any significant coupling between diagonal peaks in any spectrum from this simulation set.

It is interesting that the spectra for qm-2 and qm-4 are so similar, even in the shape of the off-diagonal component of the peaks at around 1640 cm^{-1}. The spectra for these two conformations haven't shown any particular similarities in previous spectra, though it is possible that in water the two different starting structures adopt similar conformations. The 2DIR spectrum for qm-6 has a peak with a similar shape at around 1690 cm^{-1} but appears to be much broader. qm-6 also has the tail expanding down to 1540 cm^{-1}.

The peak at 1680 cm^{-1} in the qm-5 shows more intensity than the peaks seen in the other spectra for this set, particularly in the off diagonal component. There may be some coupling between this peak and the one at 1600 cm^{-1} but it is difficult to say whether there is actually a cross-peak or if it is due to broadening of the large peak at 1680 cm^{-1}. If there is a cross-peak there then the coupling between the two peaks is relatively small. This spectrum is the first for qm-5 that doesn't appear to have three peaks in it.

The spectra for conformations qm-7 and qm-9 are largely differentiated by the broad peak between 1520 and 1600 cm^{-1} in the qm-9 spectrum. The shapes of the other peaks are very similar in both spectra, though in the qm-9 the two

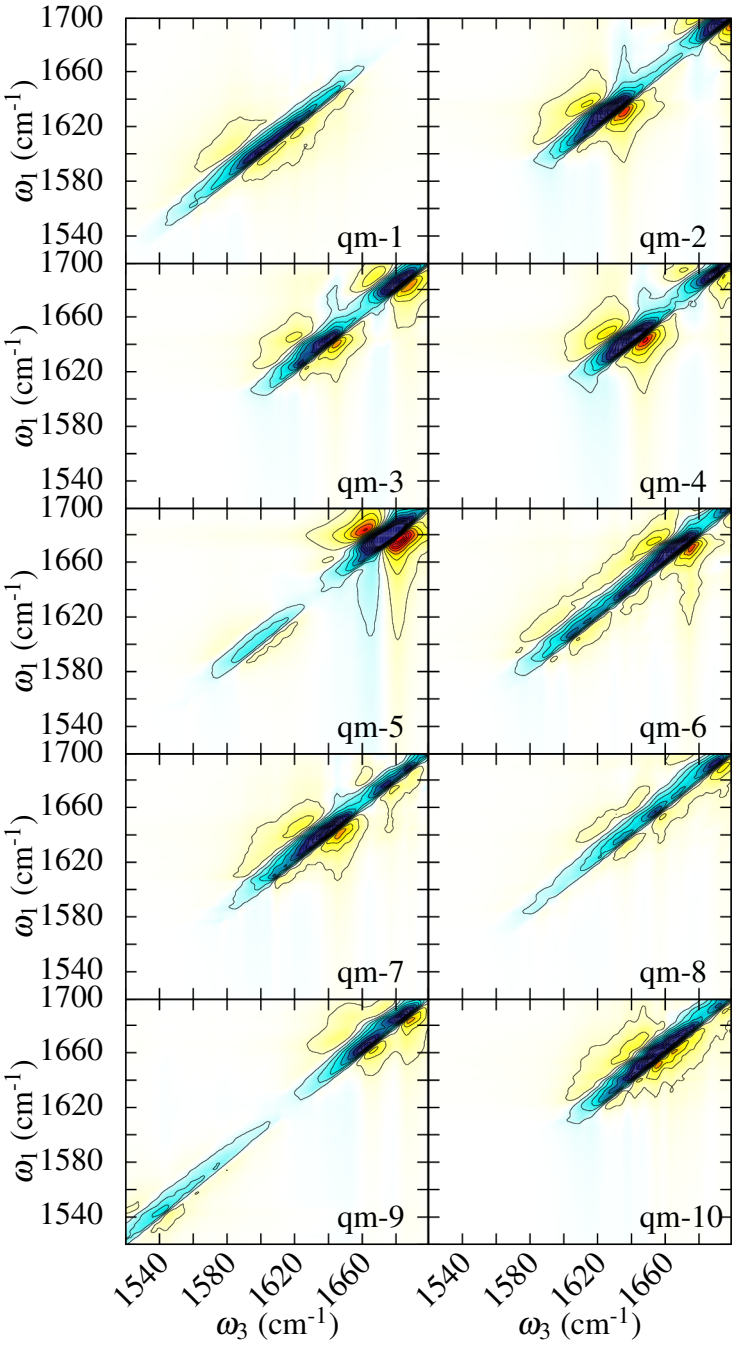


Figure 4.8: 2DIR spectra calculated in SPECTRON from simulations of enkephalin in TIP3P water.[106] Colour scale runs from blue (negative) to red (positive).

peaks are closer together and shifted towards 1700 cm^{-1}.

The spectra calculated from the TIP3P explicit water simulations show almost almost no coupling between the peaks in the spectra–there is possibly some coupling between the two peaks in qm-5 but its difficult to tell whether this is homogeneous broadening of the peak at 1680 or actual coupling. One possible explanation for this is that hydrogen bonds with water molecules preferred and preventing intramolecular and reducing the interaction—and hence the strength of coupling—between the peptide units.

The FTIR spectra for the simulations in implicit solvent are given in figure 4.9. These spectra show marked differences to the spectra shown in figures 4.5 and 4.7, indicating that solvent environment does have an effect on the spectroscopy of a small molecule, even with a relatively simple model such as is the case for the implicit solvent.

Spectra for conformations qm-3 and qm-8 are very different for this set of simulations, with the spectrum for qm-3 consisting of two sharp intense peaks and the qm-8 spectrum being the much more familiar broad signal. There is, however, a sharp peak rising out of the broader signal at 1600 cm^{-1} in the qm-8 spectrum.

The spectra for qm-4 and qm-6 also show marked differences, with qm-6 being the broad signal and qm-4 being the single sharp peak, though there is just visible a much smaller peak at 1600 cm^{-1} in the qm-4 spectrum. Like the qm-8 signal, the qm-6 spectrum has a larger peak rising out of the broader signal, this time at close to 1700 cm^{-1} and accompanied by a second peak at its shoulder.

Conformation qm-5 has sharp peaks in the FTIR spectrum, but there are now six of them rather than the three seen before. Spectra for qm-7 and qm-9 are highly similar and difficult to distinguish from one another, indicating similar dynamics in implicit solvent.

The 2DIR spectra for the implicit solvent simulations (figure 4.10) are dominated by intense sharp peaks, even for conformations whose FTIR show

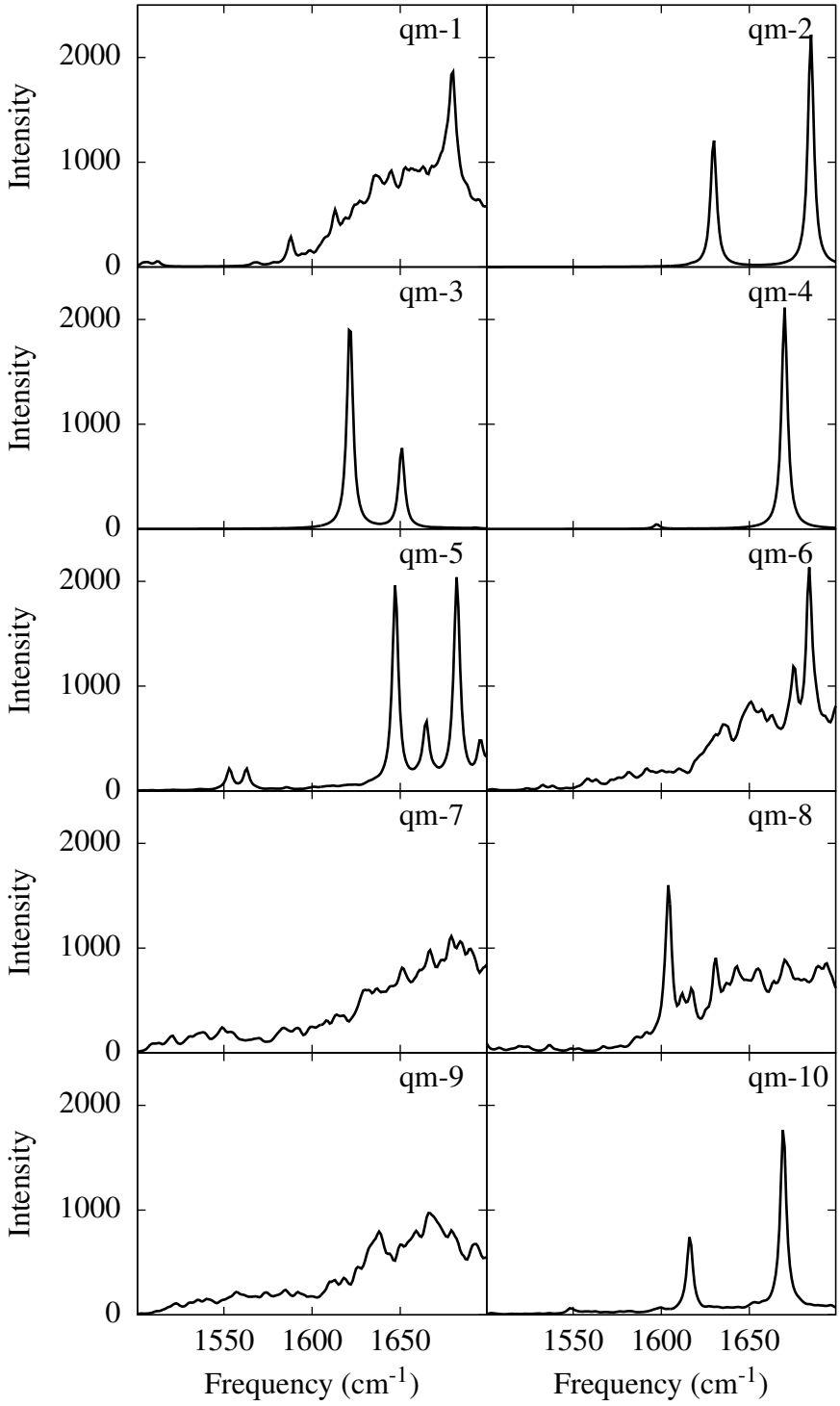


Figure 4.9: Linear absorption spectra calculated in SPECTRON from implicit solvent simulations of enkephalin.

much broader signals such as qm-1, qm-6 and qm-6.

The spectrum for qm-2 is somewhat surprising since in the FTIR it appears as two intense peaks and in the 2D is more of a continuum. There appears to be some strong coupling between peaks in the continuum for this spectrum

Conformation qm-3 and qm-8 are very similar in the 2DIR spectra for the implicit solvent simulations. They consist of two peaks with well defined off diagonal components, the lower frequency one being the more intense in both cases. The major peak in the qm-8 spectrum is at 1600 cm^{-1} and in the qm-3 it is centred at approximately 1620 cm^{-1}. It is possible there is some coupling between the two peaks but it is difficult to tell given the spread of the intense peaks. The spectra indicate similar dynamics for these conformations in the implicit solvent, something not suggested by the vacuum and TIP3P water simulations.

qm-4 and qm-6 are potentially similar in the 2DIR spectrum, appearing as single peaks, though this is at 1660 cm^{-1} for qm-4 and at 1680 cm^{-1} in qm-6. The qm-4 peak is also much more intense than in the qm-6 spectrum. There is also a potential peak in the qm-6 spectrum at 1700 cm^{-1} but it is difficult to tell since this goes outside the frequency range used for the calculation.

The 2DIR spectrum for qm-5 shows two definitive peaks at 1640 cm^{-1} and 1680 cm^{-1} that appear to be strongly coupled. The coupling pattern also indicates the presence of a third peak at 1660 cm^{-1} that isn't as visible on the diagonal. This gives a similar spectrum to that seen in figures 4.4 and 4.6.

Conformations qm-7 and qm-9 are very similar in the 2DIR spectrum as was the case in the FTIR. There are some differences in the off-diagonal components of the peaks but these spectra are qualitatively identical almost. There seems to be a small peak at 1550 cm^{-1} in the qm-7, but this is much weaker than the main signal is is difficult to pick out. The spectra for these conformations suggest highly similar dynamics in implicit solvent.

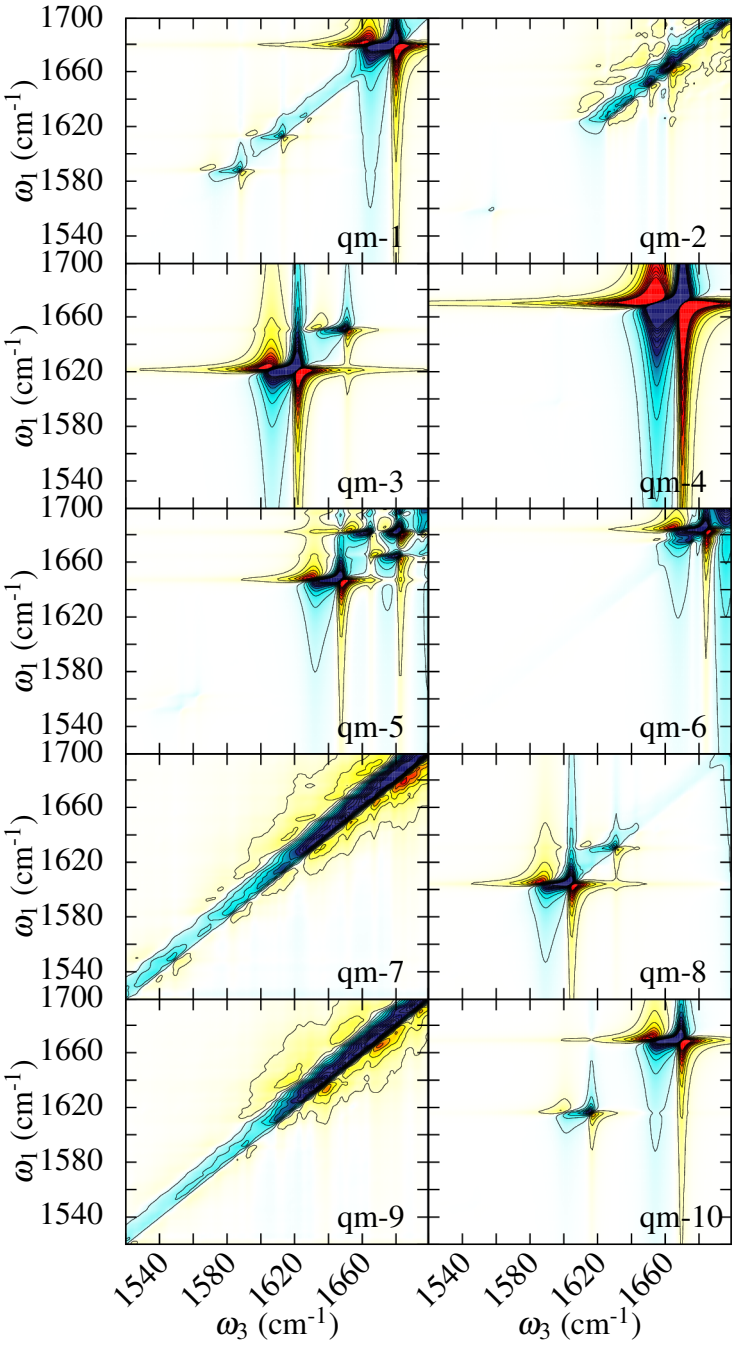


Figure 4.10: Purely absorptive 2DIR spectra calculated in SPECTRON from implicit solvent simulations of enkephalin. Colour scale runs from blue (negative) to red (positive).

4.4 Discussion

Previous theoretical studies on the IR spectroscopy of Leu-enkephalin calculated spectra for the 10 low energy conformations investigated here using both an *ab initio* method and a floating oscillator method[128]. Results from our investigation show some similarities to the spectra calculated previously, though our peaks are down shifted by approximately 50 cm^{-1} due the fact the we took 1650 cm^{-1} to be the central frequency, based on previous work modelling the amide I of NMA.[76]

Calculated FTIR spectra for the starting conformations qm-5 and qm-8 in our work show similarities to the previously calculated *ab initio* spectra. The spectrum for qm-5 has similar peak shapes and patterns to the *ab initio* calculations. The exciton based spectrum for qm-8 is almost identical in shape to the convoluted *ab initio* spectrum though again the latter has a peak at 1775 cm^{-1}. Our spectrum for qm-8 is also similar to the previous work calculated using a similar method, though some of our peaks are further apart and we have an extra peak at approximately 1610 cm^{-1}. The floating oscillator model spectrum for qm-2 is also similar to our work, though the two peaks at 1680 cm^{-1} appear reverse in our spectrum. The spectra for qm-3 are also somewhat alike in both sets of data. That some of our results show qualitative agreement with previous FTIR work for the single low energy conformers is encouraging and gives us confidence in the calculated 2DIR spectra of the same structures. That said, the real focus of the work in this chapter is the effect of dynamics on the calculated spectra.

Calculated gas phase FTIR for Leu-enkephalin show a general resemblance to experimental work done in the amide I region.[133,134] Exceptions to this statement are the calculated FTIR for simulations using qm-3, qm-6, qm-8 and qm-10 as starting points. The spectrum for qm-8 shows a broad band in the amide I region as would be expected for a peptide system, but the band is broader and weaker than those calculated for the other

simulations.This simulation shows high conformational diversity as a result of the flexible nature of the starting conformation, which could be contributing to the broad spectrum we see. Spectra for simulations based on qm-3, qm-6 and qm-10 display the opposite issue; instead of a broad band in the amide I region we see a number of sharp, intense peaks as would be expected for calculated spectra of the single conformation rather than spectra that take dynamics into account. While the simulation for qm-3 started out with an extended conformation, visualisation of the dynamics reveals that it quickly adopts a more folded conformation and remains there for the duration of the simulation time. This may be the reason for the single sharp peak in the spectrum; low conformational diversity in comparison to the simulation based on qm-8. The backbone of the peptide remains largely static throughout the simulation and since the maps used to calculate the one exciton Hamiltonian did not include side chains this may explain the single peak. The same is largely true for the simulations based on qm-6 and qm-10.

FTIR for simulations carried out in explicit TIP3P water show the broad peaks centred around 1650 cm^{-1} that we would expect from a solution phase spectrum, though some of our calculated spectra show a splitting of the main peak of uncertain origin. In addition, the spectra for qm-5 and qm-9 shows a low intensity peak in addition to the main band. The additional peak in the qm-5 spectrum lies across 1600 cm^{-1}, putting it within the expected values for an amide I band. The peak in the spectrum for qm-9, however, is shifted down to around 1550 cm^{-1}, which is slightly concerning as this lies outside the generally expected range for an amide I mode. That said, some experimental FTIR of protonated Leu-enkephalin does show a peak in this region.[133,134] Previous work assigns the band in this region as a combination of the amide II band and a tyrosine deformation.[134] Since we have not calculated the amide II band it, it is possible that our amide I band is down-shifted that far as a result of the dynamics of the tyrosine, though further work should endeavour to confirm this.

The simulations carried out in continuum solvent with the dielectric constant set to mimic acetonitrile give rise to calculated FTIR spectra with a much greater proportion of large intense peaks than were seen for the other two sets of spectra. Looking at the backbone RMSD data over time for those simulations resulting in spectra with two distinct peaks (qm-2, qm-3 and qm-10) we see the RMSD switching between two levels outside the expected variation. This might suggest the existence of two distinct but related conformations in the simulation and could explain the two distinct peaks observed in the calculated spectra. Visualisation of the trajectories seems to confirm this, with the two RMSD levels roughly corresponding to conformations where the aromatic rings are oriented together and where they are further apart.

The calculated FTIR spectrum for the simulation based on conformation qm-4 shows only one peak. Given that this simulation has the lowest mean backbone RMSD there could be a relationship between RMSD of the simulation and the observed peaks, though this is likely to be due more to the conformational diversity indicated by the RMSD rather than the RMSD itself.

The spectrum for the qm-5 simulation shows more peaks than the qm-2, qm-3 and qm-10 spectra, though this still does not appear as a broad band. The reason for the appearance of the calculated spectrum is uncertain. The simulation shows a higher backbone RMSD than for qm-4,which might be contributing to the observed spectrum. The simulation set shows a loose correlation between backbone RMSD and spectrum type, with those simulations with lower RMSDs showing fewer peak. Again, any relationship between RMSD and the calculated spectrum will be due to the underlying conformational dynamics of the spectrum rather than the RMSD per se.

Continuum solvent simulations, though less computationally expensive, are known to be less reliable than explicit solvent methods since the molecule only experiences an average effect from the electrostatics of the continuum solvent. Solvent dynamics are not taken into account. This might be reflected in

the mixed results for the calculated FTIR spectra. Future work comparing the calculated spectra for simulations in continuum solvent versus the same solvent modelled explicitly might be useful, though carrying out this work for acetonitrile would require using an alternative forcefield to the one used in this work. Alternatively a comparison between explicitly modelled and continuum water could be carried out. Also, it may be the case that implicit solvent methods are less suitable for use with the exciton method. That said, the simulations that gave calculated spectra with broad bands as opposed to large intense peaks do show some similarity to experimental spectra.

While the results seen for the implicit solvent simulations with large intense peaks could be due to the molecules remaining in one or two conformations throughout the simulation, this is not what we see when visualising the trajectories. The trajectories show high conformational diversity in the implicit solvent simulations similar to what was seen in vacuum simulations. This suggests then that, rather than only one or two conformations being present, there are one or two intense transitions that are drowning out the underlying signal and hiding any inhomogenous broadening.

The 2DIR spectra for the isolated conformers show definite coupling between modes for conformations qm-3, qm-4, qm-5, qm-7 and qm-9. There may also be a small amount of coupling in the spectrum for qm-8. There may be some relationship between hydrogen bonding and coupling pattern, but if this is the case the relationship isn't all that clear. Given that the spectra for qm-3 and qm-8 showing coupling between modes in the 2DIR spectrum despite having no hydrogen bonds in table 4.3, it is likely that other through space interactions are more important for determining coupling than hydrogen bonds.

The 2DIR spectra calculated from the simulations show the same sort of patterns as observed in the FTIR discussed previously. Many of the calculated 2DIR spectra resemble what might be expected for an experimental 2D spectrum. The exceptions to this are the 2D spectra where the corresponding

calculated FTIR show large intense peaks, which are also observed in the 2DIR spectrum for these simulations. Only a few of the 2D spectra calculated from dynamics show any definitive cross peaks. The 2DIR spectra for qm-5 in both vacuum and implicit solvent show coupling between modes. There may also be some coupling in the implicit solvent simulations for qm-2, qm-8 and qm-10, though these may simply be due to the spreading of the main peak rather than true cross peaks. Previous experimental 2DIR of Leu-enkephalin[135] focused on the interactions of tyrosine in the 1500 cm^{-1} and therefore are not directly comparable to our results. It does however, support the suggestion that peaks in the 2DIR particularly of qm-7 and qm-9 may be a result of interactions with tyrosine.

Future work could carry out multiple simulations for each starting conformation and environment and the mean spectrum calculated across the multiple trajectories. The aim of this would be to see if this improves the calculated spectra with respect to what might be expected from experiment. The major goal would be to obtain experimental 2DIR spectra for Leu-enkephalin in each of the environments studied in this chapter.

4.5 Concluding Remarks

In this work we have calculated the amide I region of Leu-enkephalin in both FTIR and 2DIR for single conformations and from MD simulations carried out in the gas phase, explicit water and continuum solvent with dielectric constant set to mimic the presence of acetonitrile. This was done by calculating the one exciton Hamiltonian using the Jansen scheme. This scheme uses the Jansen coupling maps for the nearest neighbour interaction and the transition dipole coupling scheme for long distance interactions.[112,113,123] FTIR and 2DIR spectra were then calculated from this Hamiltonian using the non-linear exciton equations in SPECTRON. The calculated FTIR spectra of the starting conformations show some similarity to previously calculated spectra.

When dynamics are incorporated we see mixed results for all three environments studied; some spectra appear as would be expected for the amide I band of a peptide (i.e. as a broad band around 1650 cm^{-1}) but others appear as a single intense peak—or multiple identifiable peaks—in the same region. This is particularly the case for the continuum solvent simulations, and we have considered the suitability of using implicit solvent methods in conjunction with the exciton method. Further work should be carried out to compare implicit solvent methods with explicit solvent simulations and the effect this has on the accuracy of the calculated FTIR and 2DIR spectrum.

For the FTIR spectra that appear as a broad band, there is a qualitative match with experimental results,[133,134] and, while previous 2DIR of Leu-enk are of the amide II region,[135] comparison with experimental FTIR leads us to consider our computed 2DIR spectra to be of reasonable quality. Where the FTIR spectra appear as instead peaks rather than broad bands, we have discussed the merits of carrying out additional simulations and calculating a mean spectrum across the different trajectories. this may have the effect of improving the accuracy of the spectrum with respect to experiment and further work should investigate this.

In addition to the further work discussed above, the investigation of calculated FTIR and 2DIR should be expanded to include the opioid receptor that is the target for Leu-enkephalin. While the calculation of the 2DIR spectrum of the receptor is currently intractable, the effect of the ligand binding on the FTIR and 2DIR spectrum of Leu-enkephalin could be investigated. In addition, the current work could be expanded to include Met-enkephalin, to investigate the effect of nature of the terminal residue on the dynamics and spectroscopy of enkephalin.

Chapter 5

Structure and Dynamics of the Enoyl-Acyl Carrier Protein Reductase in *Mycobacterium Tuberculosis*

5.1 Introduction

Tuberculosis is a serious, widespread and often fatal infectious disease usually caused by the bacterium *Mycobacterium tuberculosis*, with up to two million deaths occurring annually.[136] The disease typically attacks the lungs and is easily transmitted via respiratory fluids in the air. Since 1952,[137] the cornerstone of tuberculosis treatment has been the drug isonicotinic acid hydrazide—also known as isoniazid—due to its high activity, low cost and relatively low toxicity. It was not until relatively recently, however, that an NADH-specific enoyl-acyl carrier protein (ACP) reductase coded for by the *inhA* gene was suggested as the target for isoniazid.[138] The InhA protein forms part of the pathway for mycolic acid biosynthesis in *M. tuberculosis* and isoniazid works by inhibiting the active site within the protein.

Despite drugs such as isoniazid, tuberculosis still presents a serious issue in global health due to its prevalence and mortality. This coupled with strains that are increasingly resistant to front-line antibiotic treatments means that a better understanding of how current drugs work, how resistance arises and the discovery of novel targets for new drug development for treating tuberculosis is a high priority.

5.1.1 Function of InhA Protein and Mechanism of Action of Isoniazid

The trans-2-enoyl-acyl carrier protein reductase InhA forms an essential part of the mycolic acid biosynthesis pathway.[139,140] Mycolic acids are long chain fatty acids found in the cell walls of bacteria in the mycolata taxon, which includes *M. tuberculosis*. These fatty acids form part of a protective layer around the bacterium, leading to increased resistance to chemical damage and dehydration, preventing effective activity of antibiotic drugs. In the case of *M. tuberculosis*, the mycolic acid biosynthesis is essential for the survival and pathogenesis of the bacterium. InhA is part of the type II fatty acid synthase system and which elongates enoyl fatty acid precursors to mycolic acid. Specifically, InhA catalyzes the NADH-dependent reduction of long chain trans-2-enoyl-acyl carrier proteins (ACPs) in the mycolic acid synthesis pathway, as shown in figure 5.1. Inhibition of InhA leads to an accumulation of saturated hexacosanoic acid and results in cell lysis[141,142]

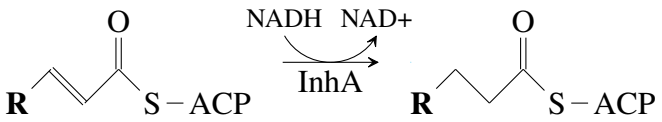


Figure 5.1: Reaction scheme for a portion of the FASII pathway, showing the InhA catalysed reduction of trans-2-enoyl-ACPs.

Isoniazid is a prodrug, which is biochemically activated through oxidation by the oxidase-peroxidase enzyme KatG. This results in the loss of the $-NH-NH_2$ functional group, and the resulting radical anion forms an adduct

with NADH. This isoniazid-NAD adduct is the biologically active ligand which binds to InhA, where it acts as a competitive inhibitor to NADH, thus halting the synthesis of mycolic acid. The structure of the isoniazid prodrug, and the biologically active isoniazid-NAD adduct are given in figure 5.2. The active form of isoniazid (the NADH adduct) is difficult to isolate and study outside of biological systems but the small molecule prodrug has been investigated using 2DIR in the solution phase.[143] This study suggested the potential for isotopic labelling of the prodrug in order to spectroscopically isolate the signal of the active form within the protein complex.

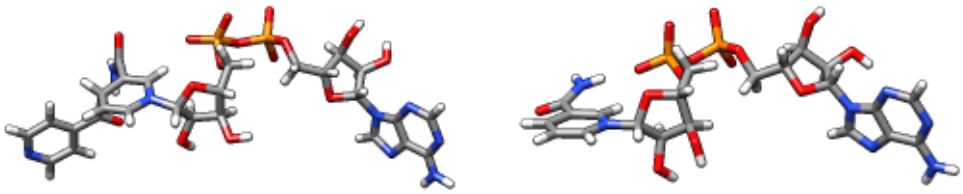
The exact mechanism of action for isoniazid turns out to be surprisingly complex, with evidence suggesting that it can also bind to KasA, a synthase condensing enzyme that forms another part of the type II fatty acid synthase system.[144,145] In addition, the way in which point mutations (usually of the active site in InhA) lead to isoniazid resistance is complicated. One mechanism of isoniazid resistance arises from a ^{94}Serine$\rightarrow$^{94}Alanine mutation, which sits within the active site of InhA. There is no significant difference in the secondary structure of wild-type InhA and the Ser^{94}Ala mutant.[146] InhA mutations dramatically reduce the binding affinity of NADH to the enzyme[147,148] whilst having only a minimal effect on the affinity of isoniazid-NAD adduct binding.[149] The mutations of the active site may permit access to a second conformational state of InhA which inhibits the action of isoniazid despite only a small decrease in affinity for the adduct. The full effect of the mutations on enzyme activity and inhibition may only be fully realised *in vivo* where protein-protein interactions between InhA and other enzymes in the fatty acid biosynthesis pathway are important.

The complexity of isoniazid's mode of action coupled with the minimal effect InhA mutation has on its binding affinity means that a more complete understanding of the structural changes that result from mutation and drug binding is of critical importance to future drug development. Particularly,

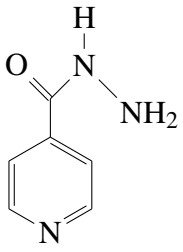
understanding the fast dynamical processes that occur during drug binding is a part of the problem that currently remains unaddressed. Measurement of these fast time scale fluctuations and hydrogen bond related dynamics is difficult experimentally, but can be observed *in silico* via MD simulations. The study of all atom MD simulations and calculated spectra could help shed light on the complex process by which isoniazid inhibits InhA and how mutations in the protein mediate isoniazid resistance. Isolating residues that have significant interactions with either NADH or the isoniazid-NAD adduct in simulations could inform experimental studies by providing targets for mutation or isotopic labelling. Understanding these interactions more fully will inform the development of new therapeutic compounds.

5.1.2 Variants of InhA: Mutants and Cofactors

Four different variants of the enoyl-ACP(CoA) reductase (InhA) from *M. tuberculosis* were studied using MD simulations and network models (as outlined in section 3.3) of the one exciton Hamiltonian in order to investigate the effect of mutation and ligand binding on the structure and dynamics of the



(a) **(b)**



(c)

Figure 5.2: Structures of the ligands (a) the isoniazid-NADH inhibitor from 1zid, (b) NADH (taken from 2aq8), and (c) chemical structure of isoniazid. Grey=carbon; red= oxygen; blue= nitrogen; white= hydrogen.

protein. The four variants chosen were as follows: the wild-type protein in complex with NADH (PDB code 2aq8[150]); a ^{2}Threonine$\rightarrow^{2}$Alanine mutant in complex with the biologically active isoniazid-NADH adduct (PDB code 1zid[139]); the isonizid resistant ^{94}Serine$\rightarrow^{94}$Alanine mutant in complex with NADH (PDB code 4dti[151]); and a ^{2}Threonine$\rightarrow^{2}$Alanine, ^{94}Serine$\rightarrow^{94}$Alanine mutant in complex with the isoniazid-NADH adduct (PDB code 2nv6[152]). Note, variant 2nv6 is given as an Asp^{2}Ala mutant in the Protein Data Bank but examination of the PDB structure file itself reveals it is in fact a Thr^{2}Ala mutant. The variants and their respective ligands are summarised in table 5.1. InhA exists as a homotetramer *in vivo*, but the crystal structures are for the monomeric unit only. The work on this chapter focusses on the dynamics of the monomer units and protein-ligand interactions rather than protein-protein interactions between the subunits.

PDB Code	**Mutation(s)**	**Ligand**	**First Residue**
2aq8	None (Wild type)	NADH	^{3}Gly
1zid	^{2}Thr $\rightarrow$ ^{2}Ala	Isoniazid-NADH inhibitor	^{2}Ala
2nv6	^{2}Thr $\rightarrow$ ^{2}Ala, ^{94}Ser $\rightarrow$ ^{94}Ala	Isoniazid-NADH inhibitor	^{2}Ala
4dti	^{94}Ser $\rightarrow$ ^{94}Ala	NADH	^{3}Gly

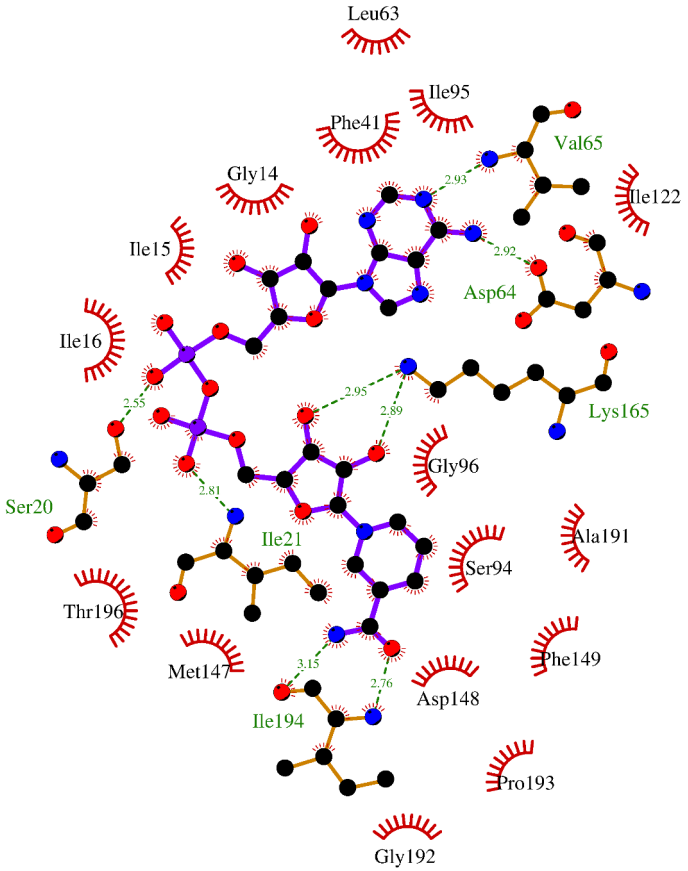
Table 5.1: Mutations and ligands present in the four variants of interest, along with the first residues present in each PDB file.

Due to varying crystallisation methods, the PDB files for the chosen variants are incomplete, missing either the first one or two residues. The first residue in each variant is given in table 5.1. This was accounted for in the construction of the network models and consequently the nodes labels in the models correspond to the residue numbers in the system. Due to this mismatch in the crystal structures, it is not possible to investigate the effect of the point mutation at residue two on the dynamics and network models of InhA.

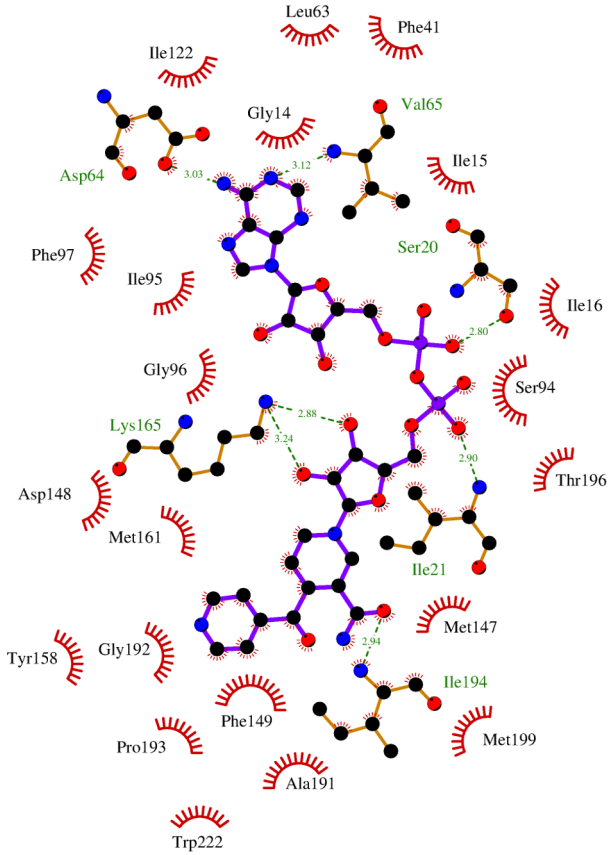
It is likely, given their location in the tertiary structure, that the first few residues in the protein have little importance with regards to primary influence

on the active site of the enzyme. Of greater concern, perhaps, is the need to understand secondary effects on the active site. Residues further away from the binding pocket may have no direct influence on the active site, but may alter the secondary structure of the protein and this may ultimately have an effect on the active site. Visualisation of the PDB structures showed that the first three residues in the protein form a highly flexible tail on the surface of the protein and are not part of any particular secondary structure elements. The decision was made to proceed with the crystal structures as any effect these residues may have on the active site is likely to be minor and other effects more likely to dominate. However, it should be noted that if these residues do have an effect on active site geometry we are unable to investigate it in this study. Comparison with experimental results may give some indication of whether these residues have any influence on the active site.

For the purposes of clarity, the four variants will be referred to as follows: wild-type (2aq8), inhibited wild-type (1zid), mutant (4dti), and inhibited mutant (2nv6). The inhibited wild-type is technically the $Thr^{2}Ala$ mutant with the inhibitor present, but since the first and second residues are missing from the starting structure for the wild-type (see table 5.1) we are treating 1zid as the wild-type protein. Strictly speaking the inhibited versions are only inhibited in the holo-protein simulations, but since the starting structures for all simulations have the ligands present this is the naming convention we will use in the rest of the chapter.



(a) Wild-type bound to NADH (from PDB 2aq8)



(b) Inhibited wild-type (from PDB 1zid)

Figure 5.3: LigPlot+[153] diagrams of a) the active site of the wild type structure bound to the biologically active NADH-isonizaid adduct and b) the active site of the wild-type bound to the inhibitor isoniazid. Residues depicted in in red show hydrophobic interactions with the ligand.

5.2 Modelling the Conformational Dynamics and Spectroscopy

5.2.1 MD Simulations

MD simulations were carried out on all four variants of InhA both bound and unbound to their respective ligands. For the simulations of the apo-form, the structures were minimised using the steepest descent algorithm and solvated in a truncated octahedron of TIP3P water[106] ensuring a minimum distance of 10 Å between the protein and the edge of the water box. 50 ps of dynamics was carried out with the protein atoms restrained in order to "soak" the peptide in the water molecules. 5 ns of equilibration was carried out in NVT ensemble at 300 K using a leapfrog algorithm. Periodic boundary conditions and Particle Mesh Ewald (PME) electrostatics[130] were used with a cut off of 10 Å to allow the Verlet cut off scheme to be used. The Verlet scheme allows GPU acceleration of the simulation in GROMACS. Heavy atom-hydrogen bonds were constrained using the LINCS algorithm[101,102] to allow a 2 fs time step. Production dynamics were carried out using the same parameters for 20 ns, with snapshots saved to the trajectory every 1 ps.

Set up of the ligand simulations was carried out with the CHARMM-GUI[88,154,155] using the CHARMM36 forcefield with the CMAP correction.[93,94] Parametrisation of the isoniazid ligand (see appendix B) was done using the CHARMM Generalised Forcefield[156] and the SwissParam server.[157] For the wild-type structure bound to NADH (2aq8) and the mutant bound to NADH (4dti), the N-terminal glycine was patched using the GLYP patch. All four structures were solvated in a truncated octahedron of TIP3P water, ensuring a minimum of 10 Å between the protein and the edge of the water box.

The solvated structures underwent 30000 steps of minimisation in CHARMM using the steepest descent and the Adopted Basis Newton-Raphson

schemes. This was followed by a 10000 step minimisation in NAMD[90] using the Conjugate Gradients scheme and a 50 ps dynamics run carried out to ensure no bad contacts between the ligand. The simulation was transferred to GROMACS[91,92] in order to take advantage of the GPU acceleration. 5 ns of equilibration was carried out under the NVT ensemble at 300 K using a leapfrog algorithm. The other simulation details were the same as for the simulations of the apo-forms.

12 independent simulations were carried out for each variant in both apo- and holo-forms: wild-type, inhibited wild-type, mutant and inhibited mutant i.e. from each starting PDB 2aq8, 1zid, 4dti and 2nv6 respectively. The mean result across the 12 simulations for each variant is what is presented in the results section.

5.2.2 Simulation of IR Spectroscopy

The Hamiltonian matrices and transition dipoles obtained using the CHARMM27 atomic partial charges were used to compute the shift of the diagonal elements of the exciton Hamiltonian and the Jansen scheme.[112,113,123] For further details see section 4.2.2. The Hamiltonian matrices and transition dipoles for the simulations were calculated from snapshots that included the simulation water box, so as to incorporate the electrostatic effect of the solvent. The spectra and network models in this chapter were calculated using uncorrected frequencies. FTIR spectra were calculated from the Hamiltonian and dipole matrices using SPECTRON[117] and the parameters given in table 5.2.

5.2.3 Visualisation of Coupling between Amide Units using Network Models

In this chapter, network models derived from the one exciton Hamiltonian matrices of the systems of interest are used to investigate the couplings between residues and approximate the information that could be gleaned from a 2DIR

Parameter	Setting
Lorentzian Linewidth (cm^{-1})	2
Initial Frequency (cm^{-1})	1500
Final Frequency (cm^{-1})	1800
Number of Frequencies	300

Table 5.2: SPECTRON input parameters for calculating linear absorption spectra of InhA. The Lorentzian linewidth controls inhomogeneous broadening of the signal; the signal is computed in the range defined by the initial and final frequencies, i.e., 1500 to 1800 cm^{-1}. 1800 cm^{-1} was chosen as the final frequency after initial calculations showed the amide I band extending slightly above 1700 cm^{-1}.

spectrum. While 2DIR spectra of proteins provide insight into the dynamics and three-dimensional structure, calculations of large proteins are still intractable with current computational methods. In the case of the exciton method, which utilises the one- and two-exciton block of the Hamiltonian matrix, the computational cost of calculating 2D spectra grows rapidly with system size. The number of two-quantum states scales as $N(N+1)/2$ and the diagonalisation of the matrix scales as $O(N^6)$,[158] where N is the number of oscillators in the system. Approximations within the NISE method can reduce this scaling to $O(N^3)$[159] but this still means 2DIR calculations are largely still intractable for systems larger than 150 residues. Since InhA is a 268 residue protein it is unsuitable for calculations involving the two-exciton Hamiltonian.

Network models can be used to probe the hidden structure of a complex system by presenting large sets of data in an intuitive, visual manner.[160–162] Network models have already been used to model systems with various applications in chemistry and biology, and previous work has been done with network models of proteins to identify groups of coupled oscillators.[163] The calculation of the one-exciton block of the Hamiltonian matrix is much simpler than the two-exciton block, and network models were built from the couplings between residues present in the one-exciton Hamiltonian. The aim of the work is to use the network models—and the couplings themselves—to identify areas of interest for investigation in experimental 2DIR and to support the

information already gained from experiment. We also wanted to see whether the network models from these simulations would show any correlation with previous work carried out on InhA.

The network models were generated using the couplings between the residues from the Hamiltonian matrix generated as outlined in the previous section. The residue numbers were used as the nodes of the model, and the couplings themselves used as the weight of the edges between the nodes. The layout of the network was found using the Force Atlas algorithm as implemented in Gephi.[164]

Forces are assigned to the nodes (in this case, each individual peptide unit) and edges (the couplings between peptides) of a graph. Spring-like forces based on Hooke's law are used to attract nodes towards each other, while repulsive forces similar to electrically charged particles based on Coulomb's law are used to separate the nodes. The parameters for the Force Atlas algorithm are given in table 5.3. The resultant networks were filtered to ensure that no couplings of less than 2 cm^{-1} were visible and all visible nodes had at least one edge. The model was coloured in a RGB rainbow scheme from the N-terminus to the C-terminus to reflect the colouring scheme of the VMD cartoon structure in figure 5.4a.

Parameter	**Setting**
Inertia	0.1
Repulsion Strength	200
Attraction Strength	10
Maximum Displacement	10
Auto-stabilise function	On
Auto stabilise strength	80
Auto stabilise sensibility	0.2
Gravity	30
Speed	1

Table 5.3: Force Atlas Algorithm parameters used to generate the network model layout. These are the default parameters for the Force Atlas algorithm in Gephi[164] and were left as is because these parameters quickly produced a network graph that resembled the tertiary structure of each variant as shown in figure 5.4

5.3 Results

5.3.1 Analysis of Network Model and Exciton Hamiltonian Elements

5.3.1.1 Crystal Structures

Figure 5.4 shows that the network models closely resemble the structure of the protein when the layout is generated using the Force Atlas algorithm. Figure 5.4a shows a VMD cartoon of the structure and figure 5.4b shows the Force Atlas generated network model. Both figures are colour coded in a similar fashion and the α-helices can be picked out as clusters of nodes with many short range interactions between them. The β-sheet is more difficult to see, but is apparent in the long range interaction figure in 5.4c as a number of differently coloured nodes with interactions between them. The similarities between the PDB structures and the network model indicates that the parameters used in the models are of sufficient quality to draw conclusions about the structure and dynamics of the protein from the models. Figure 5.4 also indicates that α-helices are best visualised when all interactions are included in the network model, and β-sheets are clearer when only longer range interactions are considered.

The network models for the other three variants of InhA showed high similarity with the model of the inhibited wild-type. The four structures have only point mutations to differentiate them and at low temperatures neither these mutations nor the different ligands result in large differences between the structures. The high structural similarity is shown in the RMSD data (table 5.4), with the all-atom RMSD for the crystal structures being less than 1 Å.

Visualising the RMSD per residue (figure 5.4) between each variant indicates that the largest differences exist between the inhibited wild-type (1zid) and the other three variants. The wild-type (2aq8), mutant (4dti) and inhibited mutant (2nv6) all show a high degree of similarity. The inhibited wild-type

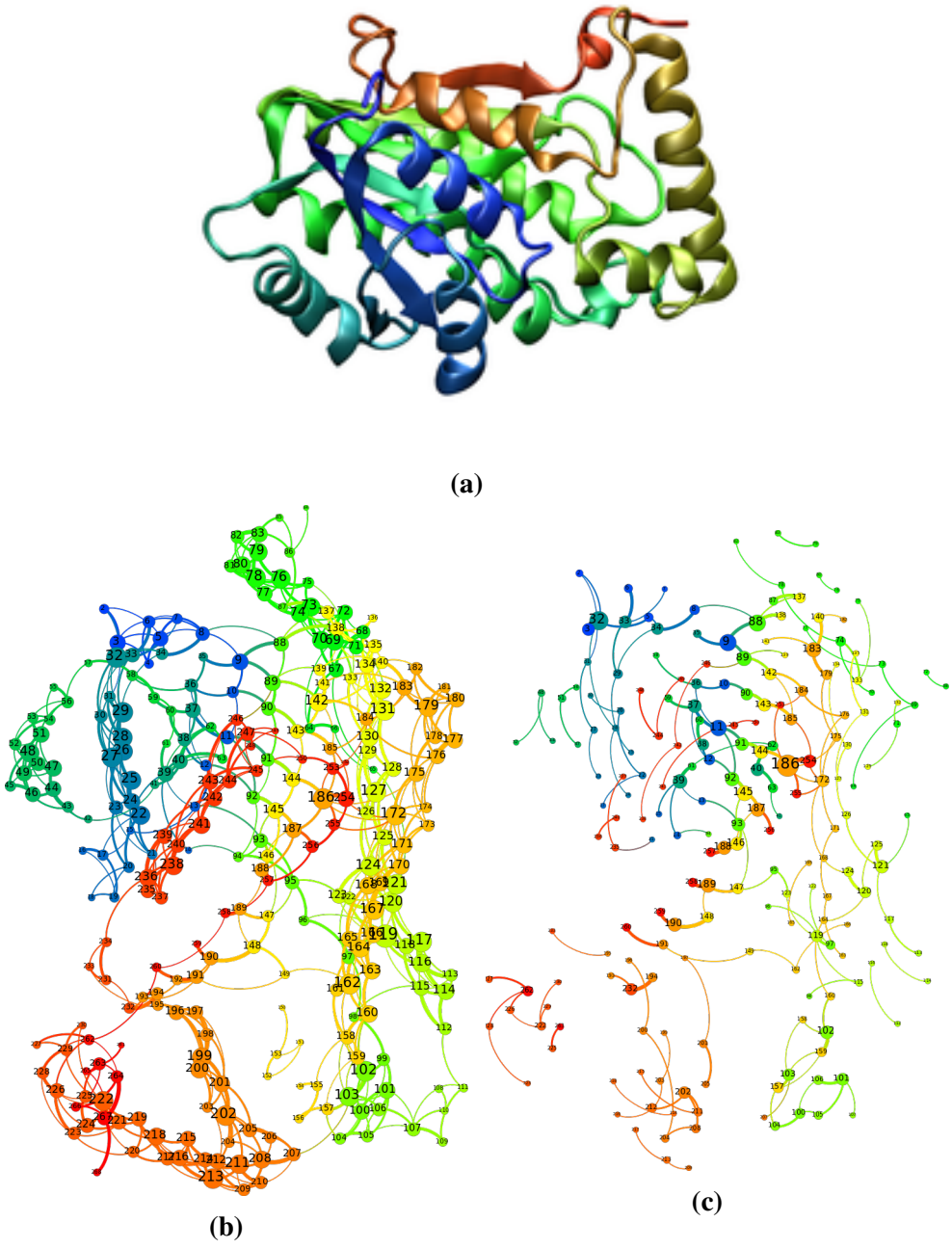


Figure 5.4: a) VMD[165] cartoon of inhibited wild-type structure (taken from PDB 1zid—ligand not shown). Protein is coloured red to blue from the N-terminus to C-terminus. b) Network model extracted from the crystal structure with ligand included. All interactions included in the model. c) Network model with long range interactions only.

plots show a single residue with an RMSD of 3.5 Å, much larger than that seen for the rest of the structures. Interestingly, this residue is not residue 94, the site of the point mutation conferring isoniazid resistance, but residue 104 which is a glycine residue not in the active site. This residue is on the surface of the protein and the large RMSD is likely due to the inherent conformational flexibility of the residue and its position on the surface where it is exposed to solvent effects.

	RMSD (Å)		
Structures Compared	C_α	Backbone	All-atom
Inhibited wild-type (1zid), Wild-type (2aq8)	0.5	0.5	0.8
Inhibited wild-type (1zid), Inhibited mutant (2nv6)	0.4	0.4	0.7
Inhibited wild-type (1zid), Mutant (4dti)	0.4	0.4	0.7
Wild-type (2aq8), Inhibited mutant (2nv6)	0.3	0.3	0.5
Wild-type (2aq8), Mutant (4dti)	0.3	0.3	0.3
Inhibited mutant (2nv6), Mutant (4dti)	0.1	0.1	0.1

Table 5.4: Mean RMSD between the crystal structures of InhA.

The couplings between residues for the crystal structures were also similar; the 20 largest interactions were between 8 and 9 cm^{-1} and the differences between variants were small. The exception to this was the ^{3}Gly, ^{32}Gln interaction which was an order of magnitude larger than most of the other differences between the inhibited wild-type (1zid) and the others. This was attributed to the fact ^{3}Gly is the first residue for two of the variants and being a glycine is conformationally flexible. Examination of the PDB files indicated that in the inhibited wild-type (1zid) this residue was oriented in one direction and in the opposite direction for the other three, which could explain the large change in coupling value. It is primarily short range interactions that have large coupling values, and there are few active site residues with large couplings. Since the crystal structures contain the ligands it is to be expected that a similar result will be seen for the holo-protein simulations and there may be increased coupling between these residues in the apo-protein simulations.

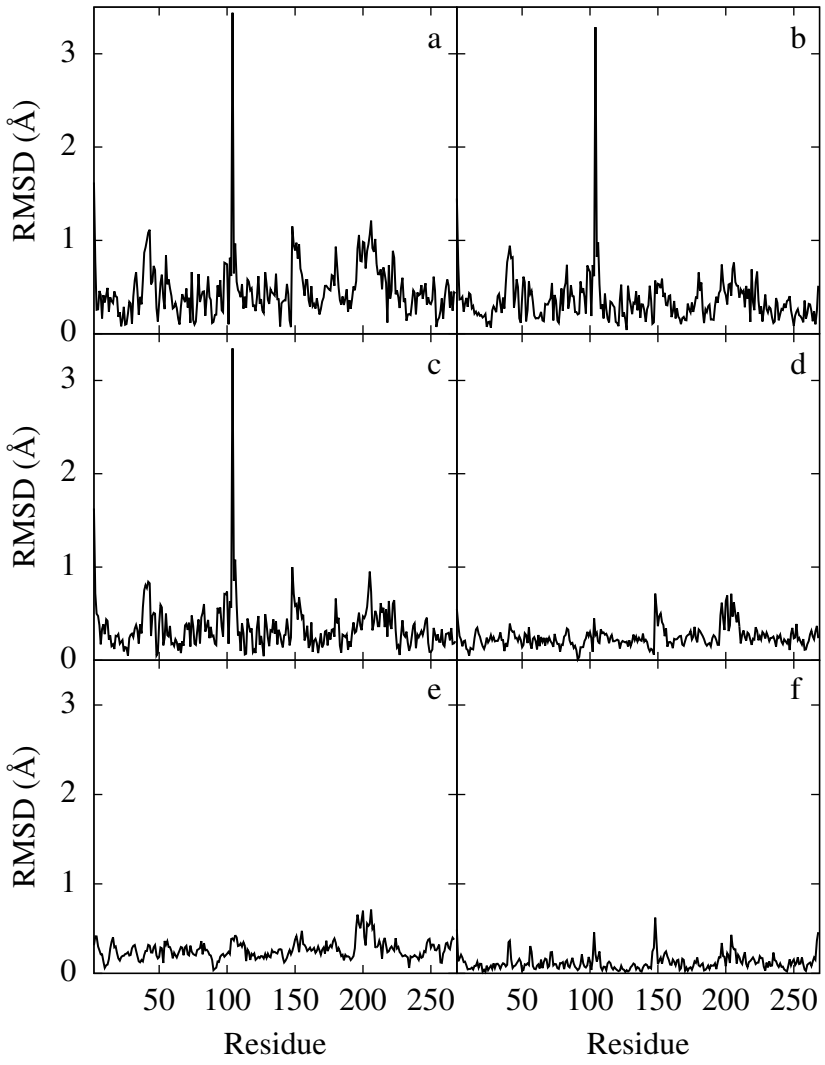


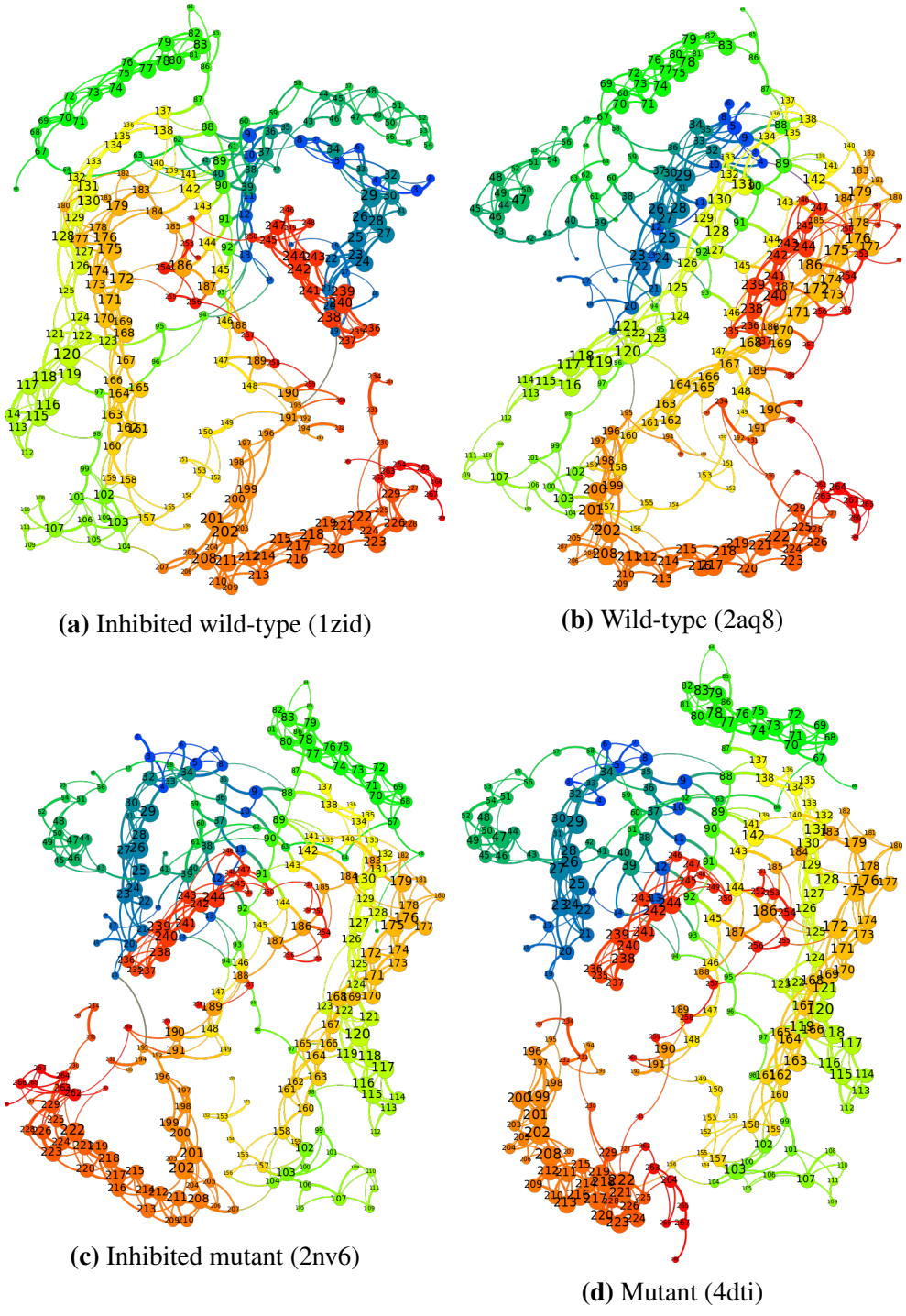
Figure 5.5: RMSD per residue for α carbons of the crystal structures of InhA. a) Inhibited wild-type/wild-type (1zid/2aq8); b) Inhibited wild-type/inhibited mutant (1zid/2nv6); c) Inhibited wild-type/mutant (1zid/4dti); d) wild-type/inhibited mutant (2aq8/2nv6); e) wild-type/mutant (2aq8/4dti); f) inhibited mutant/mutant (2nv6/4dti).

5.3.1.2 Apo-protein

Figure 5.6 shows the network models of all four variants of InhA that have undergone 5 ns of dynamics. The models were constructed from the mean exciton Hamiltonian calculated from the final frame of 12 independent simulations. Figure 5.5 shows that the networks are broadly similar, and the same is observed in figure 5.6 and 5.7. This indicates that the four variants have undergone similar dynamical processes in the absence of the binding ligand and the mutations do not cause large disruptions in the overall structure of the protein.

Table 5.5 shows some numerical values of the difference data shown in figure 5.6 displaying the largest differences in coupling between variants only. The largest differences in coupling are between the inhibited wild-type (1zid) and the other three variants. The largest difference in coupling between the inhibited mutant (2nv6) and the mutant (4dti) is 3 cm^{-1}, suggesting that these share the same structure and dynamics. The wild-type (2aq8) and inhibited wild-type (1zid) might be expected to show coupling differences on the same order as the inhibited mutant (2nv6) and the mutant (4dti) but this is not the case. These structures show a large difference in coupling between residues 149 and 150 and the other coupling differences in the table are larger than those seen for the Ser^{94}Ala mutants. This could be due to the effects of the mutation at residue 2, though no interactions with this residue are seen in the data.

The 10 largest interactions from each network model are given in table 5.5 and again indicate broad similarities between the models. Long range interactions appear much more frequently than when looking at the crystal structures, suggesting that the coupling of amide carbonyls between relatively distant residues may have an important effect on dynamical processes. Long range interactions tend to have larger coupling values than in the crystal structures, but without the ligands, the point mutations in the active site do not have a large influence on the overall dynamics of the protein. In the absence of



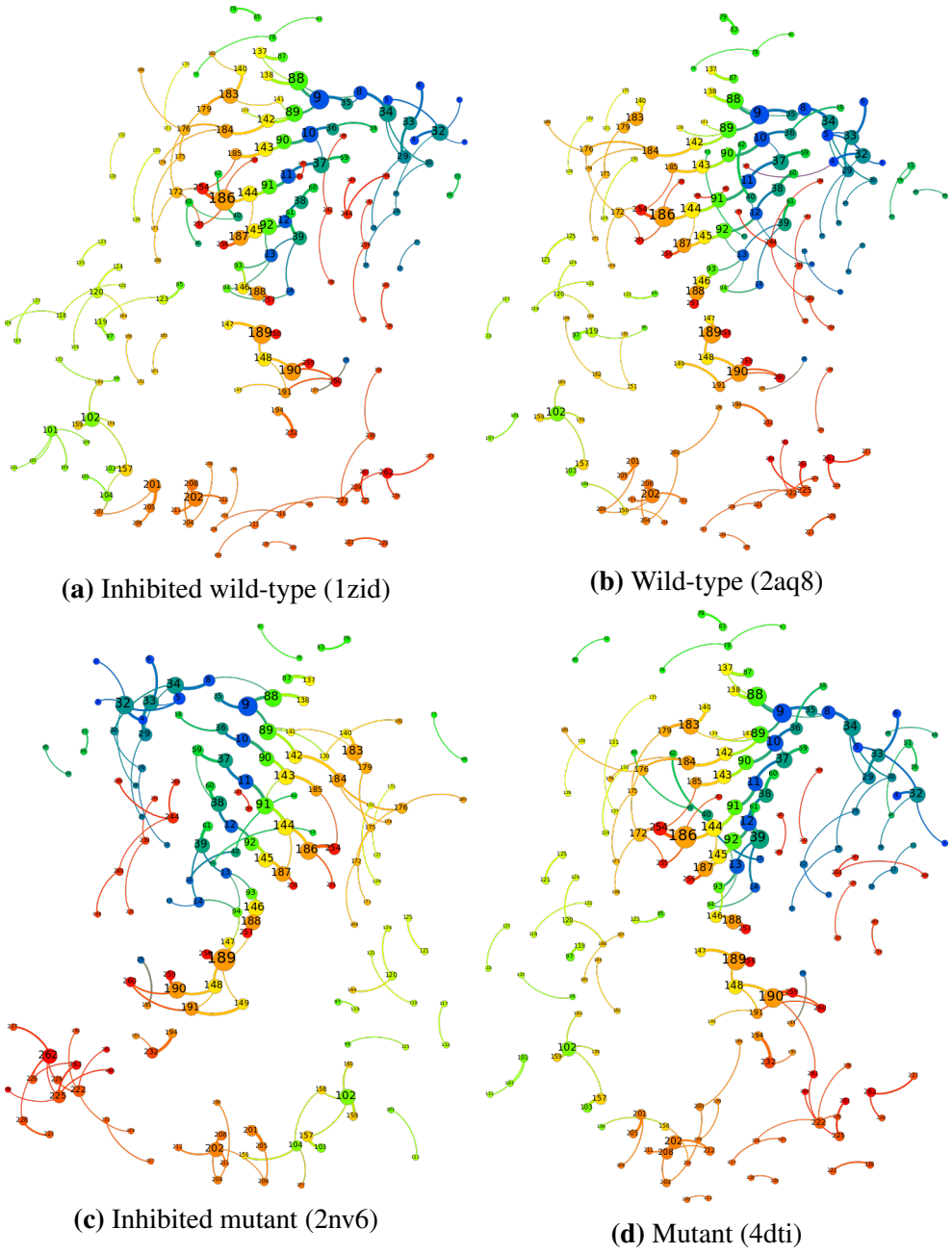
(a) Inhibited wild-type (1zid)

(b) Wild-type (2aq8)

(c) Inhibited mutant (2nv6)

(d) Mutant (4dti)

Figure 5.6: Network models extracted from systems that have undergone equilibration with all interactions included in the model.



(a) Inhibited wild-type (1zid)

(b) Wild-type (2aq8)

(c) Inhibited mutant (2nv6)

(d) Mutant (4dti)

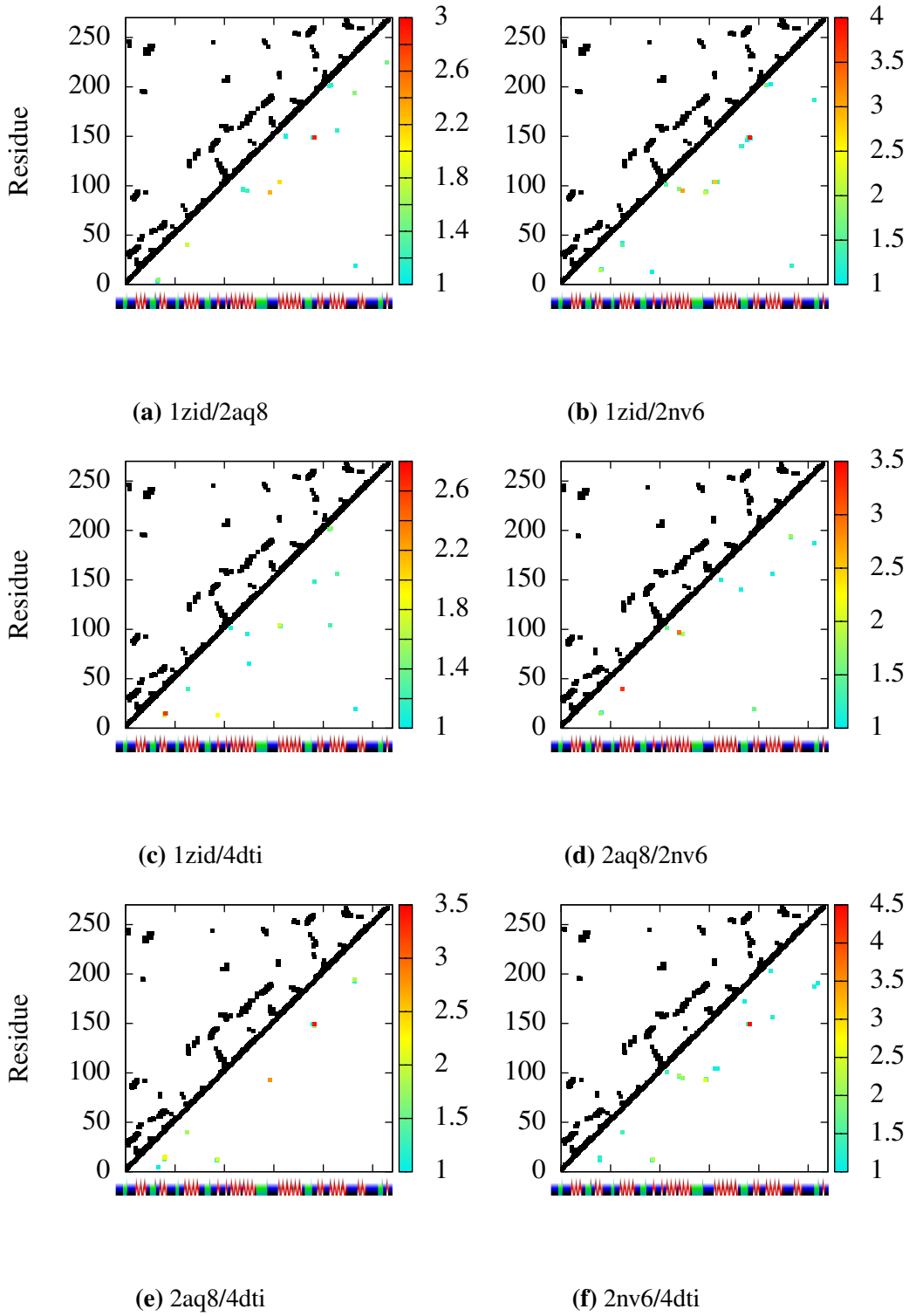
Figure 5.7: Network models extracted from systems that have undergone equilibration with only $i, i\pm5$ interactions included in the model.

Interaction (Residue)	Structure PDB code *1zid*	*2aq8*	*2nv6*	*4dti*
2, 3	7			
8, 34	7	7	7	7
9, 88		7	7	
10, 90	7	7	7	
11, 91	7	7	7	7
38, 60				7
39, 61			7	7
52, 54				7
82, 83	9	8	8	8
87, 137			7	7
88, 138	7	7	7	7
91, 144		7	7	7
92, 145	7			7
113, 116	7	7		
118, 121		7		
141, 142	8	9		
143, 185	7			
146, 188			7	

Table 5.5: The 10 largest interactions extracted from each network and compared across all equilibrated structures. The coupling, J, is given in wavenumbers (cm^{-1}). Long range interactions highlighted in bold. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

the binding ligand the carbonyl stretches of the active site residues do not appear to be coupled to any significant extent, suggesting that in the absence of a ligand these residues do not interact strongly with one another.

Figure 5.7 shows the contact/difference maps constructed from the long range network graphs. The upper left of the plot is a contact map constructed from the crystal structure of one of the variants being compared (the one mentioned first in the legend). A contact is assumed to exist and thus shown on the map when the C_α atoms of residues are less than 10 Å apart. The bottom right gives the absolute value of the difference in coupling between the two variants given in the legend. A representation of the secondary structure is given on the x-axis. For the most part the differences in coupling are small and occur in regions that are indicated in the contact map, indicating that they are



(a) 1zid/2aq8 **(b)** 1zid/2nv6

(c) 1zid/4dti **(d)** 2aq8/2nv6

(e) 2aq8/4dti **(f)** 2nv6/4dti

Figure 5.8: Contact/difference maps for the equilibrated apo-protein with only long range interactions considered. The colour code for the figures is the absolute value (in cm^{-1}) of the differences between the interactions. Upper quadrant is the contact map, x-axis is a representation of the secondary structure elements: red=α-helix, green/cyan=β-sheet, dark blue=random coil. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

due to these residues' fluctuating dynamics.

Table 5.7 gives the largest differences in coupling but for residues occurring in either β-sheets or α-helices. Active site residues appear more frequently in the β-sheet data than in the α-helix data. Short range interactions appear more frequently in the α-helix data than in the β-sheet data while the β-sheets have more long range interactions than the α-helices. This is expected given the nature of the hydrogen bond patterns that occur in these secondary structure elements. Residue ^{94}Ser/Ala, the site of the mutation conferring isoniazid resistance, appears in the β-sheet data, but the only significant change in its coupling between variants is between it and residue ^{93}His.

The data for the apo-protein simulations indicates broad similarities in the structure and dynamics of the four variants. This is to be expected since the four variants share near identical sequences. This data suggests that there is one conformational state adopted by the unbound protein, whether mutated or not, and highlights several residues to compare to data from the holo-protein simulations.

1zid, 2aq8		*1zid, 2nv6*		*1zid, 4dti*		*2aq8, 2nv6*		*2aq8, 4dti*		*2nv6, 4dti*	
Interaction	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$
149, 150	8	149, 150	9	100, 103	4	40, 62	3	149, 150	7	40, 62	3
148, 149	4	148, 149	4	252, 253	4	41, 43	3	252, 253	4	41, 43	3
103, 106	4	103, 106	4	103, 106	4	97, 119	3	149, 191	3	97, 119	3
100, 103	4	104, 106	4	41, 43	3	42, 43	3	148, 149	3	42, 43	3
264, 267	3	149, 191	4	174, 175	3	265, 266	3	251, 252	3	265, 266	3
41, 42	3	41, 43	4	15, 40	3	196, 198	2	93, 146	3	196, 198	2
149, 191	3	95, 123	3	104, 106	3	41, 42	2	264, 267	2	41, 42	2
46, 47	3	103, 104	3	251, 252	3	93, 94	2	154, 155	2	93, 94	2
104, 106	2	104, 156	3	41, 42	2	158, 159	2	15, 40	2	158, 159	2
103, 104	2	264, 267	3	64, 67	2	105, 106	2	41, 43	2	105, 106	2

Table 5.6: The 10 largest differences between each network model of the equilibrated structures; all interactions included. The coupling, J, is given in wavenumbers (cm $^{-1}$). Active site residues underlined. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

1zid, 2aq8		*1zid, 2nv6*		*1zid, 4dti*		*2aq8, 2nv6*		*2aq8, 4dti*		*2nv6, 4dti*	
Interaction	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$
β-sheet residues											
148, 149	4	148, 149	4	*40, 62*	3	*40, 62*	3	149, 191	3	149, 191	4
149, 191	3	149, 191	4	93, 94	2	93, 94	2	148, 149	3	148, 149	3
93, 146	2	*145, 146*	2	141, 142	2	141, 142	2	*93, 146*	3	145, 146	3
39, 40	2	93, 94	2	15, 40	2	15, 40	2	15, 40	2	*93, 146*	3
40, 62	2	*93, 146*	2	256, 257	1	*256, 257*	1	13, 93	2	93, 94	2
α-helix residues											
103, 106	4	103, 106	4	*100, 103*	4	41, 43	3	*264, 267*	2	*265, 266*	3
41, 43	4	41, 43	4	103, 106	4	97, 119	3	41, 43	2	*264, 265*	3
95, 123	3	95, 123	3	41, 43	3	42, 43	3	*174, 175*	2	*174, 175*	2
103, 104	3	103, 104	3	174, 175	3	*265, 266*	3	*112, 113*	2	42, 43	2
264, 267	3	*264, 267*	3	64, 67	2	196, 198	2	*112, 115*	2	97, 119	2

Table 5.7: The 10 largest differences between each equilibrated apo-protein network model; residues in α-helices and β-sheets only. The coupling, J, is given in wavenumbers (cm $^{-1}$). Active site residues are underlined and β-β or α-α interactions are in italics. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

5.3.1.3 Holo-protein

Figure 5.8 shows the network models generated from the holo-protein simulations with the same protocol as the apo-simulations. The networks are similar in layout, and look much the same as their respective apo-protein networks. The model for the inhibited mutant (2nv6) appears to show helix 2 in slightly different position to the other three models. Figure 5.9 shows the long distance networks which display the same pattern. For the mutant with inhibitor model (2nv6) in figure 5.9 the β-sheet is easily observable, and there seems to be a cluster of interactions to the left of the sheet which seems to include residues in helix 2. This needs to be investigated further.

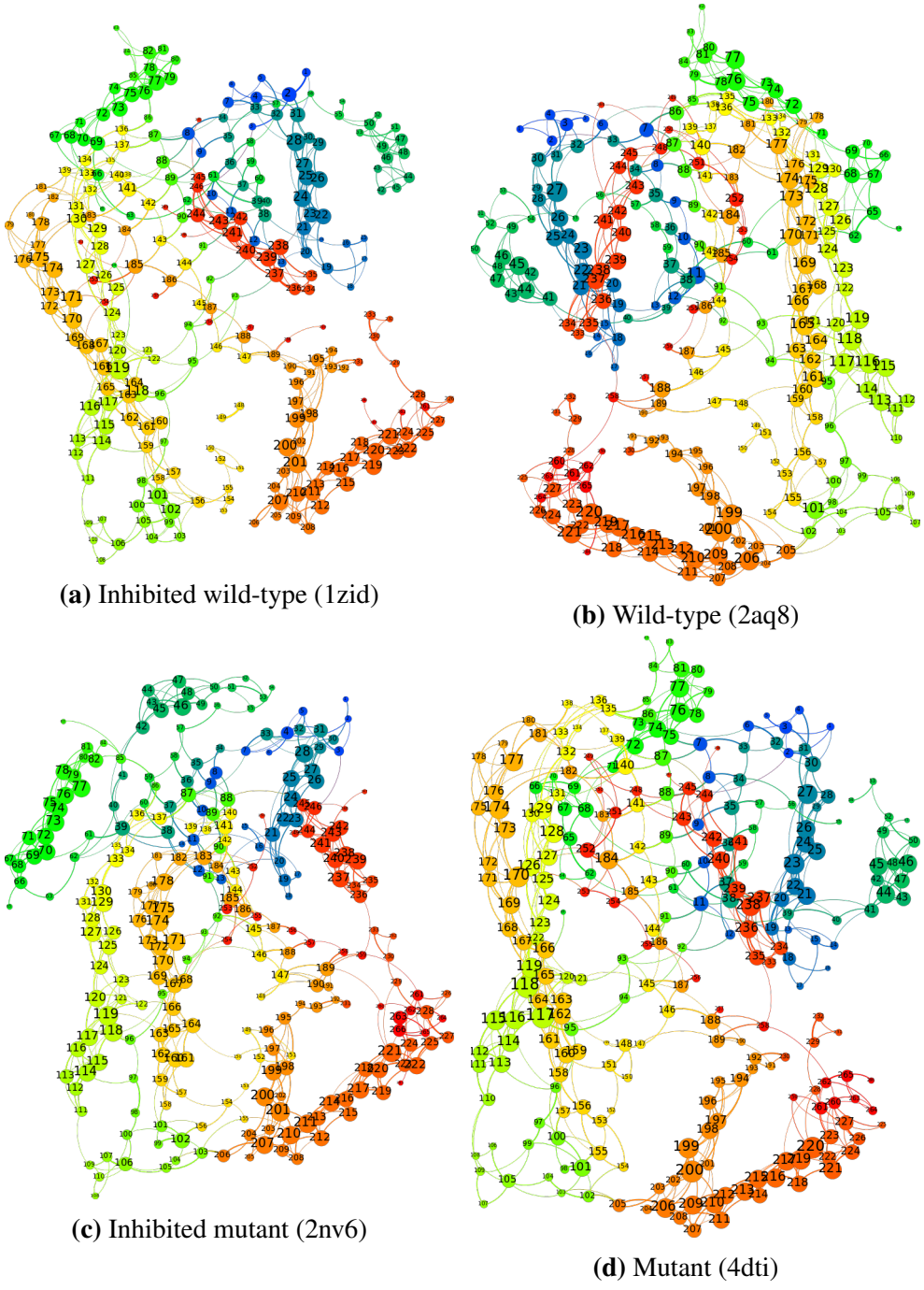
Table 5.8 gives the largest couplings in the network model of each variant. These are larger than those seen in table 5.5, with some of the couplings exceed 10 cm^{-1}. There is also a larger proportion of short range interactions present in this data, and we see some active site residues appearing. The nearest residue to the Ser^{94}Ala mutation conferring resistance is residue ^{97}Phe, which appears in the table for three of the variants, but not for the mutant bound to isoniazid.

The difference maps in figure 5.10 show many more large variations than were seen in figure 5.7, indicating that the binding of the ligands generates much more conformational diversity between the variants. There are fewest differences between the wild type and Ser^{94}Ala mutant both bound to NADH (2aq8 and 4dti respectively) which indicates that the structure and dynamics of the bound proteins are similar; this is reassuring since the function of these two variants should be identical. There are also large differences between the mutant with inhibitor and the other variants, both wild type and mutation, suggesting that there may indeed be significant structural changes resulting from the isoniazid-NAD adduct binding to the mutant.

Interestingly, residue ^{94}Ser/Ala is absent from table 5.9, indicating that despite the point mutation the dynamics of that residue are largely unaffected. Residue ^{95}Ile, on the other hand, does appear in the tables, suggesting that it

might be the dynamics of this (and other residues surrounding the active site) that may engender the resistance to isoniazid. Since the point mutation is from a more sterically bulky residue (serine) to a less sterically bulky residue (alanine) this may be allowing residue ^{95}Ile to move in to the active site and interact in a manner which results in isoniazid becoming less effective at inhibiting the protein. Previous work has shown evidence that the mutation does not affect binding affinity of the inhibitor, so if residue ^{95}Ile is involved in isoniazid resistance it is not preventing the ligand from binding to the active site. Visualisation of the simulations revealed some differences in the position of ^{95}Ile between the wild-type InhA and the Ser94Ala mutant, though this was only a small movement. Given this it is likely that, while ^{95}Ile may help mediate the effect, overall resistance to isoniazid is likely to be due to the movement of a number of residues as a result of the Ser94Ala mutation.

Table 5.3.1.2 shows the differences between the four equilibrated holo-structures, focussing on the secondary structure elements. Both sets of secondary structure elements show a similar patterns in that the coupling differences are larger between the wild-type (2aq8) and inhibited wild-type (1zid) and growing smaller moving to the comparison between the wild-type and mutant (4dti). The differences grow larger again when comparing the mutant and inhibited mutant (2nv6). There are a few residues that occur repeatedly in this data; residues 147–149 occur frequently in the data for the α-helix data, with residue 148 in particular turning up for almost every comparison. For the β-sheet data, residues 122 and 222 occur several times. In general the data shows more β-β interactions than α-αinteractions, which indicates differences in the β-sheet between variants. The β-sheet may then have a role to play in mediating isoniazid resistance.



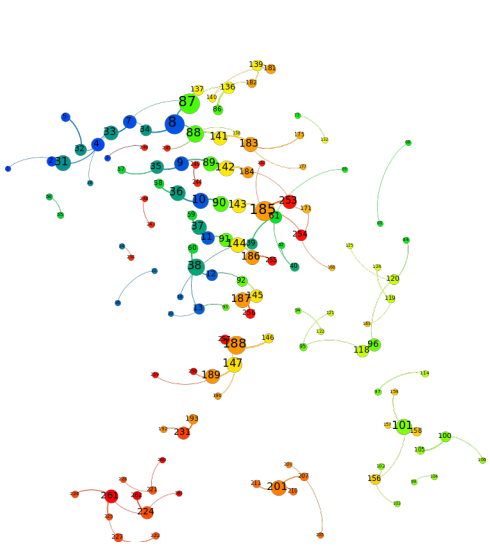
(a) Inhibited wild-type (1zid)

(b) Wild-type (2aq8)

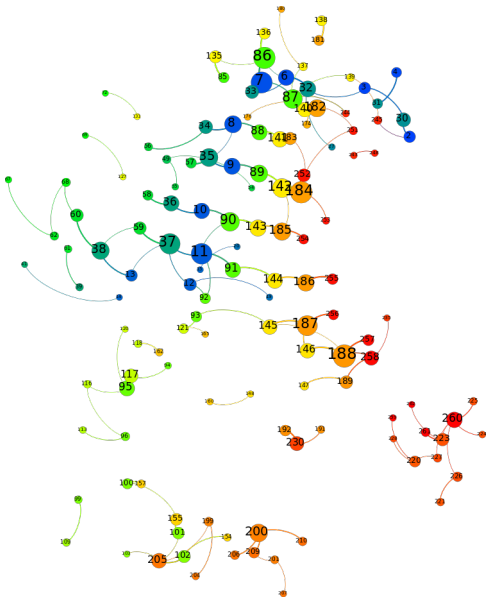
(c) Inhibited mutant (2nv6)

(d) Mutant (4dti)

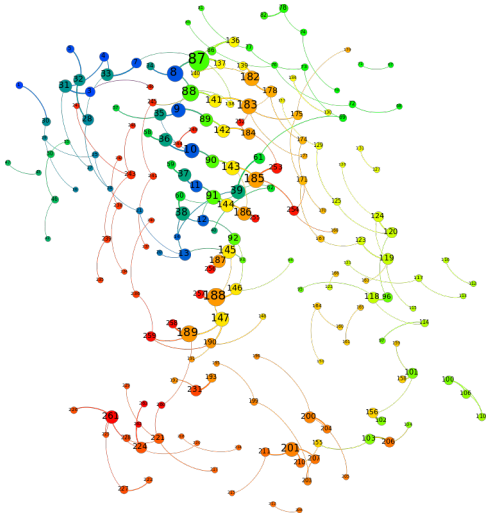
Figure 5.9: Network models extracted from holo-systems that have undergone equilibration with all interactions included in the model.



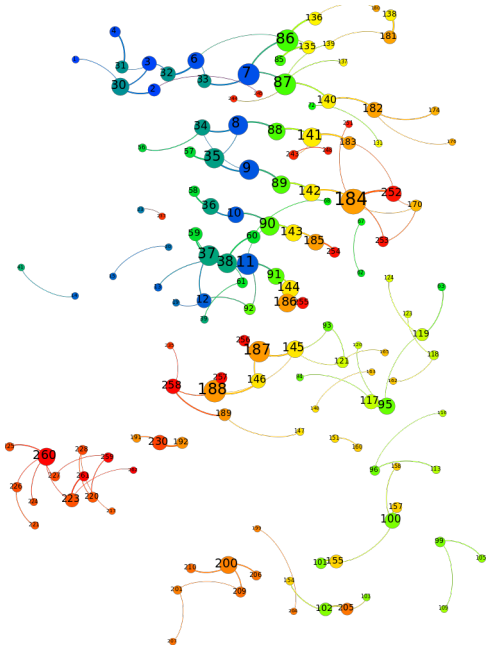
(a) Inhibited wild-type (1zid)



(b) Wild-type (2aq8)



(c) Inhibited mutant (2nv6)

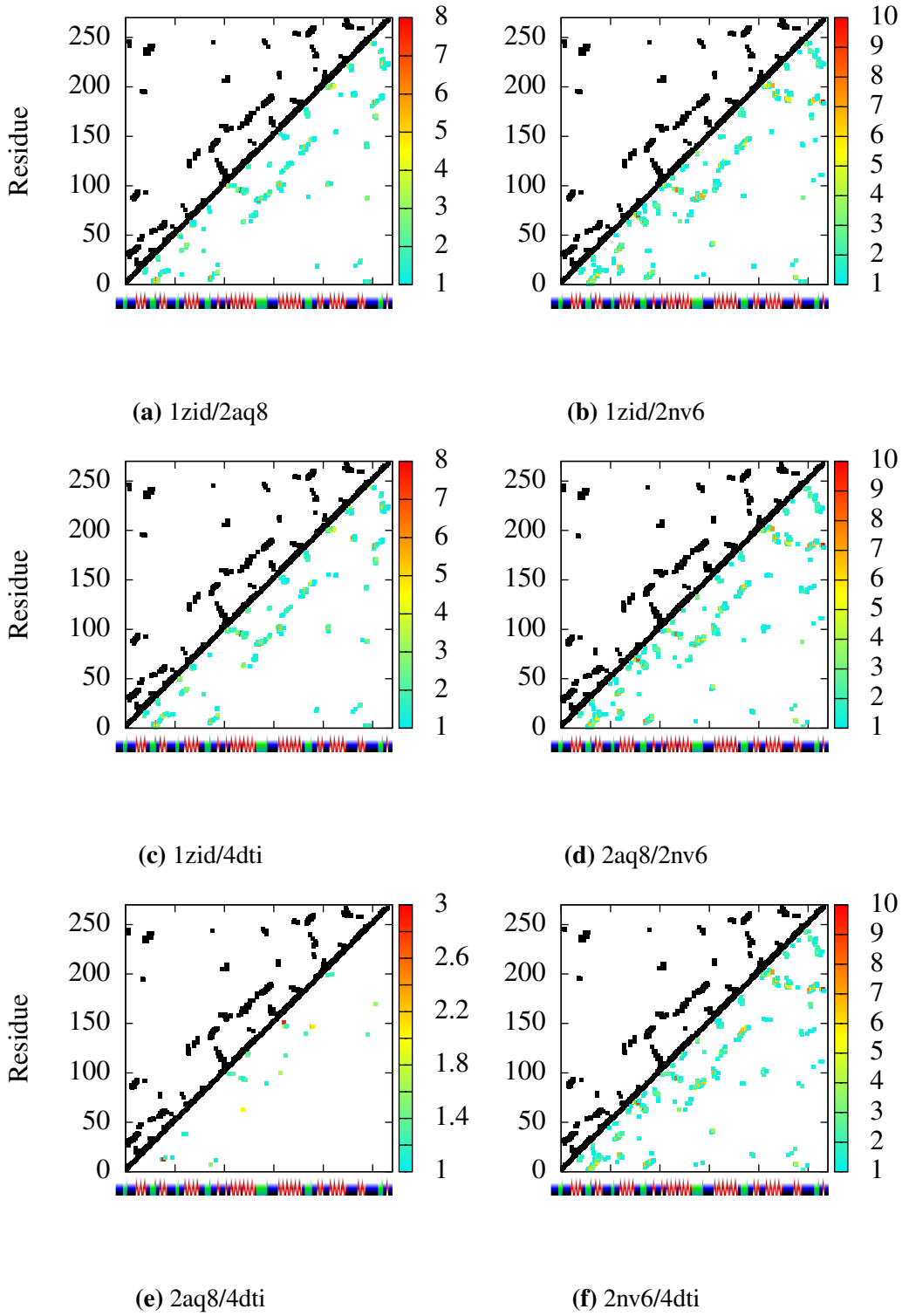


(d) Mutant (4dti)

Figure 5.10: Network models extracted from holo-systems that have undergone equilibration with only $i,i\pm5$ interactions included in the model.

Interaction (Residue)	Structure PDB code *1zid*	*2aq8*	*2nv6*	*4dti*
2 , 3	10			
3 , 4		8		
8 , 34				8
11 , 91				7
40 , 41	9			
41 , 42		9		
82 , 83	10	9	9	10
88 , 138			8	
97 , 119	8	8		8
100 , 103			8	8
102 , 103	9		9	
141 , 142	9	8		9
147 , 189			7	
149 , 150	11	11		
150 , 151		9		
175 , 178			7	
186 , 254				8
188 , 257				7
191 , 192	9		8	
223 , 226	10	8	8	
228 , 229	8	9	8	9
229 , 230		8		
231 , 234				8
264 , 267			9	

Table 5.8: The 10 largest interactions extracted from each network and compared across all equilibrated holo-structures. The coupling, J, is given in wavenumbers (cm $^{-1}$). Long range interactions highlighted in bold and active site residues underlined. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.



(a) 1zid/2aq8 **(b)** 1zid/2nv6

(c) 1zid/4dti **(d)** 2aq8/2nv6

(e) 2aq8/4dti **(f)** 2nv6/4dti

Figure 5.11: Contact/difference maps for the equilibrated holo-protein with only long range interactions considered. Z-axis of the figures is the absolute value (in cm^{-1}) of the differences between the interactions. Upper quadrant is the contact map, x-axis is a representation of the secondary structure elements: red=α-helix, green/cyan=β-sheet, dark blue=random coil. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

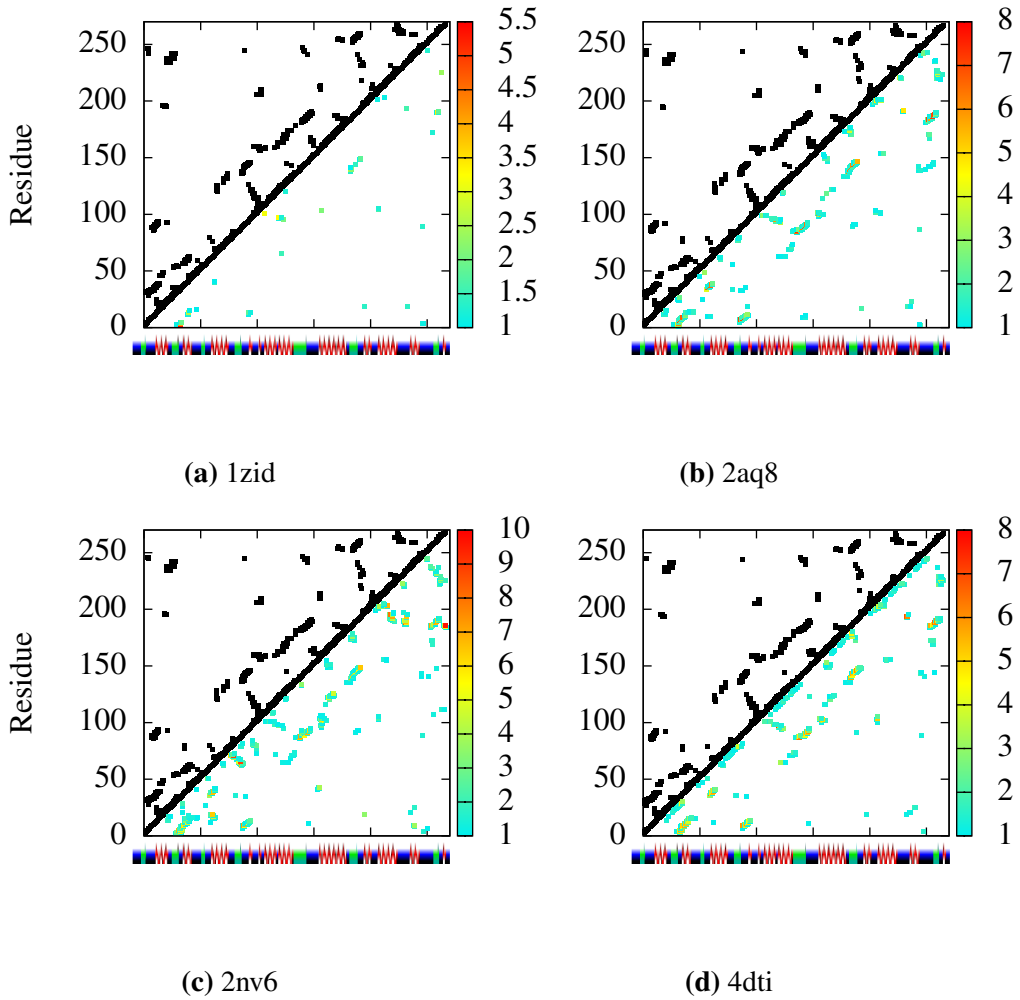
1zid, 2aq8		*1zid, 2nv6*		*1zid, 4dti*		*2aq8, 2nv6*		*2aq8, 4dti*		*2nv6, 4dti*	
Interaction	$\|\Delta J\|$	**Interaction**	$\|\Delta J\|$	**Interaction**	$\|\Delta J\|$	**Interaction**	$\|\Delta J\|$	**Interaction**	$\|\Delta J\|$	**Interaction**	$\|\Delta J\|$
147 , 148	11	149 , 150	10	80 , 81	9	149 , 150	11	138 , 139	10	82 , 83	10
148 , 149	10	2 , 3	10	222 , 225	9	148 , 149	4	185 , 265	10	228 , 229	9
222 , 225	9	264 , 267	9	226 , 227	8	13 , 14	4	177 , 178	10	141 , 142	9
205 , 206	8	3 , 32	6	81 , 82	8	64 , 67	4	70 , 78	9	231 , 234	8
80 , 81	8	100 , 103	6	148 , 149	8	156 , 157	4	65 , 84	9	97 , 119	8
226 , 227	8	103 , 106	6	95 , 117	8	72 , 75	4	172 , 173	9	100 , 103	8
95 , 117	8	104 , 207	6	139 , 140	8	71 , 72	4	85 , 86	8	186 , 254	8
81 , 82	8	104 , 106	6	184 , 252	8	71 , 72	4	75 , 76	8	8 , 34	8
62 , 65	8	264 , 266	5	230 , 233	8	70 , 73	4	187 , 230	8	11 , 91	7
190 , 191	8	44 , 47	5	190 , 191	8	264 , 267	4	229 , 232	8	188 , 257	7

Table 5.9: The 10 largest differences between each network model of the equilibrated structures; all interactions included. The coupling, J, is given in wavenumbers (cm $^{-1}$). Active site residues underlined. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

1zid, 2aq8		*1zid, 2nv6*		*1zid, 4dti*		*2aq8, 2nv6*		*2aq8, 4dti*		*2nv6, 4dti*	
Interaction	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$
β-sheet residues											
147 , 148	11	148 , 149	10	148 , 149	8	62 , 65	8	*147 , 148*	6	139 , 182	7
148 , 149	10	*147 , 148*	5	*139 , 140*	8	139 , 182	7	*11 , 12*	5	*139 , 140*	7
62 , 65	8	*39 , 40*	5	*190 , 191*	8	*190 , 191*	6	148 , 151	3	*190 , 191*	6
190 , 191	8	40 , 41	4	185 , 253	6	*139 , 140*	6	13 , 38	3	62 , 65	6
139 , 140	7	*189 , 258*	4	148 , 151	6	*147 , 148*	6	*61 , 62*	3	185 , 253	6
α-helix residues											
222 , 225	9	*263 , 266*	9	*80 , 81*	9	*263 , 266*	9	*119 , 122*	5	*263 , 266*	9
80 , 81	8	2 , 31	9	*222 , 225*	9	226 , 227	8	151 , 160	3	226 , 227	9
226 , 227	8	99 , 102	6	226 , 227	8	95 , 117	8	51 , 52	3	80 , 81	8
95 , 117	8	102 , 105	6	*81 , 82*	8	99 , 102	8	*118 , 122*	3	95 , 117	8
81 , 82	8	103 , 206	6	95 , 117	8	80 , 81	7	*49 , 52*	2	99 , 102	8

Table 5.10: The five largest differences between each equilibrated holo-protein network model; residues in α-helices and β-sheets only. The coupling, J, is given in wavenumbers (cm $^{-1}$). Active site residues are underlined and β-β or α-α interactions are in italics. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

5.3.2 Effect of Ligand Binding



(a) 1zid **(b)** 2aq8

(c) 2nv6 **(d)** 4dti

Figure 5.12: Contact difference map comparing each of the unbound variants to their bound counterparts. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

Figure 5.12 shows the differences in coupling between residues on binding of the ligand. The plot for 1zid shows the fewest differences indicating that the interactions between amide units of the residues are unaffected by the binding of the ligand i.e. that the behaviour of the bound form is the same as the unbound.There are large differences between the bound and unbound forms of the wild type (2aq8) and $Ser^{94}Ala$ mutant which indicates that the behaviour of these two variants is similar once NADH is bound—the binding affinity of the $Ser^{94}Ala$ variant (4dti) for NADH is lower than that of the wild type. What is interesting is that the $Ser^{94}Ala$ variant bound to isoniazid (2nv6) also shows a

large number of differences in a similar pattern to those in the wild-type (2aq8) and $Ser^{94}Ala$ (4dti) plots. It would be expected that, if the differences in coupling on binding are significant, they should result in some visible difference in the network graph generated from those couplings.

Table 5.11 shows the difference between the bound and unbound forms of each variant of InhA. The differences on binding are of similar magnitude for three of the variants, with the structure based on the bound wild-type (1zid) showing the smallest change in coupling between the apo- and holo- forms. The inhibited mutant variant (2nv6) shows generally larger changes in inter-residue coupling upon ligand binding, with the exception of the 82, 83 interaction in the mutant (4dti). Interestingly, the largest changes in coupling upon ligand binding are largely in residues outside the active site, with only active-site residue 41 appearing multiple times in the table.

The data in table 5.12 shows the changes in residue coupling between the holo- and apo- forms of each variants, focussing on the secondary structure interactions. Residue 94—the residue responsible for conferring isoniazid resistance in the mutant—appears in the table for the inhibited mutant (2nv6) data fro the β-sheets. Otherwise there are few active site residues in the dataset. Again there are a larger number of β-β interactions in the table than α-α interactions, suggesting that changes occur in the β-sheet upon ligand binding.

1zid		*2aq8*		*2nv6*		*4dti*	
Interaction	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$
3 , 32	5	221 , 224	9	223 , 226	9	82 , 83	10
253 , 254	5	139 , 140	9	14 , 209	9	228 , 229	9
40 , 41	5	226 , 227	8	24 , 74	9	141 , 142	9
155 , 156	5	80 , 81	8	245 , 253	8	149 , 151	8
194 , 195	4	86 , 136	7	129 , 226	8	223 , 227	8
255 , 256	4	6 , 32	7	82 , 83	8	228 , 230	8
248 , 249	4	184 , 252	7	124 , 251	8	97 , 119	7
247 , 250	4	186 , 255	7	22 , 41	8	141 , 143	7
4 , 32	4	189 , 190	7	39 , 130	8	231 , 234	7
122 , 123	4	229 , 232	7	82 , 107	8	207 , 208	7

Table 5.11: The 10 largest differences between the network models for the apo- and holo-protein simulations. The coupling, J, is given in wavenumbers (cm $^{-1}$) and active site residues underlined. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

1zid		*2aq8*		*2nv6*		*4dti*	
Interaction	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$	**Interaction**	$\lvert\Delta J\rvert$
β-sheet residues							
4 , 32	4	*139 , 140*	9	*40 , 62*	4	*141 , 142*	9
190 , 259	3	186 , 255	7	149 , 191	3	*141 , 143*	7
190 , 191	3	*189 , 190*	7	93 , 94	3	188 , 257	7
257 , 258	3	*89 , 142*	7	*61 , 62*	3	87 , 138	6
140 , 182	3	*9 , 89*	7	185 , 253	2	37 , 59	6
α-helix residues							
3 , 32	5	*221 , 224*	9	122 ,123	4	*82 , 83*	10
253 , 254	5	226 , 227	8	*107 ,108*	4	223 , 227	8
255 , 256	4	*80 , 81*	8	*43 ,46*	4	97 , 119	7
248 , 249	4	6 , 32	7	*123 ,124*	3	207 , 208	7
247 , 250	4	*173 , 176*	7	97 ,119	3	82 , 84	7

Table 5.12: The five largest differences between the apo- and holo-protein network models for residues in α-helices and β-sheets only. The coupling, J, is given in wavenumbers (cm $^{-1}$). Active site residues are underlined and β-β or α-α interactions are in italics. PDB codes: 1zid=inhibited wild-type; 2aq8=wild-type; 2nv6=inhibited mutant; 4dti=mutant.

5.3.3 FTIR Spectra

While the calculation of 2DIR spectra for a protein the size of InhA is largely intractable with current computational limitations, the calculation of the FTIR spectrum is relatively straightforward by comparison. The mean FTIR spectra (across 12 simulations) for each variant in both bound and unbound states were calculated to investigate whether the binding of the ligand has any noticeable affect on the spectrum that could provide markers for experimental investigations. FTIR spectra are shown in figure 5.13.

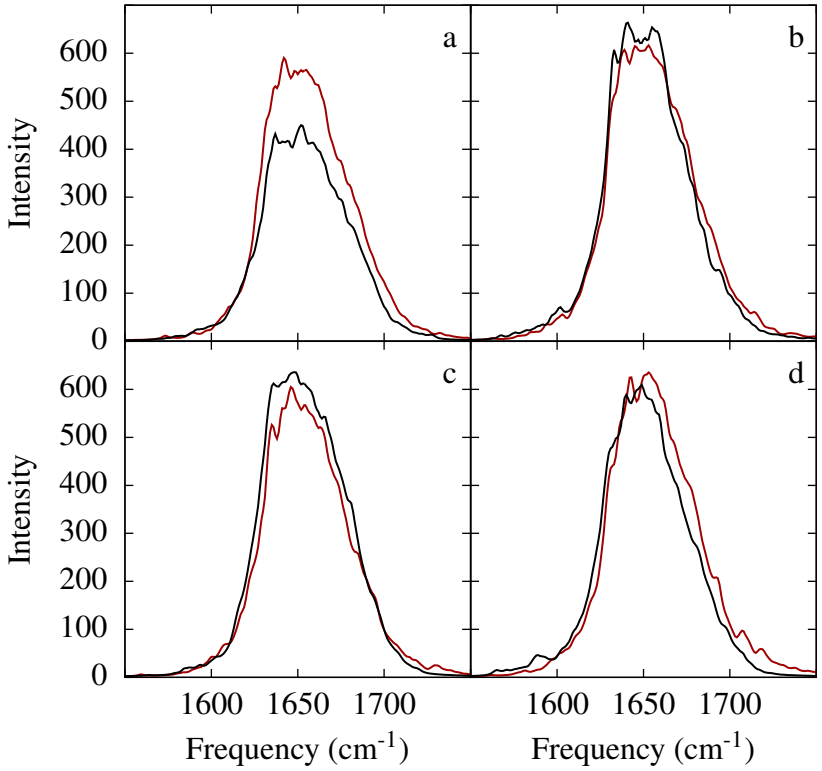


Figure 5.13: FTIR calculated from dynamics using SPECTRON for four InhA variants: a) $Asp^{2}Ala$ mutant with isoniazid (1zid), b) wild-type protein with NADH (2aq8), c) $Ser^{94}Ala$ mutant bound to isoniazid (2nv6) and d) $Ser^{94}Ala$ mutant bound to NADH (4dti). The red line denotes the apo- form and the black line is the holo- form.

The FTIR spectra for the apo- and holo- show broad similarities, as might be expected for proteins of this size with largely the same structure. The most notable differences are in figure 5.12a when the wild type protein is bound to the inhibitor. Binding of isoniazid results in a reduction in the intensity of the spectral line by almost a third. Tying this reduction to structural changes could yield useful information on the mechanism of action of the inhibitor, and any important residues for binding. Aside from the change in intensity the shape of the peak in figure 5.12a remains relatively unchanged.

The FTIR spectrum for the $Ser^{94}Ala$ mutant in complex with the inhibitor shows the opposite trend, with a slight increase in peak intensity upon binding of the ligand. This could be indicative of broader structural changes as a result of a mutation that may be contributing to isoniazid resistance. The change in intensity is relatively small so any structural changes contributing to the change in signal are likely to be small. There is also a small change in intensity upon ligand binding for the wild type with NADH (2aq8).

The spectrum for the $Ser^{94}Ala$ mutant bound to NADH (4dti) shows similar line shape and intensity for both the bound and unbound version, though there does appear to be small down shift in the frequency of the band upon binding of the ligand.

5.4 Discussion

The network models constructed from the inter-residue coupling data show a resemblance to models of the protein constructed from three-dimensional coordinates. There are a few differences that can be attributed to the network models being wholly two-dimensional in nature while the coordinate models attempt to represent the 3D structure, but overall they look the same for most of the protein systems. The α-helices can be clearly seen as long "strings" of nodes that are very close together with similar residue numbers. The β-sheet is somewhat harder to visualise in the all interaction models but can be discerned

in the long range models as a number of nodes relatively far apart with edges between them. The ability to see the secondary structure elements—and the similarity between the models and VMD cartoon structures—indicates that the coupling data used to construct these network models is reliable. The models themselves provide an intuitive visualisation of the secondary structure of the protein that are easier to compare than cartoon models constructed from coordinates. These models can be used to guide a deeper analysis of the structure and dynamics of the protein.

In the case of the apo-protein simulations, the network models are highly similar. Given the high backbone similarity of the different variants and the fact that the primary structures of the variants only differ by one or two residues this is reassuring. It can be assumed, on the basis of the network model similarity, that the four apo- variants undergo similar dynamical processes during the time frame studied, at least in a macro sense. We see a general "relaxation" of the 3D structure for all four variants in the absence of the bound ligand. Table 5.5 shows slightly smaller coupling values than the corresponding table for the holo-simulations (table 5.8) supporting the idea of a relaxation of the structure. There are few active site residues in table 5.5 indicating that in the apo-protein these residues sit further apart than when the ligand is present. When the ligand is bound we see larger differences in the shorter range interactions (*i,i+4* or less) and larger couplings in table 5.8. This is in contrast to the apo-protein data which shows the longer range couplings as having larger values as seen in table 5.5.

Figure 5.9 shows the long range interaction models for the holo- variants of InhA. In this figure the β-sheet of the inhibited mutant (2nv6) is harder to identify than for the other models. In addition to the nodes and edges identifying the sheet, there are a number of long range interactions between the β-sheet residues and residues in other parts of the protein, specifically in the region of residues 30–50 and 240–260 which corresponds to helices 2 and 8.

This is not seen in either the apo-protein models in figure 5.6 or in the wild type protein bound to the inhibitor (1zid) i.e. the inactive protein. It is possible that these long range interactions may sufficiently change the 3D structure of the protein to allow InhA—and the other proteins in the biosynthesis pathway—to remain active despite the binding of isoniazid. Previous work[166] suggested there may be a second biologically active conformation of InhA and we may be seeing evidence of this here.

Other work[167] has shown evidence of a binding loop in the region of residues 180–200 formed of helix 9 and β-strand 6. Residues in this region appear in the coupling data several times. Table 5.8, shows the largest coupling interactions present in the holo-protein simulations. Residues from the binding loop appear in this table on four occasions with the interactions being: 147, 189; 186, 254; 188, 257 and 191, 192. These interactions are mostly between residues in β-strand 6 and residues towards the C-terminus of the protein. For the apo-protein in table 5.5 there are only two interactions within the region of the binding loop: 145, 185 and 146, 188. For the apo-protein the interaction is between the binding loop and β-strand 5. There are also a number of residues from the binding loop present in table 5.9, in particular the 190, 191 interaction which is also present in some of the other tables. There are also a few residues in this region in table 5.11, particularly for the wild type protein, indicating that this region may indeed have a role in mediating the binding of the ligand. This would suggest that our simulations support the work in the literature.

Direct comparison of the apo- and holo- simulations in figure 5.11 reveals very few changes in coupling between residues on binding of the inhibitor to the wild-type (1zid), though with the first two residues missing this variant essentially has the same sequence as the wild type protein (2aq8). This is an interesting result as the 1zid variant with isoniazid bound is the inhibited protein but the majority of the changes in coupling on binding are less than 1 cm^{-1} and are not shown in the figure as they can can be considered the same.

The other three variants show greater changes in coupling upon binding of the ligand, in particular the wild type (2aq8) and the Ser^{94}Ala mutant (4dti) which were simulation in the absence and presence of the native ligand NADH. The Ser^{94}Ala mutant with the inhibitor (2nv6) shows more changes in coupling than 1zid but fewer than either wild type or mutant with isoniazid (2aq8 and 4dti respectively). Experimental work has indicated that the largest changes in the 2DIR spectrum occur on binding of NADH to the wild type protein and our results from the coupling maps seems to support that, though our calculated FTIR show the greatest change upon binding of isoniazid to the inhibited wild-type (1zid). Experimental 2DIR has also suggested a general "relaxation" of the β-sheet having a role in mediating resistance to isoniazid. Possible evidence for this from our simulations is the slightly larger differences in coupling for β-sheet residues between variants in the holo-simulations, though this effect is small.

The small changes in coupling on binding of the inhibitor may be evidence that the mechanism of action of isoniazid may be more complex than the inhibitor competitively inhibiting the active site or causing changes in the structure of InhA. There is evidence in the literature[168] that the mechanism of isoniazid resistance involves complex interactions between InhA and other proteins in the mycolic acid biosynthesis pathway. It is possible that this is why were are seeing only small changes in coupling—the mutations of the protein and the binding of the inhibitor may be having an effect on protein-protein interactions within the biosynthesis pathway that we are unable to see with the current investigation. This possibility presents an opportunity for further work.

One aim of this work was to identify residues that might be suitable candidates for isotope labelling in 2DIR experiments. There are a few residues that show interesting results or appear in the coupling tables multiple times, indicating these might be potential targets for labelling studies. Residue ^{95}Ile is a possible candidate for labelling studies. It is one of the residues next to the

point mutation conferring isoniazid resistance (residue 94) and is interesting because it appears in the coupling tables several times while residue 94 (either serine or alanine depending on the variant) does not. We have speculated that since the Ser^{94}Ala mutation replaces a more sterically bulky residue with a less sterically bulky one, this might allow residue ^{95}Ile to move into the active site. We have seen some evidence of this upon visualisation of the simulations, but it would be useful to investigate this experimentally.

Another possible target is residue ^{41}Phe, which is an active site residue that appears in a number of tables. In particular it appears in table 5.11 for the Ser^{94}Ala mutant bound to the inhibitor (2nv6), indicating that it may play some role in isoniazid resistance that could be investigated further. Residue ^{149}Phe is another active site residue that appears to have particular strong changes in coupling between the holo-simulations as well as upon binding of the ligand (tables 5.11 and 5.12). It is also part of the β-sheet and isotope labelling may allow further support for the idea of the sheet relaxing. Residue ^{191}Ala is apart of the proposed binding loop[167] and appears a number of the coupling tables. The coupling between this residue and others seems to vary depending on the mutation of InhA and the nature of the ligand bound and it could therefore be an interesting probe for the mechanism of resistance.

5.5 Concluding Remarks

One of the major goals of this work was to investigate the viability of using network models as a supplement or replacement for calculated 2DIR spectra in cases where calculation of the spectrum (especially via the exciton method) is computationally expensive or intractable. We have done this by constructing network models of the inter-residue coupling from the one-exciton Hamiltonian matrix. These network models provide a visually intuitive method of looking at coupling patterns, but more useful is analysis of the couplings themselves. By looking at the the differences in coupling patterns between variants and upon

ligand binding we have identified trends in the coupling that support previous experimental work. In particular we see evidence that a binding loop consisting of residues between 180 and 200 may have an effect on mediating isoniazid resistance. In addition there is some evidence from our coupling data that supports the idea of a second biologically active conformation of InhA which is potentially reached via a relaxation of the β-sheet.

We also wanted to identify potential targets for isotope labelled 2DIR studies of InhA. Residues ^{41}Phe, ^{95}Ile, ^{149}Phe and ^{191}Ala have been identified as possible targets. Each of these residues lies within either the active site or the binding loop and shows large differences in coupling between variants depending on the binding state of the protein. Isotope labelled 2DIR could elicit information on the role these residues may play in either the mechanism of action of the inhibitor isoniazid or the mechanism of isoniazid resistance.

A major goal of further work would be to calculate 2DIR spectra for this system and its variants. This would allow full comparison with experimental results and has the potential to verify the conclusions drawn from the coupling maps and network models. It would also be useful to look further into the dynamics of the system, for example by developing an intuitive way to visualise the coupling data for multiple snapshots of the simulation. Longer simulations of these variants would also be useful to investigate whether or not we can see direct evidence in the simulation of either the binding loop or the proposed second conformation of InhA. The ultimate goal would be to carry out MD simulations not just of InhA but also the other proteins in the pathway to investigate the protein-protein interaction and the effect these may have on the spectroscopy of the protein(s).

Chapter 6

Concluding Remarks

In this thesis we have expanded upon previous work on the amide I spectra of Leu-enkephalin by calculating FTIR and 2DIR spectra using the exciton method. This was done for the low energy conformations investigated previously and for MD trajectories carried out in three separate environments; gas phase, explicit water and implicitly modelled acetonitrile. In addition to the calculated spectra on this small peptide system, we have investigated using network models and the residue-residue couplings from the one exciton Hamiltonian to probe the dynamics of InhA, a protein system for which calculated 2DIR spectra are currently intractable. We have used these models to support previous experimental work, as well as suggest potential targets for isotope labelled 2DIR experiments of InhA.

In chapter 4 we calculated the amide I region of Leu-enkephalin in both FTIR and 2DIR for single conformations and from MD simulations carried out in the gas phase, explicit water and continuum solvent environments. This was done using by calculating the one exciton Hamiltonian using the Jansen scheme. FTIR and 2DIR spectra were calculated using the non-linear exciton equations in SPECTRON. The calculated FTIR spectra of the starting conformations show some similarity to previously calculated spectra, giving us confidence in the exciton method regarding its suitability for calculating qualitatively accurate spectra of small peptides.

When dynamics are incorporated we saw mixed results for all three environments studied; some spectra appear as would be expected for the amide I band of a peptide (i.e. as a broad band around 1650 cm^{-1}) but others appear as a single intense peak—or multiple identifiable peaks—in the same region. This is particularly the case for the continuum solvent simulations, and we have considered the suitability of using implicit solvent methods in conjunction with the exciton method. For the FTIR spectra that appear as a broad band, there is a good qualitative match with experimental results, and as such the 2DIR spectra can be considered to be of reasonable quality.

One of the major goals of chapter 5 was to investigate the viability of using network models as a supplement or replacement for calculated 2DIR spectra in cases where calculation of the spectrum (especially via the exciton method) is computationally expensive or intractable. We have done this by constructing network models of the inter-residue coupling from the one-exciton Hamiltonian matrix. These network models provide a visually intuitive method of looking at coupling patterns, but more useful is analysis of the couplings themselves. By looking at the the differences in coupling patterns between variants and upon ligand binding we have identified trends in the coupling that support previous experimental work. In particular we see evidence that a binding loop consisting of residues between 180 and 200 may have an effect on mediating isoniazid resistance. In addition there is some evidence from our coupling data that supports the idea of a second biologically active conformation of InhA which is potentially reached via a relaxation of the β-sheet.

We also wanted to identify potential targets for isotope labelled 2DIR studies of InhA, which we have done. Residues ^{41}Phe, ^{95}Ile, ^{149}Phe and ^{191}Ala have been identified as possible targets. Each of these residues lies within either the active site or the binding loop and shows large differences in coupling between variants depending on the binding state of the protein. Isotope labelled 2DIR could elicit information on the role these residues may play in either the

mechanism of action of the inhibitor isoniazid or the mechanism of isoniazid resistance.

Further work should be carried out on Leu-enkephalin to compare implicit solvent methods with explicit solvent simulations and the effect this has on the accuracy of the calculated FTIR and 2DIR spectrum. Where the FTIR spectra appear as peaks rather than broad bands, we have discussed the merits of carrying out additional simulations and calculating a mean spectrum across the different trajectories. This may have the effect of improving the accuracy of the spectrum with respect to experiment and further work should investigate this. In addition to the further work discussed above, the investigation of calculated FTIR and 2DIR should be expanded to include the opioid receptor that is the target for Leu-enkephalin. While the calculation of the 2DIR spectrum of the receptor is currently intractable, the effect of the ligand binding on the FTIR and 2DIR spectrum of Leu-enkephalin could be investigated. In addition, the current work could be expanded to include Met-enkephalin, to investigate the effect of nature of the terminal residue on the dynamics and spectroscopy of enkephalin.

A major goal of further work on InhA would be to calculate 2DIR spectra for this system and its variants. This would allow full comparison with experimental results and has the potential to verify the conclusions drawn from the coupling maps and network models. It would also be useful to look further into the dynamics of the system, for example by developing an intuitive way to visualise the coupling data for multiple snapshots of the simulation. Longer simulations of these variants would also be useful to investigate whether or not we can see direct evidence in the simulation of either the binding loop or the proposed second conformation of InhA. The ultimate goal would be to carry out MD simulations not just of InhA but also the other proteins in the pathway to investigate the protein-protein interaction and the effect these may have on the spectroscopy of the protein(s).

Appendices

Appendix A

Mathematical and Quantum Mechanical Concepts

A.1 Mathematical Concepts

A.1.1 Complex Numbers

A complex number is a number that can be expressed in the form of $a+bi$ where a and b are real numbers and i is the imaginary unit which satisfies $i^2 = -1$ or $i = \sqrt{-1}$. For this expression, a is considered the *real part* of the complex number and b is the *imaginary part*.

The complex conjugate of a complex number is the number where the real part is equal to the real part of the original complex number and the and imaginary part is equal and opposite in sign. For example, for the complex number $z = x+iy$, the complex conjugate would be $x-iy$. The complex conjugate of a complex number, z, is denoted as $|z|$ or z^*.

A.1.2 Introduction to Matrices

A matrix is a mathematical array of numbers, symbols or expressions that can be manipulated or interpreted in a number of ways. Typically matrices are referred to by the number of rows and columns e.g. a 2×3 matrix would appear

as follows:

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} \tag{A.1}$$

and a square matrix would have the same number of columns and rows.

The individual entries in the matrix (a, b, c etc.) are known as the matrix *elements*. A large number of transformations and manipulations can be done with matrices, staring with addition and subtraction, (provided the matrices are the same size) and growing more complex. Matrix mathematics has applications in most scientific fields; in this thesis they are most often used to represent the Hamiltonian of a system under study.

The *eigenvectors* and *eigenvalues* of a matrix are defined by the equation

$$Av = \lambda v \tag{A.2}$$

where v is a non-zero vector called the eigenvector and λ is a number called the eigenvalue. It is possible for λ to also be an eigenvalue of matrix A if $A - \lambda I_n$ is not invertible.

The *determinant* of a square matrix, A, denoted as $|A|$ or $det(A)$, is a number that encodes certain properties of the matrix. For a 2×2 matrix the determinant is given as

$$det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc \tag{A.3}$$

This can be generalised to matrices of any size using more complex formulae. Determinants are used extensively in electronic structure methods.

The *trace* of an $n \times n$ square matrix is the sum of the main diagonal elements. That is, for matrix A,

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \tag{A.4}$$

the trace of the matrix, $\mathrm{Tr}(A) = a + e + i$.

A *diagonal* matrix is one in which all elements outside of the main diagonal are zero, though the diagonal elements themselves do not have to be zero. For example

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{A.5}$$

is a diagonal matrix.

The matrix given in equation A.5 is also an identity matrix, a special kind of diagonal matrix where all diagonal elements are equal to 1 and all off-diagonal elements are equal to zero. Identity matrices are designated such because multiplication with it leaves a matrix unchanged.

An $n \times n$ square matrix, A, is *invertible* if there exists a corresponding matrix, B, such that

$$AB = BA = I_n \tag{A.6}$$

where I_n is the $n \times n$ identity matrix and A and B are multiplied using ordinary matrix multiplication. Matrix B is then the *inverse* of A and is denoted A^{-1}.

Diagonalization is the process of finding the corresponding diagonal matrix for matrix that is diagonalizable. Diagonalizing a matrix is also equivalent to finding the matrix's eigenvalues, which turn out to be precisely the entries of the diagonalized matrix. A matrix is diagonalizable if there exists an invertible matrix P such that $P^{-1}DP$ is a diagonal matrix, where P is a matrix composed of the eigenvectors of A, D is the diagonal matrix and P^{-1} is the

inverse of P. As long as P is a square matrix, an initial matrix equation such as

$$AX = Y \tag{A.7}$$

can be written as

$$DP^{-1}X = P^{-1}Y \tag{A.8}$$

Since P^{-1} is being applied to both X and Y, solving the original matrix equation is equivalent to solving

$$DX' = Y' \tag{A.9}$$

where $X' \equiv P^{-1}X$ and $Y' \equiv P^{-1}Y$. This transforms the system into the simplest possible form and reduces the number of parameters from $n \times n$ for an arbitrary matrix to n for a diagonal matrix, while obtaining the characteristic properties of the initial matrix. Matrix diagonalization occurs frequently in physics and diagonalization of the exciton Hamiltonian is an important step in calculating 2DIR spectra using the exciton method in SPECTRON.[117]

A.1.3 Gaussian and Lorentzian Distributions

A Gaussian (or normal) distribution is a commonly used probability distribution that is often used to represent real-valued random variables when the distribution is unknown. For example, when modelling an FTIR spectrum one might obtain the frequency and intensity of a peak but the band shape will be unknown; a distribution such as a Gaussian may be used to model the band shape. A Gaussian distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2)} \tag{A.10}$$

where μ is the mean value of the distribution and σ is the standard deviation.

Alternatively, in the case of modelling IR peaks, one might use a Lorentzian distribution instead. This is the distribution used by SPECTRON[117] for calculating FTIR. A Lorentzian distribution is given by

$$f(x) = \frac{1}{\pi} \frac{\frac{1}{2}\Gamma}{(x - x_0)^2 + (\frac{1}{2}\Gamma)^2} \tag{A.11}$$

where x_0 is the centre of the peak and Γ is a parameter specifying the width of the peak.

A.1.4 Fourier Transforms

The Fourier transform is a reversible, linear transformation with many important properties and applications. For a function, $f(x)$, the Fourier transform can be denoted $F(s)$. The product of x and s is dimensionless. In the case of Fourier transforms in infrared spectroscopy, x is a measure of time (the time-domain signal) and s is the inverse of time, or frequency ν (the frequency-domain signal).

The Fourier transform is given by

$$F(s) \equiv \int_{-\infty}^{\infty} f(x) e^{-2\pi i s x} \, \mathrm{d}x \tag{A.12}$$

for any real number s where x is time. This is the forward transform, and the inverse transform is given as

$$f(x) \equiv \int_{-\infty}^{\infty} F(s) e^{-2\pi i s x} \, \mathrm{d}s \tag{A.13}$$

Alternative definitions of the Fourier transform may have different signs in the exponential. Due to their reversible nature the Fourier transform is often denoted with $\Leftrightarrow$ e.g. $F(s) \Leftrightarrow f(x)$.

A.2 Quantum Mechanical Concepts

A.2.1 Electronic Structure Methods

In quantum mechanics one is generally interested in solving the Schrödinger equation. For a system of n particles the time-independent Schrödinger equation takes the form of

$$\hat{H}\Psi = E\Psi \tag{A.14}$$

where $\hat{H}$ is the Hamiltonian operator, Ψ is the wave function being operated on and E is the energy of the wave function Ψ. The Hamiltonian operator characterises the total energy for a given wave function and is often expressed in terms of operators corresponding to the kinetic and potential energy of the system e.g.

$$\hat{H} = \hat{T} + \hat{V} \tag{A.15}$$

where $\hat{V}$ is the potential energy operator and $\hat{T}$ is the kinetic energy operator.

The major drawback of the Schrödinger equation is that it cannot be solved exactly for systems larger than a single hydrogen atom, which means that an exact solution only exists for the hydrogen atom, which has only a single electron. Since we are interested in many-electron systems, methods must be used that find *approximate* solutions to the Schrödinger equation that describe real systems of interest.

One important approximation is the Born-Oppenheimer approximation, which assumes that the wavefunction can be separated into its nuclear and electronic components. This allows us to completely neglect the coupling between nuclear and electronic motion, which simplifies the solution of the Schrödinger equation. The nuclear positions and charges can be treated as fixed parameters and only the electronic portion of the wavefunction needs to be solved.

A.2.2 Hartree-Fock Theory

The Hartree-Fock method is one way in which an approximate solution to the time-independent Schrödinger equation for a system with multiple electrons can be found.The many-electron wavefunction is expressed as the product of single electron wavefuctions and the electrons are considered to occupy single-particle orbitals. These orbitals are assumed to be independent of one another and each electron experiences only an average interactions with the other electrons in the system. The Hartree-Fock equations need to be solved iteratively, since the solution of one electronic orbital necessarily depends on the solution of the others. Higher levels of quantum theory e.g. Density Functional Theory and Coupled Cluster theory can be used to improve upon the results obtained by Hartree-Fock theory.

The Hartree-Fock equations give an approximation of the wave function for a single particle. In order to find numerical solutions to the Schrödinger equation for systems with many particles the orbitals are expanded in a basis set. A basis set is a group of basis functions that can be summed together to represent an unknown quantity, such as the orbitals in the Hartree-Fock equations. In order to exactly find the unknown function an infinite set of basis functions would be required. Since this is obviously impossible to compute, a finite set of basis functions are used to achieve an approximate value for the unknown function.

There are two main types of basis sets used in quantum chemistry packages: those that use Slater-type orbitals and those that use Gaussian-type orbitals. Slater-type orbitals give rise to a better representation of electronic behaviour near the nucleus and less STO are needed than GTO to achieve the same level of accuracy. However, integrals of STO are difficult to evaluate, and for some cases impossible.correct short-range and long-range behaviour. It is common to use sufficient linear combinations of Gaussian-type orbitals to mimic a Slater-type.

A.2.3 Density Functional Theory

Another method of finding an approximate solution to the Schödinger wave equation is density functional theory (DFT). In DFT, the properties of a many-electron system are determined using functionals—functions of another function—in this case the electron density. DFT therefore is the method of using functionals of the electron density to approximate the many-body wave-function.

Modern DFT methods have their theoretical underpinnings in the Hohenberg-Kohn theorems which are as follows:

1 *If two systems of electrons, one trapped in a potential* $v_1(\vec{r})$*, and the other in* $v_2(\vec{r})$*, have the same ground state density* $n(\vec{r})$ *then* $v_1(\vec{r}) - v_1(\vec{r}) = constant$.

2 *For any positive integer, N, and potential* $v(\vec{r})$*, a density functional,* $F[n]$ *exists such that* $E_{(v,N)}[n] = F[n] + \int v(\vec{r})n(\vec{r})\,\mathrm{d}^3r$ *obtains its minimal value at the ground-state density of N electrons in the potential* $v(\vec{r})$*. The minimal value of* $E_{(v,N)}[n]$ *is then the ground state energy of this system.*

The first Hohenberg-Kohn theorem demonstrates that the ground state properties of a many-electron system are uniquely determined by an electron density that depends on only 3 spatial coordinates. As a result of this, the many-body problem of N electrons with $3N$ spatial coordinates is reduced to 3 spatial coordinates through the use of functionals of the electron density. The second theorem defines an energy functional for the system and proves that the correct ground state electron density minimizes this energy functional.

Within this framework the intractable many-body problem of interacting electrons in a static external potential becomes a tractable problem of non-interacting electrons in an effective potential. This effective potential includes the effects of the Coulomb interactions between electrons including exchange and correlation interactions. For most systems exact functionals for exchange an correlation are not known and must be approximated.

Appendix B

Parameters for Isoniazid NADH adduct

```
;;
;; Generated by CHARMM-GUI (http://www.charmm-gui.org) v1.6
;;
;; psf2itp.py
;;
;; Correspondance:
;; j712l362@ku.edu or wonpil@ku.edu
;;
;; GROMACS topology file for HETA
;;
[ moleculetype ]
; name nrexcl
HETA 3
[ atoms ]
; nr type resnr residu atom cgnr charge mass
1 OG3O3 300 ZID O10 1 -0.618 15.9994 ; qtot -0.618
2 PG1 300 ZID P2 2 1.468 30.9738 ; qtot 0.850
3 OG2P1 300 ZID O9 3 -0.817 15.9994 ; qtot 0.033
4 OG2P1 300 ZID O8 4 -0.817 15.9994 ; qtot -0.784
5 OG304 300 ZID O7 5 -0.786 15.9994 ; qtot -1.570
6 PG0 300 ZID P1 6 1.746 30.9738 ; qtot 0.176
7 OG2P1 300 ZID O1 7 -0.643 15.9994 ; qtot -0.467
8 OG3O3 300 ZID O3 8 -0.476 15.9994 ; qtot -0.943
9 CG321 300 ZID C1 9 -0.156 12.0110 ; qtot -1.099
10 HGA2 300 ZID H2 10 0.090 1.0080 ; qtot -1.009
11 HGA2 300 ZID H3 11 0.090 1.0080 ; qtot -0.919
12 CG3C51 300 ZID C2 12 0.114 12.0110 ; qtot -0.805
13 CG3C51 300 ZID C3 13 0.141 12.0110 ; qtot -0.664
14 HGA1 300 ZID H4 14 0.090 1.0080 ; qtot -0.574
15 HGA1 300 ZID H5 15 0.090 1.0080 ; qtot -0.484
16 OG311 300 ZID O5 16 -0.649 15.9994 ; qtot -1.133
17 HGP1 300 ZID H6 17 0.421 1.0080 ; qtot -0.712
18 CG3C51 300 ZID C4 18 0.137 12.0110 ; qtot -0.575
19 OG311 300 ZID O6 19 -0.649 15.9994 ; qtot -1.224
20 HGA1 300 ZID H7 20 0.090 1.0080 ; qtot -1.134
21 HGP1 300 ZID H8 21 0.421 1.0080 ; qtot -0.713
22 CG3C51 300 ZID C5 22 0.116 12.0110 ; qtot -0.597
23 NG2R51 300 ZID N1 23 -0.054 14.0070 ; qtot -0.651
24 HGA1 300 ZID H9 24 0.090 1.0080 ; qtot -0.561
25 CG2RC0 300 ZID C10 25 0.428 12.0110 ; qtot -0.133
26 CG2RC0 300 ZID C7 26 0.281 12.0110 ; qtot 0.148
27 NG2R50 300 ZID N2 27 -0.709 14.0070 ; qtot -0.561
28 CG2R53 300 ZID C6 28 0.338 12.0110 ; qtot -0.223
29 HGR52 300 ZID H10 29 0.131 1.0080 ; qtot -0.092
30 CG2R64 300 ZID C8 30 0.457 12.0110 ; qtot 0.365
31 NG2R62 300 ZID N4 31 -0.741 14.0070 ; qtot -0.376
32 CG2R64 300 ZID C9 32 0.500 12.0110 ; qtot 0.124
33 NG2R62 300 ZID N5 33 -0.750 14.0070 ; qtot -0.626
34 HGR62 300 ZID H13 34 0.127 1.0080 ; qtot -0.499
35 NG2S3 300 ZID N3 35 -0.769 14.0070 ; qtot -1.268
36 HGP4 300 ZID H11 36 0.379 1.0080 ; qtot -0.889
```

```
37 HGP4 300 ZID H12 37 0.379 1.0080 ; qtot -0.510
38 OG3C51 300 ZID O4 38 -0.414 15.9994 ; qtot -0.924
39 OG311 300 ZID O2 39 -0.590 15.9994 ; qtot -1.514
40 HGP1 300 ZID H1 40 0.420 1.0080 ; qtot -1.094
41 CG321 300 ZID C11 41 -0.081 12.0110 ; qtot -1.175
42 CG3C51 300 ZID C12 42 0.111 12.0110 ; qtot -1.064
43 HGA2 300 ZID H14 43 0.090 1.0080 ; qtot -0.974
44 HGA2 300 ZID H15 44 0.090 1.0080 ; qtot -0.884
45 HGA1 300 ZID H16 45 0.090 1.0080 ; qtot -0.794
46 CG3C51 300 ZID C13 46 0.137 12.0110 ; qtot -0.657
47 OG311 300 ZID O12 47 -0.650 15.9994 ; qtot -1.307
48 HGA1 300 ZID H17 48 0.090 1.0080 ; qtot -1.217
49 HGP1 300 ZID H18 49 0.421 1.0080 ; qtot -0.796
50 CG3C51 300 ZID C14 50 0.134 12.0110 ; qtot -0.662
51 OG311 300 ZID O13 51 -0.651 15.9994 ; qtot -1.313
52 HGA1 300 ZID H19 52 0.090 1.0080 ; qtot -1.223
53 HGP1 300 ZID H20 53 0.421 1.0080 ; qtot -0.802
54 CG3C53 300 ZID C15 54 0.114 12.0110 ; qtot -0.688
55 NG2R61 300 ZID N6 55 -0.073 14.0070 ; qtot -0.761
56 HGA1 300 ZID H21 56 0.090 1.0080 ; qtot -0.671
57 CG2R62 300 ZID C21 57 0.108 12.0110 ; qtot -0.563
58 CG2R62 300 ZID C20 58 -0.038 12.0110 ; qtot -0.601
59 HGR63 300 ZID H26 59 0.189 1.0080 ; qtot -0.412
60 HGR63 300 ZID H25 60 0.129 1.0080 ; qtot -0.283
61 CG2R62 300 ZID C19 61 0.191 12.0110 ; qtot -0.092
62 CG2O5 300 ZID C22 62 0.422 12.0110 ; qtot 0.330
63 OG2D3 300 ZID O15 63 -0.461 15.9994 ; qtot -0.131
64 CG2R61 300 ZID C25 64 0.005 12.0110 ; qtot -0.126
65 CG2R61 300 ZID C26 65 -0.116 12.0110 ; qtot -0.242
66 HGR61 300 ZID H29 66 0.115 1.0080 ; qtot -0.127
67 CG2R61 300 ZID C27 67 0.177 12.0110 ; qtot 0.050
68 NG2R60 300 ZID N8 68 -0.598 14.0070 ; qtot -0.548
69 HGR62 300 ZID H30 69 0.121 1.0080 ; qtot -0.427
70 CG2R61 300 ZID C23 70 0.177 12.0110 ; qtot -0.250
71 CG2R61 300 ZID C24 71 -0.116 12.0110 ; qtot -0.366
72 HGR62 300 ZID H27 72 0.121 1.0080 ; qtot -0.245
73 HGR61 300 ZID H28 73 0.115 1.0080 ; qtot -0.130
74 CG2R62 300 ZID C17 74 0.091 12.0110 ; qtot -0.039
75 CG2O1 300 ZID C18 75 0.671 12.0110 ; qtot 0.632
76 OG2D1 300 ZID O14 76 -0.400 15.9994 ; qtot 0.232
77 NG2S2 300 ZID N7 77 -0.820 14.0070 ; qtot -0.588
78 HGP1 300 ZID H24 78 0.350 1.0080 ; qtot -0.238
79 HGP1 300 ZID H23 79 0.350 1.0080 ; qtot 0.112
80 CG2R62 300 ZID C16 80 0.126 12.0110 ; qtot 0.238
81 HGR63 300 ZID H22 81 0.159 1.0080 ; qtot 0.397
82 OG3C51 300 ZID O11 82 -0.397 15.9994 ; qtot 0.000
[ bonds ]
; ai aj funct b0 Kb
1 2 1
1 41 1
2 3 1
2 4 1
2 5 1
5 6 1
6 7 1
6 8 1
6 39 1
8 9 1
9 10 1
9 11 1
9 12 1
12 13 1
12 14 1
12 38 1
13 15 1
13 16 1
13 18 1
16 17 1
18 19 1
18 20 1
18 22 1
19 21 1
22 23 1
```

```
22 24 1
22 38 1
23 25 1
23 28 1
25 26 1
25 33 1
26 27 1
26 30 1
27 28 1
28 29 1
30 31 1
30 35 1
31 32 1
32 33 1
32 34 1
35 36 1
35 37 1
39 40 1
41 42 1
41 43 1
41 44 1
42 45 1
42 46 1
42 82 1
46 47 1
46 48 1
46 50 1
47 49 1
50 51 1
50 52 1
50 54 1
51 53 1
54 55 1
54 56 1
54 82 1
55 57 1
55 80 1
57 58 1
57 59 1
58 60 1
58 61 1
61 62 1
61 74 1
62 63 1
62 64 1
64 65 1
64 71 1
65 66 1
65 67 1
67 68 1
67 69 1
68 70 1
70 71 1
70 72 1
71 73 1
74 75 1
74 80 1
75 76 1
75 77 1
77 78 1
77 79 1
80 81 1
[ pairs ]
; ai aj funct c6 c12
1 6 1
1 45 1
1 46 1
1 82 1
2 7 1
2 8 1
2 39 1
2 42 1
2 43 1
```

```
2 44 1
3 6 1
3 41 1
4 6 1
4 41 1
5 9 1
5 40 1
5 41 1
6 10 1
6 11 1
6 12 1
7 9 1
7 40 1
8 13 1
8 14 1
8 38 1
8 40 1
9 15 1
9 16 1
9 18 1
9 22 1
9 39 1
10 13 1
10 14 1
10 38 1
11 13 1
11 14 1
11 38 1
12 17 1
12 19 1
12 20 1
12 23 1
12 24 1
13 21 1
13 23 1
13 24 1
14 15 1
14 16 1
14 18 1
14 22 1
15 17 1
15 19 1
15 20 1
15 22 1
15 38 1
16 19 1
16 20 1
16 22 1
16 38 1
17 18 1
18 25 1
18 28 1
19 23 1
19 24 1
19 38 1
20 21 1
20 23 1
20 24 1
20 38 1
21 22 1
22 26 1
22 27 1
22 29 1
22 33 1
23 30 1
23 32 1
24 25 1
24 28 1
25 29 1
25 31 1
25 34 1
25 35 1
25 38 1
```

```
26 29 1
26 32 1
26 36 1
26 37 1
27 31 1
27 33 1
27 35 1
28 30 1
28 33 1
28 38 1
30 33 1
30 34 1
31 36 1
31 37 1
32 35 1
41 47 1
41 48 1
41 50 1
41 54 1
42 49 1
42 51 1
42 52 1
42 55 1
42 56 1
43 45 1
43 46 1
43 82 1
44 45 1
44 46 1
44 82 1
45 47 1
45 48 1
45 50 1
45 54 1
46 53 1
46 55 1
46 56 1
47 51 1
47 52 1
47 54 1
47 82 1
48 49 1
48 51 1
48 52 1
48 54 1
48 82 1
49 50 1
50 57 1
50 80 1
51 55 1
51 56 1
51 82 1
52 53 1
52 55 1
52 56 1
52 82 1
53 54 1
54 58 1
54 59 1
54 74 1
54 81 1
55 60 1
55 61 1
55 75 1
56 57 1
56 80 1
57 62 1
57 74 1
57 81 1
57 82 1
58 63 1
58 64 1
58 75 1
```

```
58 80 1
59 60 1
59 61 1
59 80 1
60 62 1
60 74 1
61 65 1
61 71 1
61 76 1
61 77 1
61 81 1
62 66 1
62 67 1
62 70 1
62 73 1
62 75 1
62 80 1
63 65 1
63 71 1
63 74 1
64 68 1
64 69 1
64 72 1
64 74 1
65 70 1
65 73 1
66 68 1
66 69 1
66 71 1
67 71 1
67 72 1
68 73 1
69 70 1
72 73 1
74 78 1
74 79 1
75 81 1
76 78 1
76 79 1
76 80 1
77 80 1
80 82 1
[ angles ]
; ai aj ak funct th0 cth S0 Kub
2 1 41 5
1 2 3 5
1 2 4 5
1 2 5 5
3 2 4 5
3 2 5 5
4 2 5 5
2 5 6 5
5 6 7 5
5 6 8 5
5 6 39 5
7 6 8 5
7 6 39 5
8 6 39 5
6 8 9 5
8 9 10 5
8 9 11 5
8 9 12 5
10 9 11 5
10 9 12 5
11 9 12 5
9 12 13 5
9 12 14 5
9 12 38 5
13 12 14 5
13 12 38 5
14 12 38 5
12 13 15 5
12 13 16 5
```

```
12 13 18 5
15 13 16 5
15 13 18 5
16 13 18 5
13 16 17 5
13 18 19 5
13 18 20 5
13 18 22 5
19 18 20 5
19 18 22 5
20 18 22 5
18 19 21 5
18 22 23 5
18 22 24 5
18 22 38 5
23 22 24 5
23 22 38 5
24 22 38 5
22 23 25 5
22 23 28 5
25 23 28 5
23 25 26 5
23 25 33 5
26 25 33 5
25 26 27 5
25 26 30 5
27 26 30 5
26 27 28 5
23 28 27 5
23 28 29 5
27 28 29 5
26 30 31 5
26 30 35 5
31 30 35 5
30 31 32 5
31 32 33 5
31 32 34 5
33 32 34 5
25 33 32 5
30 35 36 5
30 35 37 5
36 35 37 5
12 38 22 5
6 39 40 5
1 41 42 5
1 41 43 5
1 41 44 5
42 41 43 5
42 41 44 5
43 41 44 5
41 42 45 5
41 42 46 5
41 42 82 5
45 42 46 5
45 42 82 5
46 42 82 5
42 46 47 5
42 46 48 5
42 46 50 5
47 46 48 5
47 46 50 5
48 46 50 5
46 47 49 5
46 50 51 5
46 50 52 5
46 50 54 5
51 50 52 5
51 50 54 5
52 50 54 5
50 51 53 5
50 54 55 5
50 54 56 5
50 54 82 5
```

```
55 54 56 5
55 54 82 5
56 54 82 5
54 55 57 5
54 55 80 5
57 55 80 5
55 57 58 5
55 57 59 5
58 57 59 5
57 58 60 5
57 58 61 5
60 58 61 5
58 61 62 5
58 61 74 5
62 61 74 5
61 62 63 5
61 62 64 5
63 62 64 5
62 64 65 5
62 64 71 5
65 64 71 5
64 65 66 5
64 65 67 5
66 65 67 5
65 67 68 5
65 67 69 5
68 67 69 5
67 68 70 5
68 70 71 5
68 70 72 5
71 70 72 5
64 71 70 5
64 71 73 5
70 71 73 5
61 74 75 5
61 74 80 5
75 74 80 5
74 75 76 5
74 75 77 5
76 75 77 5
75 77 78 5
75 77 79 5
78 77 79 5
55 80 74 5
55 80 81 5
74 80 81 5
42 82 54 5
[ dihedrals ]
; ai aj ak al funct phi0 cp mult
2 1 41 42 9
2 1 41 43 9
2 1 41 44 9
3 2 1 41 9
4 2 1 41 9
5 2 1 41 9
1 2 5 6 9
3 2 5 6 9
4 2 5 6 9
2 5 6 7 9
2 5 6 8 9
2 5 6 39 9
5 6 8 9 9
7 6 8 9 9
5 6 39 40 9
7 6 39 40 9
8 6 39 40 9
9 8 6 39 9
6 8 9 10 9
6 8 9 11 9
6 8 9 12 9
8 9 12 13 9
8 9 12 14 9
8 9 12 38 9
```

10 9 12 13 9
10 9 12 14 9
10 9 12 38 9
11 9 12 13 9
11 9 12 14 9
11 9 12 38 9
9 12 13 15 9
9 12 13 16 9
9 12 13 18 9
14 12 13 15 9
14 12 13 16 9
14 12 13 18 9
9 12 38 22 9
13 12 38 22 9
14 12 38 22 9
15 13 12 38 9
16 13 12 38 9
18 13 12 38 9
12 13 16 17 9
15 13 16 17 9
12 13 18 19 9
12 13 18 20 9
12 13 18 22 9
15 13 18 19 9
15 13 18 20 9
15 13 18 22 9
16 13 18 19 9
16 13 18 20 9
16 13 18 22 9
17 16 13 18 9
13 18 19 21 9
20 18 19 21 9
13 18 22 23 9
13 18 22 24 9
13 18 22 38 9
19 18 22 23 9
19 18 22 24 9
19 18 22 38 9
20 18 22 23 9
20 18 22 24 9
20 18 22 38 9
21 19 18 22 9
18 22 23 25 9
18 22 23 28 9
24 22 23 25 9
24 22 23 28 9
25 23 22 38 9
28 23 22 38 9
22 23 25 26 9
22 23 25 33 9
28 23 25 33 9
22 23 28 27 9
22 23 28 29 9
25 23 28 27 9
25 23 28 29 9
26 25 23 28 9
23 25 26 27 9
23 25 26 30 9
23 25 33 32 9
26 25 33 32 9
27 26 25 33 9
30 26 25 33 9
25 26 27 28 9
25 26 30 31 9
25 26 30 35 9
27 26 30 31 9
27 26 30 35 9
28 27 26 30 9
26 27 28 29 9
23 28 27 26 9
26 30 31 32 9
26 30 35 36 9
26 30 35 37 9

```
31 30 35 36 9
31 30 35 37 9
32 31 30 35 9
30 31 32 33 9
30 31 32 34 9
25 33 32 31 9
25 33 32 34 9
12 38 22 18 9
12 38 22 23 9
12 38 22 24 9
1 41 42 45 9
1 41 42 46 9
1 41 42 82 9
43 41 42 45 9
43 41 42 46 9
43 41 42 82 9
44 41 42 45 9
44 41 42 46 9
44 41 42 82 9
41 42 46 47 9
41 42 46 48 9
41 42 46 50 9
45 42 46 47 9
45 42 46 48 9
45 42 46 50 9
41 42 82 54 9
45 42 82 54 9
46 42 82 54 9
47 46 42 82 9
48 46 42 82 9
50 46 42 82 9
42 46 47 49 9
48 46 47 49 9
42 46 50 51 9
42 46 50 52 9
42 46 50 54 9
47 46 50 51 9
47 46 50 52 9
47 46 50 54 9
48 46 50 51 9
48 46 50 52 9
48 46 50 54 9
49 47 46 50 9
46 50 51 53 9
52 50 51 53 9
46 50 54 55 9
46 50 54 56 9
46 50 54 82 9
51 50 54 55 9
51 50 54 56 9
51 50 54 82 9
52 50 54 55 9
52 50 54 56 9
52 50 54 82 9
53 51 50 54 9
50 54 55 57 9
50 54 55 80 9
56 54 55 57 9
56 54 55 80 9
57 55 54 82 9
80 55 54 82 9
54 55 57 58 9
54 55 57 59 9
54 55 80 74 9
54 55 80 81 9
57 55 80 74 9
57 55 80 81 9
58 57 55 80 9
59 57 55 80 9
55 57 58 60 9
55 57 58 61 9
59 57 58 60 9
59 57 58 61 9
```

```
57 58 61 62 9
57 58 61 74 9
60 58 61 62 9
60 58 61 74 9
58 61 62 63 9
58 61 62 64 9
58 61 74 75 9
58 61 74 80 9
62 61 74 75 9
62 61 74 80 9
63 62 61 74 9
64 62 61 74 9
61 62 64 65 9
61 62 64 71 9
63 62 64 65 9
63 62 64 71 9
62 64 65 66 9
62 64 65 67 9
62 64 71 70 9
62 64 71 73 9
65 64 71 70 9
65 64 71 73 9
66 65 64 71 9
67 65 64 71 9
64 65 67 68 9
64 65 67 69 9
66 65 67 68 9
66 65 67 69 9
65 67 68 70 9
69 67 68 70 9
67 68 70 71 9
67 68 70 72 9
68 70 71 73 9
72 70 71 73 9
64 71 70 68 9
64 71 70 72 9
61 74 75 76 9
61 74 75 77 9
61 74 80 81 9
75 74 80 81 9
76 75 74 80 9
77 75 74 80 9
74 75 77 78 9
74 75 77 79 9
76 75 77 78 9
76 75 77 79 9
55 80 74 61 9
55 80 74 75 9
42 82 54 50 9
42 82 54 55 9
42 82 54 56 9
[ dihedrals ]
; ai aj ak al funct q0 cq
35 37 36 30 2
62 64 61 63 2
75 74 77 76 2
```

References

[1] RCSB Protein Data Bank. PDB Current Holdings Breakdown. *http://www.rcsb.org/pdb/statistics/holdings.do* [Accessed 25/08/2015].

[2] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–99, 1963.

[3] S. C. Lovell, I. W. Davis, W. B. Adrendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by C Alpha geometry: Phi, Psi and C Beta deviation. *Proteins*, 50:437–450, 2003.

[4] C.M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

[5] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[6] K.A. Dill and J.L. MacCallum. The Protein-Folding Problem, 50 Years On. *Science*, 338:1042–1046, 2012.

[7] E. Fischer. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dt. Chem. Ges*, 27:2985–2993, 1894.

[8] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, 44:98–104, 1958.

[9] C. Micheletti. Comparing proteins by their internal dynamics: Exploring structurefunction relationships beyond static structural alignments. *Phys. Life Rev.*, 10:1–26, 2013.

[10] P.C. Biggin. Protein dynamics — a moving target: Comment on "Comparing proteins by their internal dynamics: Exploring structurefunction relationships beyond static structural alignments" by C. Micheletti . *Phys. Life Rev.*, 10:27–28, 2013.

[11] S. Hammes-Schiffer and J. Klinman. Emerging Concepts about the Role of Protein Motion in Enzyme Catalysis. *Acc. Chem. Res.*, 48:899–899, 2015.

[12] J.A. McCammon. Molecular dynamics of the bovine pancreatic trypsin inhibitor. In H.J.C. Berendsen, editor, *Report of the 1976 Workshop, Models for Protein Dynamics*. 1977.

[13] J. A. McCammon, B. R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[14] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14:437–449, 2006.

[15] G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A.M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497:643–646, 2013.

[16] T. Reddy, D. Shorthouse, D.L. Parton, E. Jefferys, P.W. Fowler, M. Chavent, M. Baaden, and M.S.P. Sansom. Nothing to sneeze at: A dynamic and integrative computational model of an influenza A virion. *Structure*, 23:584–597, 2015.

[17] G. Jayachandran, V. Vishal, and V. S. Pande. Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Chem. Phys.*, 124, 2006.

[18] J. A. Baker and J. D. Hirst. Molecular dynamics simulations using graphics processing units. *Mol. Inform.*, 30:498–504, 2011.

[19] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.

[20] A. E. Warshel. Bicycle-pedal model for the first step in the vision process. *Nature*, 260:679–683, 1976.

[21] A. Huerta-Viga, D. J. Shaw, and S. Woutersen. pH dependence of the conformation of small peptides investigated with two-dimensional vibrational spectroscopy. *J. Phys. Chem. B*, 114:15212–15220, 2010.

[22] D. P. Weliky and R. Tycko. Determination of peptide conformations by two-dimensional magic angle spinning NMR exchange spectroscopy with rotor synchronization. *J. Am. Chem. Soc.*, 118:8487–8488, 1996.

[23] C. Liang, J. Knoester, and T. L. C. Jansen. Proton transport in a membrane protein channel: Two-dimensional infrared spectrum modeling. *J. Phys. Chem. B*, 116:6336–6345, 2012.

[24] Z. Ganim, A. Tokmakoff, and A. Vaziri. Vibrational excitons in ionophores: experimental probes for quantum coherence-assisted ion transport and selectivity in ion channels. *New J. Phys.*, 13:113030, 2011.

[25] C. Falvo, W. Zhuang, Y. S. Kim, P. H. Axelsen, R. M. Hochstrasser, and S. Mukamel. Frequency distribution of the amide-I vibration sorted by residues in amyloid fibrils revealed by 2D-IR measurements and simulations. *J. Phys. Chem. B*, 116:3322–3330, 2012.

[26] A. Barth. Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1767:1073–1101, 2007.

[27] S. Krimm and J. Bandekar. Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Protein Chem.*, 38:181–364, 1986.

[28] L.P. DeFlores, Z. Ganim, R.A. Nicodemus, and A. Tokmakoff. Amide I'-II' 2D IR spectroscopy provides enhanced protein secondary structural sensitivity. *J. Am. Chem. Soc.*, 131:3385–3391, 2009.

[29] C.R. Baiz, M. Reppert, and A. Tokmakoff. Introduction to protein 2D IR spectroscopy. In M.D. Fayer, editor, *Ultrafast Infrared Vibrational Spectroscopy*. Taylor & Francis.

[30] A. W. Smith and A. Tokmakoff. Amide I two-dimensional infrared spectroscopy of beta-hairpin peptides. *J. Chem. Phys.*, 126:045109, 2007.

[31] J. Dimaio, T. M. D. Nguyen, C. Lemieux, and P. W. Schiller. Synthesis And Pharmacological Characterization In vitro Of Cyclic Enkephalin Analogs - Effect Of Conformational Constraints On Opiate Receptor Selectivity. *J. Med. Chem.*, 25:1432–1438, 1982.

[32] V. Boguslavsky, V. J. Hruby, D. F. O'Brien, A. Misicka, and A. W. Lipkowski. Effect of peptide conformation on membrane permeability. *J. Pept. Res*, 61:287–297, 2003.

[33] R.E. Hill, N.T. Hunt, and J.D. Hirst. Studying biomacromolecules with two-dimensional infrared spectroscopy. In C. Christov, editor,

Biomolecular Spectroscopy: Advances from Integrating Experiments and Theory, volume 93, pages 1–36. Academic Press, 2013.

[34] P. Hamm, M. H. Lim, and R. M. Hochstrasser. Structure of the amide I band of peptides measured by femtosecond nonlinear-infrared spectroscopy. *J. Phys. Chem. B*, 102:6123–6138, 1998.

[35] M. C. Asplund, M. T. Zanni, and R. M. Hochstrasser. Two-dimensional infrared spectroscopy of peptides by phase-controlled femtosecond vibrational photon echoes. *Proc. Natl. Acad. Sci. USA*, 97:8219–8224, 2000.

[36] V. Cervetto, J. Helbing, J. Bredenbeck, and P. Hamm. Double-resonance versus pulsed Fourier transform two-dimensional infrared spectroscopy: An experimental and theoretical comparison. *J. Chem. Phys.*, 121:5935–5942, 2004.

[37] S. Mukamel. *Principles of Nonlinear Optics and Spectroscopy*. Oxford University Press, New York, 1995.

[38] P. Hamm and M. T. Zanni. *Concepts and methods of 2D infrared spectroscopy*. Cambridge University Press, Cambridge, 2011.

[39] N. T. Hunt. 2D-IR spectroscopy: ultrafast insights into biomolecule structure and function. *Chem. Soc. Rev.*, 38:1837–1848, 2009.

[40] Y. Tanimura and S. Mukamel. 2-dimensional femtosecond vibrational spectroscopy of liquids. *J. Chem. Phys.*, 99:9496–9511, 1993.

[41] A.W. Smith, J. Lessing, Z. Ganim, C.S. Peng, A. Tokmakoff, S. Roy, T.L.C. Jansen, and J. Knoester. Melting of a beta-Hairpin Peptide Using Isotope-Edited 2D IR Spectroscopy and Simulations. *J. Phys. Chem. B*, 114:10913–10924, 2010.

[42] C. R. Baiz, C. S. Peng, M. E. Reppert, K. C. Jones, and A. Tokmakoff. Coherent two-dimensional infrared spectroscopy: Quantitative analysis of protein secondary structure in solution. *Analyst*, 137:1793–1799, 2012.

[43] H. S. Chung and A. Tokmakoff. Temperature-dependent downhill unfolding of ubiquitin. I. nanosecond-to-millisecond resolved nonlinear infrared spectroscopy. *Proteins*, 72:474–487, 2008.

[44] S. Roy, T. L. C. Jansen, and J. Knoester. Structural classification of the amide I sites of a beta-hairpin with isotope label 2DIR spectroscopy. *Phys. Chem. Chem. Phys.*, 12:9347–9357, 2010.

[45] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik. Tryptophan zippers: Stable, monomeric beta-hairpins. *Proc. Natl. Acad. Sci. USA*, 98:5578–5583, 2001.

[46] J. P. Wang, J. X. Chen, and R. M. Hochstrasser. Local structure of beta-hairpin isotopomers by FTIR, 2D IR, and ab initio theory. *J. Phys. Chem. B*, 110:7545–7555, 2006.

[47] J. P. Wang, W. Zhuang, S. Mukamel, and R. M. Hochstrasser. Two-dimensional infrared spectroscopy as a probe of the solvent electrostatic field for a twelve residue peptide. *J. Phys. Chem. B*, 112:5930–5937, 2008.

[48] J. Lessing, S. Roy, M. Reppert, M. Baer, D. Marx, T. L. C. Jansen, J. Knoester, and A. Tokmakoff. Identifying residual structure in intrinsically disordered systems: A 2D IR spectroscopic study of the GVGXPGVG peptide. *J. Am. Chem. Soc.*, 134:5032–5035, 2012.

[49] T. L. C. Jansen and J. Knoester. Nonadiabatic effects in the two-dimensional infrared spectra of peptides: Application to alanine dipeptide. *J. Phys. Chem. B*, 110:22910–22916, 2006.

[50] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comp. Chem.*, 25:1605–1612, 2004.

[51] A. M. Woys, A. M. Almeida, L. Wang, C. C. Chiu, M. McGovern, J. J. de Pablo, J. L. Skinner, S. H. Gellman, and M. T. Zanni. Parallel beta-sheet vibrational couplings revealed by 2D IR spectroscopy of an isotopically labeled macrocycle: Quantitative benchmark for the interpretation of amyloid and protein infrared spectra. *J. Am. Chem. Soc.*, 134:19118–19128, 2012.

[52] A.T. Krummel and M.T. Zanni. Evidence for coupling between nitrile groups using DNA templates: A promising new method for monitoring structures with infrared spectroscopy. *J. Phys. Chem. B*, 112:1336–1338, 2008.

[53] H. Torii and M. Tasumi. Model-calculations on the amide-I infrared bands of globular-proteins. *J. Chem. Phys.*, 96:3379–3387, 1992.

[54] P. E. Wright, H. J. Dyson, and R. A. Lerner. Conformation of peptide-fragments of proteins in aqueous-solution - implications for initiation of protein folding. *Biochemistry*, 27:7167–7175, 1988.

[55] F. Eker, X. L. Cao, L. Nafie, and R. Schweitzer-Stenner. Tripeptides adopt stable structures in water. a combined polarized visible Raman, FTIR, and VCD spectroscopy study. *J. Am. Chem. Soc.*, 124:14330–14341, 2002.

[56] Y. Xiang, L. Duan, and J. Z. H. Zhang. Folding dynamics of a small protein at room temperature via simulated coherent two-dimensional infrared spectroscopy. *Phys. Chem. Chem. Phys.*, 12:15681–15688, 2010.

[57] T. Hayashi, W. Zhuang, and S. Mukamel. Electrostatic DFT map for the complete vibrational amide band of NMA. *J. Phys. Chem. A*, 109:9747–9759, 2005.

[58] S. Mukamel and D. Abramavicius. Many-body approaches for simulating coherent nonlinear spectroscopies of electronic and vibrational excitons. *Chem. Rev.*, 104:2073–2098, 2004.

[59] P. Hamm, J. Helbing, and J. Bredenbeck. Two-dimensional infrared spectroscopy of photoswitchable peptides. In *Annual Review of Physical Chemistry*.

[60] C. J. McKnight, P. T. Matsudaira, and P. S. Kim. NMR structure of the 35-residue villin headpiece subdomain. *Nat. Struct. Biol.*, 4:180–184, 1997.

[61] J. Kubelka, E. R. Henry, T. Cellmer, J. Hofrichter, and W. A. Eaton. Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proc. Natl. Acad. Sci. USA*, 105:18655–18662, 2008.

[62] Z. Getahun, C. Y. Huang, T. Wang, B. De Leon, W. F. DeGrado, and F. Gai. Using nitrile-derivatized amino acids as infrared probes of local environment. *J. Am. Chem. Soc.*, 125:405–411, 2003.

[63] J. K. Chung, M. C. Thielges, and M. D. Fayer. Dynamics of the folded and unfolded villin headpiece (HP35) measured with ultrafast 2D IR vibrational echo spectroscopy. *Proc. Natl. Acad. Sci. USA*, 108:3578–3583, 2011.

[64] M. M. Waegele, R. M. Culik, and F. Gai. Site-specific spectroscopic reporters of the local electric field, hydration, structure, and dynamics of biomolecules. *J. Phys. Chem. Lett.*, 2:2598–2609, 2011.

[65] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter. Sub-microsecond protein folding. *J. Mol. Biol.*, 359:546–553, 2006.

[66] C. Kolano, J. Helbing, M. Kozinski, W. Sander, and P. Hamm. Watching hydrogen-bond dynamics in a beta-turn by transient two-dimensional infrared spectroscopy. *Nature*, 444:469–472, 2006.

[67] Y. S. Kim and R. M. Hochstrasser. Applications of 2D IR spectroscopy to peptides, proteins, and hydrogen-bond dynamics. *J. Phys. Chem. B*, 113:8231–8251, 2009.

[68] S. H. Shim, R. Gupta, Y. L. Ling, D. B. Strasfeld, D. P. Raleigh, and M. T. Zanni. Two-dimensional IR spectroscopy and isotope labeling defines the pathway of amyloid formation with residue-specific resolution. *Proc. Natl. Acad. Sci. USA*, 106:6614–6619, 2009.

[69] J. T. King and K. J. Kubarych. Site-specific coupling of hydration water and protein flexibility studied in solution with ultrafast 2D-IR spectroscopy. *J. Am. Chem. Soc.*, 134:18705–18712, 2012.

[70] R. V. Rariy and A. M. Klibanov. Correct protein folding in glycerol. *Proc. Natl. Acad. Sci. USA*, 94:13520–13523, 1997.

[71] A. Ghosh, J. Qiu, W. F. DeGrado, and R. M. Hochstrasser. Tidal surge in the M2 proton channel, sensed by 2D IR spectroscopy. *Proc. Natl. Acad. Sci. USA*, 108:6115–6120, 2011.

[72] C. Liang, T. L. C. Jansen, and J. Knoester. Proton transport in biological systems can be probed by two-dimensional infrared spectroscopy. *J. Chem. Phys.*, 134:044502, 2011.

[73] G. S. Engel, T. R. Calhoun, E. L. Read, T. K. Ahn, T. Mancal, Y. C. Cheng, R. E. Blankenship, and G. R. Fleming. Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature*, 446:782–786, 2007.

[74] S. Y. Noskov and B. Roux. Ion selectivity in potassium channels. *Biophys. Chem.*, 124:279–291, 2006.

[75] O. S. Makin and L. C. Serpell. Structural characterisation of islet amyloid polypeptide fibrils. *J. Mol. Biol.*, 335:1279–1288, 2004.

[76] S. Ham, J. H. Kim, H. Lee, and M. H. Cho. Correlation between electronic and molecular structure distortions and vibrational properties. II. amide I modes of NMA-nD_2O complexes. *J. Chem. Phys.*, 118:3491–3498, 2003.

[77] S. Ham, S. Cha, J. H. Choi, and M. Cho. Amide I modes of tripeptides: Hessian matrix reconstruction and isotope effects. *J. Chem. Phys.*, 119:1451–1461, 2003.

[78] H. Torii and M. Tasumi. Ab initio molecular orbital study of the amide I vibrational interactions between the peptide groups in di- and tripeptides and considerations on the conformation of the extended helix. *J. Raman Spectros.*, 29:81–86, 1998.

[79] W. Zhuang, N. G. Sgourakis, Z. Li, A. E. Garcia, and S. Mukamel. Discriminating early stage A beta 42 monomer structures using chirality-induced 2DIR spectroscopy in a simulation study. *Proc. Natl. Acad. Sci. USA*, 107:15687–15692, 2010.

[80] P. Cao, F. Meng, A. Abedini, and D. P. Raleigh. The ability of rodent islet amyloid polypeptide to inhibit amyloid formation by human islet amyloid polypeptide has important implications for the mechanism of amyloid formation and the design of inhibitors. *Biochemistry*, 49:872–881, 2010.

[81] C. T. Middleton, P. Marek, P. Cao, C. C. Chiu, S. Singh, A. M. Woys, J. J. de Pablo, D. P. Raleigh, and M. T. Zanni. Two-dimensional infrared spectroscopy reveals the complex behaviour of an amyloid fibril inhibitor. *Nat. Chem.*, 4:355–360, 2012.

[82] Y. S. Kim, L. Liu, P. H. Axelsen, and R. M. Hochstrasser. 2D IR provides evidence for mobile water molecules in beta-amyloid fibrils. *Proc. Natl. Acad. Sci. USA*, 106:17751–17756, 2009.

[83] D. B. Strasfeld, Y. L. Ling, R. Gupta, D. P. Raleigh, and M. T. Zanni. Strategies for extracting structural information from 2D IR spectroscopy of amyloid: Application to islet amyloid polypeptide. *J. Phys. Chem. B*, 113:15679–15691, 2009.

[84] L. Wang, C. T. Middleton, M. T. Zanni, and J. L. Skinner. Development and validation of transferable amide I vibrational frequency maps for peptides. *J. Phys. Chem. B*, 115:3713–3724, 2011.

[85] S. Luca, W. M. Yau, R. Leapman, and R. Tycko. Peptide conformation and supramolecular organization in amylin fibrils: Constraints from solid-state NMR. *Biochemistry*, 46:13505–13522, 2007.

[86] L. Wang, C. T. Middleton, S. Singh, A. S. Reddy, A. M. Woys, D. B. Strasfeld, P. Marek, D. P. Raleigh, J. J. de Pablo, M. T. Zanni, and J. L. Skinner. 2DIR Spectroscopy of Human Amylin Fibrils Reflects Stable beta-Sheet Structure. *J. Am. Chem. Soc.*, 133:16062–16071, 2011.

[87] S. D. Moran, S. M. Decatur, and M. T. Zanni. Structural and sequence analysis of the human gamma D-Crystallin amyloid fibril core using 2D IR spectroscopy, segmental C-13 labeling, and mass spectrometry. *J. Am. Chem. Soc.*, 134:18410–18416, 2012.

[88] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM - a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[89] B. R. Brooks, III Brooks, C. L., Jr. Mackerell, A. D., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30:1545–1614, 2009.

[90] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 26:1781–1802, 2005.

[91] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4:435–447, 2008.

[92] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29:845–854, 2013.

[93] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha,

D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102:3586–3616, 1998.

[94] A. D. Mackerell, M. Feig, and C. L. Brooks. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.*, 25:1400–1415, 2004.

[95] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.*, 25:1656–1676, 2004.

[96] D. I. Freedberg, R. M. Venable, A. Rossi, T. E. Bull, and R. W. Pastor. Discriminating the helical forms of peptides by NMR and molecular dynamics simulation. *J. Am. Chem. Soc.*, 126:10478–10484, 2004.

[97] M. Buck, S. Bouguet-Bonnet, R. W. Pastor, and A. D. MacKerell. Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme. *Biophys. J.*, 90:L36–L38, 2006.

[98] L. Verlet. Computer experiments on classical fluids .I. thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159:98–103, 1967.

[99] W.C. Swope, H.C Andersen, P.H. Berens, and Wilson K.R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76:637–649, 1982.

[100] J.-P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327 – 341, 1977.

[101] B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraaije. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463–1472, 1997.

[102] B. Hess. P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.*, 4:116–122, 2008.

[103] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Dinola, and J. R. Haak. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[104] S. Nosé. A molecular-dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.

[105] W. G. Hoover. Canonical dynamics - equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[106] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[107] W. L. Jorgensen and J. D. Madura. Temperature and size dependence for Monte-Carlo simulations of TIP4P water. *Mol. Phys.*, 56:1381–1392, 1985.

[108] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112:8910–8922, 2000.

[109] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78:1–20, 1999.

[110] W. P. Im, M. S. Lee, and C. L. Brooks. Generalized Born model with a simple smoothing function. *J. Comput. Chem.*, 24:1691–1702, 2003.

[111] D. Bashford and D. A. Case. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, 51:129–152, 2000.

[112] T.L.C. Jansen, A.G. Dijkstra, T.M. Watson, J.D. Hirst, and J. Knoester. Modeling the amide I bands of small peptides. *J. Chem. Phys.*, 125, 2006.

[113] S. Roy, J. Lessing, G. Meisl, Z. Ganim, A. Tokmakoff, J. Knoester, and T.L.C. Jansen. Solvent and conformation dependence of amide I vibrations in peptides and proteins containing proline. *J. Chem. Phys.*, 135, 2011.

[114] J. K. Carr, A. V. Zabuga, S. Roy, T. R. Rizzo, and J. L. Skinner. Assessment of amide I spectroscopic maps for a gas-phase peptide using IR-UV double-resonance spectroscopy and density functional theory calculations. *J. Chem. Phys.*, 140:224111, 2014.

[115] R. D. Gorbunov, D. S. Kosov, and G. Stock. Ab initio-based exciton model of amide I vibrations in peptides: definition, conformational dependence, and transferability. *J Chem Phys*, 122(22):224904, 2005.

[116] P. Hamm, M. Lim, W. F. DeGrado, and R. M. Hochstrasser. The two-dimensional IR nonlinear spectroscopy of a cyclic penta-peptide in relation to its three-dimensional structure. *Proc. Natl. Acad. Sci. U.S.A.*, 96(5):2036–2041, Mar 1999.

[117] W. Zhuang, D. Abramavicius, T. Hayashi, and S. Mukamel. Simulation protocols for coherent femtosecond vibrational spectra of peptides. *J. Phys. Chem. B*, 110:3362–3374, 2006.

[118] D. Abramavicius and S. Mukamel. Coherent third-order spectroscopic probes of molecular chirality. *J. Chem. Phys.*, 122:134305, 2005.

[119] C. Fang, J. Wang, Y.S. Kim, A.K. Charnley, W. Barber-Armstrong, A.B. Smith, S.M. Decatur, and R.M. Hochstrasser. Two-dimensional infrared spectroscopy of isotopomers of an alanine rich alpha-helix. *J. Phys. Chem. B*, 108:10415–10427, 2004.

[120] Torii H. Time-domain calculations of the polarized raman and two-dimensional infrared spectra of liquid n,n-dimethylformamide. *Chem. Phys. Lett.*, 414:417, 2005.

[121] Jansen T. l. C. and Knoester J. Nonadiabatic effects in the two-dimensional infrared spectra of peptides: Alanine dipeptide. *J. Phys. Chem. B*, 110:22910, 2006.

[122] Jansen T. l. C., Zhuang W., and Mukamel S. Stochastic liouville equation simulation of multidimensional vibrational line shapes of trialanine. *J. Chem. Phys.*, 121:10577, 2004.

[123] T. L. C. Jansen and J. Knoester. A transferable electrostatic map for solvation effects on amide I vibrations and its application to linear and two-dimensional spectroscopy. *J. Chem. Phys.*, 124:044502, 2006.

[124] R.J. Wilson. *Introduction to Graph Theory*. Addison Wesley; 4 edition, 1996.

[125] P. Fletcher. *Foundations of discrete mathematics*. PWS-KENT Pub. Co, Boston, 1991.

[126] G. Strang. *Linear algebra and its applications*. Thomson, Brooks/Cole, Belmont, CA, 2006.

[127] J. Hughes, T.W. Smith, H.W. Kosterlitz, L.A. Fothergill, B.A. Morgan, and H.R. Morris. Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature*, 258:577–579, 1975.

[128] T. M. Watson and J. D. Hirst. Vibrational analysis of capped [Leu]enkephalin. *Phys. Chem. Chem. Phys.*, 6:2580–2587, 2004.

[129] R.D. Adamson, P.M.W. Gill, and J.A Pople. Empirical density functionals. *Chem. Phys. Lett.*, 284:6–11, 1998.

[130] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[131] S. Páll and B. Hess. A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.*, 184:2641–2650, 2013.

[132] D. Qui, P. Shenkin, F. Hollinger, and W. Still. The GB/SA continuum model for solvation. a fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A.*, 101:3005–3014, 1997.

[133] N.L. Burke, J.G. Redwine, J.C. Dean, S.A. McLuckey, and T.S. Zwier. UV and IR spectroscopy of cold protonated leucine enkephalin. *Int. J. Mass Spectrom.*, 378:196 – 205, 2015.

[134] N.C. Polfer, J. Oomens, S. Suhai, and B. Paizs. Infrared spectroscopy and theoretical studies on gas-phase protonated Leu-enkephalin and its fragments: Direct experimental evidence for the mobile proton. *J. Am. Chem. Soc.*, 129:5887–5897, 2007.

[135] S. Sul, Y. Feng, U. Le, D.J. Tobias, and N-H Ge. Interactions of tyrosine in Leu-enkephalin at a membranewater interface: An ultrafast two-dimensional infrared study combined with density functional calculations and molecular dynamics simulations. *J. Phys. Chem. B*, 114:1180–1190, 2010.

[136] P. Glaziou, K. Floyd, and M. Raviglione. Global Burden and Epidemiology of Tuberculosis. *Clin. Chest Med.*, 30:621–636, 2009.

[137] J. Bernstein, W.A. Lott, B.A. Steinberg, and H.L. Yale. Chemotherapy of experimental tuberculosis .5. Isonicotinic acid hydrazide (nydrazid) and related compounds. *Am. Rev. Tuberc. Pulm.*, 65:357–364, 1952.

[138] E. Dubnau, P. Fontán, R. Manganelli, S. Soares-Appel, and I. Smith. Mycobacterium tuberculosis genes induced during infection of human macrophages. *Infect. Immun.*, 70:2787–2795, 2002.

[139] D.A. Rozwarski, G.A. Grant, D.H.R. Barton, W.R. Jacobs, and J.C. Sacchettini. Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*. *Science*, 279:98–102, 1998.

[140] Y. Zhang, B. Heym, B. Allen, D. Young, and S. Cole. The catalase peroxidase gene and isoniazid resistance of mycobacterium-tuberculosis. *Nature*, 358:591–593, 1992.

[141] K. Takayama, H. K. Schnoes, E. L. Armstrong, and R. W. Boyle. Site of inhibitory action of isoniazid in synthesis of mycolic acids in mycobacterium-tuberculosis. *J. Lipid Res.*, 16:308–317, 1975.

[142] C. Vilcheze, H. R. Morbidoni, T. R. Weisbrod, H. Iwamoto, M. Kuo, J. C. Sacchettini, and W. R. Jacobs. Inactivation of the inhA-encoded fatty acid synthase II (FASII) enoyl-acyl carrier protein reductase induces accumulation of the FASI end products and cell lysis of Mycobacterium Smegmatis. *J. Bacteriol.*, 182:4059–4067, 2000.

[143] D.J. Shaw, K. Adamczyk, P.W.J.M. Frederix, N. Simpson, K. Robb, G.M. Greetham, M. Towrie, A.W. Parker, P.A. Hoskisson, and N.T. Hunt. Multidimensional infrared spectroscopy reveals the vibrational and solvation dynamics of isoniazid. *J. Chem. Phys.*, 142:212401, 2015.

[144] A. Banerjee, E. Dubnau, A. Quemard, V. Balasubramanian, K. S. Um, T. Wilson, D. Collins, G. Delisle, and W. R. Jacobs. InhA, a gene encoding a target for isoniazid and ethionamide in mycobacterium-tuberculosis. *Science*, 263:227–230, 1994.

[145] L. Kremer, L. G. Dover, H. R. Morbidoni, C. Vilcheze, W. N. Maughan, A. Baulard, S. C. Tu, N. Honore, V. Deretic, J. C. Sacchettini, C. Locht, W. R. Jacobs, and G. S. Besra. Inhibition of InhA activity, but not KasA activity, induces formation of a KasA-containing complex in mycobacteria. *J. Biol. Chem.*, 278:20547–20554, 2003.

[146] D. A. Rozwarski, G. A. Grant, D. H. R. Barton, W. R. Jacobs, and J. C. Sacchettini. Modification of the NADH of the isoniazid target (InhA) from mycobacterium tuberculosis. *Science*, 279, 1998.

[147] A. Dessen, A. Quemard, J. S. Blanchard, W. R. Jacobs, and J. C. Sacchettini. Crystal-structure and function of the isoniazid target of mycobacterium-tuberculosis. *Science*, 267:1638–1641, 1995.

[148] L. A. Basso, R. J. Zheng, J. M. Musser, W. R. Jacobs, and J. S. Blanchard. Mechanisms of isoniazid resistance in mycobacterium tuberculosis: Enzymatic characterization of enoyl reductase mutants identified in isoniazid-resistant clinical isolates. *J. Infect. Dis.*, 178:769–775, 1998.

[149] R. Rawat, A. Whitty, and P. J. Tonge. The isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the Mycobacterium Tuberculosis enoyl reductase: Adduct affinity and drug resistance. *Proc. Natl. Acad. Sci. USA*, 100:13881–13886, 2003.

[150] J.S. Oliveira, J.H. Pereira, F. Canduri, N.C. Rodrigues, O.N de Souza, W.F. de Azevedo Jr, L.A. Basso, and D.S Santos. Crystallographic and pre-steady-state kinetics studies on binding of NADH to wild-type and isoniazid-resistant enoyl-ACP(CoA) reductase enzymes from mycobacterium tuberculosis. *J. Mol. Biol.*, 359:646–666, 2006.

[151] R.C. Hartkoorn, C. Sala, J. Neres, F. Pojer, S. Magnet, R. Mukherjee, S. Uplekar, S. Boy-Röttger, K-H. Altmann, and S.T. Cole. Towards a new tuberculosis drug: pyridomycin — nature's isoniazid. *EMBO Mol. Med.*, 4:1032–1042, 2012.

[152] C. Vilcheze, F. Wang, M. Arai, M.H. Hazbon, R. Colangeli, L. Kremer, T.R. Weisbrod, D. Alland, J.C. Sacchettini, and W.R. Jacobs Jr. Transfer of a point mutation in *Mycobacterium tuberculosis* inhA resolves the target of isoniazid. *Nat. Med.*, 12:1027–1029, 2006.

[153] R.A. Laskowski and M.B. Swindells. LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.*, 51:2778–2786, 2011.

[154] S. Jo, T. Kim, V.G. Iyer, and W. Im. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.*, 29:1859–1865, 2008.

[155] J. Lee, X. Cheng, J. Swails, M.S. Yeom, P. Eastman, J. Lemkul, S. Wei, Y. Qi, S. Jo, V. Pande, D.A Case, A.D MacKerell Jr, C.L. Brooks III, J.M Klauda, and W. Im. CHARMM-GUI input generation for NAMD,

GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM force fields. *In preparation.*, 2015.

[156] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. MacKerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31:671–690, 2010.

[157] V. Zoete, M.A. Cuendet, A. Grosdidier, and O. Michielin. Swissparam: A fast force field generation tool for small organic molecules. *J. Comput. Chem.*, 32:2359–2368, 2011.

[158] E Anderson. *LAPACK users' guide*. Society for Industrial and Applied Mathematics, Philadelphia, 1999.

[159] C. Liang and T. L. C. Jansen. An efficient n-3-scaling propagation scheme for simulating two-dimensional infrared and visible spectra. *J. Chem. Theory Comput.*, pages 1706–1713, 2012.

[160] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[161] S. H. Strogatz. Exploring complex networks. *Nature*, 410:268–276, 2001.

[162] R. Albert and A. L. Barabasi. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[163] C. R. Baiz, M. Reppert, and A. Tokmakoff. Amide I two-dimensional infrared spectroscopy: Methods for visualizing the vibrational structure of large proteins. *J. Phys. Chem. A*, 117:5955–5961, 2013.

[164] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media, 2009.

[165] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graphics*, 14:33–38, 1996.

[166] N.A. Kruh, R. Rawat, P. Ruzsicska, and P.J. Tonge. Probing mechanisms of resistance to the tuberculosis drug isoniazid: Conformational changes caused by inhibition of InhA, the enoyl reductase from *Mycobacterium Tuberculosis*. *Protein Sci.*, 16:1617–1627, 2007.

[167] H-J Li, C-T Lai, P. Pan, W. Yu, N. Liu, G.R. Bommineni, M. Garcia-Diaz, C. Simmerling, and P.J. Tonge. A structural and energetic model for the slow-onset inhibition of the *mycobacterium tuberculosis* enoyl-acp reductase inha. *ACS Chem. Biol.*, 9:986–993, 2014.

[168] R. Rawat, A. Whitty, and P.J. Tonge. The Isoniazid-NAD adduct is a slow, tight-binding inhibitor of InhA, the *Mycobacterium Tuberculosis* enoyl reductase: Adduct affinity and drug resistance. *Proc. Natl. Acad. Sci. USA*, 100:1388113886, 2003.