Computational Infrared Spectroscopy: Reproducing Kernel- and Multipolar-Based Force Field Simulations for Site-Selective Dynamics of Proteins

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Seyedeh Maryam Salehi

aus dem Iran

Basel, 2022

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel

edoc.unibas.ch

 This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von:
Prof. Dr. Markus Meuwly
Prof. Dr. Anatole von Lilienfeld
Prof. Dr. Thomas la Cour Jansen

Basel, den 19. October 2021

Prof. Dr. Marcel Mayor

Dekan

To My Father

And if the Wine you drink, the Lip you press,

End in the Nothing all Things end in—Yes-

Then fancy while Thou art, Thou art but what

Thou shalt be—Nothing—Thou shalt not be less

Omar Khayyam

Drawing, Illustration to the *"Rubaiyat of Omar Khayyam":* Trans. Edward Fitzgerald, by Edmund Joseph Sullivan, London, 1912.

Abstract

Characterizing the structural and functional dynamics of complex systems in the condensed phase requires fine spatial and/or temporal resolution which is a challenging problem, demanding vibrational probes that confer possible functional and steric variation on local properties. Vibrational time occurs on the femtosecond domain and frequencies are dependent on spatial arrangement and the characteristics of the constituent atoms. Therefore, vibrational spectroscopy has become an essential tool to study the structure and dynamics of various biological systems at the molecular level. However, achieving site-specific information of biological molecules of interest, such as proteins, is impossible for many cases or problematic to rely on the intrinsic vibrational modes. To overcome this limitation, the focus of this work is the development and application of several intrinsic backbone and side chain vibrational probes that can be easily incorporated into proteins and be used to site-specifically investigate their structural or environmental properties using reproducing kernel- and multipolar-based force field simulations.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Markus Meuwly who gave me the opportunity to do my PhD in a wonderful group, great university, and beautiful country. It was a great experience and I really appreciate your consistent support and guidance throughout my study.

Further, I would like to thank my committee members, Prof. Thomas la Cour Jansen and Prof. Anatole von Lilienfeld, who kindly accepted to co-referee my thesis, and Prof. Stefan Willitsch for chairing the defense.

I am extremely grateful to my collaborators who accompanied me through the common projects: Dr. Debasish Koner, Dr. Polydefkis Diamantis, Silvan Käser, Dr. Kai Töpfer. Prof. Ursula Röthlisberger, and Prof. Peter Hamm.

Special thanks go to all present and past group members for all kinds of support and the great times we shared. Particularly, thanks to Marco and Taylan for all the support and wonderful moments we had together. I enjoyed a lot spending time with you. At last, many thanks to Eric for proofreading part of my thesis.

I am fortunate to have a lovely family and great friends. I am deeply thankful for your endless support. You have always stood behind me and this means everything to me.

Finally, I would like to thank two amazing persons in my life, my father and my partner, whom I am speechless to thank. “Babae”, without your support and encouragement I would have not achieved what I have today. “Keyvan”, without your love and hope I would have not been as happy as I am today. I am so lucky to have you both.

Contents

Chapter 1

Introduction

Proteins are key molecular machines in virtually every biological process, performing their function by interacting with other peptides, proteins, nucleic acids, ions or other small molecules.[1] Proteins are remarkable macromolecules able to transmit signals, regulate metabolic processes, activate or prevent enzymatic reactions, transform chemical energy into mechanical and electrical forces, replicate DNA, assemble macromolecular complexes, and schedule cell death.[2] The abilities of proteins are tightly related to their structural and dynamic properties which are ultimately controlled by numerous inter- and intra-molecular interactions. Any disruption of a specific interaction leads to various diseases, and knowledge of protein-ligand mechanism facilitate drug discovery.[3,4] Therefore, achieving a quantitative and molecular-level understanding of cellular processes requires a comprehensive investigation of protein-ligand interactions.

A molecule's dynamic and steric properties depend on the chemically bonded constituent atoms which are the vital unit of all molecules. Accordingly, understanding the intrinsic nature of its constituent bonds is a prerequisite to obtaining molecular knowledge. Moreover, each bond inherently is sensitive to its local environment, and thus its features provide information about the surrounding environment. Vibrational spectroscopy is one of the most direct approaches to study molecular bonds. Furthermore, molecules are not static and they exhibit dynamic behavior, interconverting between various states with an extensive range of timescales. Thus, due to the fast time scale of vibrational spectroscopy, it is an invaluable instrument for the characterization and resolution of the fastest

interconverting states, and the frequency, intensity, lineshape, and number of vibrational absorptions of a given bond provides indispensable insight into the local structure, dynamics, and environment.

Two-dimensional infrared (2D-IR) spectroscopy provides a powerful approach to investigate the structural dynamics of numerous biomolecular systems with high special and temporal resolution. As a vibrational spectroscopy method, it directly investigates chemical bonds vibrations and how the various vibrations of a molecule together with its local environment interact with each other. With subpicosecond time resolution and observation of spectral features such as the fluctuation of fundamental vibrational frequencies of a ligand, probe molecule, or biological macromolecule, the coupling between inter- and intra-molecular degrees of freedom such as structural characteristics or hydrogen bonding networks in the condensed phases can be interrogated.

On the other hand, 2D-IR spectroscopy has more benefits compared to the standard linear infrared spectroscopies such as Fourier transform infrared (FT-IR) spectroscopy. The vibrational spectrum of 2D-IR is spread into a second dimension which provides more information on molecular couplings and the dynamics of the system. Spectral features such as frequencies, intensities, lineshapes, and the evolution of these features with time are used to gain more insights into molecular structures and environmental dynamics. Thus, 2D-IR spectroscopy provides ideal time resolution to measure structural change.

Over the past decade, infrared vibrational spectroscopy has been used with small molecules to provide detailed characterizations of molecular and electronic structure, interaction with other molecules, and solvation energy.[5] Study of small molecules is possible, owing to the limited number of bonds that make the spectral features clearly observable. However, characterization of larger biomolecules, i.e., anything larger than the smallest protein is extensively inhibited and lacks site-specific resolution due to the huge number of similar bonds that cause spectral congestion.[6] To alleviate this limitation, significant effort has been focused on the development and application of various infrared reporters[7,8] that absorb in the frequency range between ∼1800 and ∼2500. This spectral region which is

also known as the "transparent window"[6] is free of native signals, discriminating the probe absorption from a huge protein background as is shown in Figure 1.1. Such IR probes have successfully led to valuable results. For example, using localized nitrile to clarify the role of electrostatic fields in enzymatic reactions[9,10] or to elucidate the mode of drug binding,[11,12] utilizing isotope edited carbonyl to define the mechanism of protein folding and amyloid formation[13,14] or the function and structure of membrane protein[15,16]. Furthermore, there are also other molecular groups such as thiocyanate,[17] cyanamide,[18] sulfhydryl vibrations of cysteines,[19] deuterated carbons,[20] carbonyl vibrations of metal-carbonyls,[21] cyanophenylalanine (PheCN),[22] and azidohomoalanine (AHA)[23] which have been explored extensively.

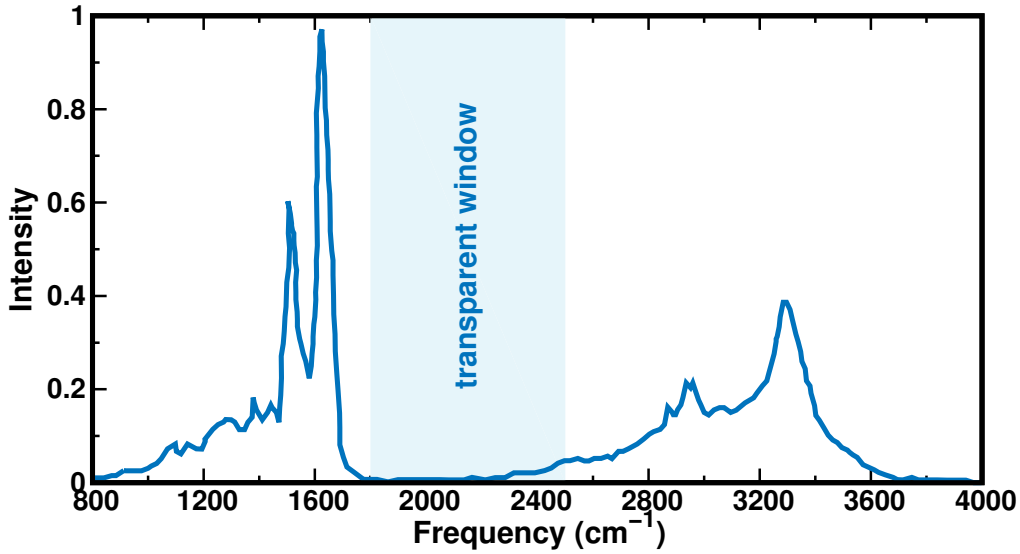


Figure 1.1: Infrared spectrum of Myoglobin[24] as a typical protein representing the area of the "transparent window".

1.1 Site-Specific Probes

Site-specific probes should have several criteria to be considered ideal for a vibrational mode.[25] First and most importantly, it must be locally sensitive to the environment and dependent on the physical property of interest, such as the electric field. In addition, the region of the IR spectrum should be located in an uncongested region and since the vibrational couplings and the spectral overlap can entangle the results, the probe should only be sensitive to its immediate environment. Usage of low sample concentration is another advantage that comes from strong transition dipole moment. Furthermore, it is important that the

probe can be easily incorporated with the minimum structural change of the reference system. Finally, for the incorporation of the probe into the original system, a method must exist to make it practically applicable. For incorporation of the probe to the transparent window, there are various methods, selection of which depends on different factors such as the size of the biological system, the specific probe, and also the site of incorporation.[20]

Site-specific IR probes of proteins can be divided into backbone- and side chain–based probe categories. Due to the chemical structure and protein backbone features, there are a limited number of backbone based IR probes. However, protein backbone vibrational modes include plenty of structural information. In particular, the C=O stretching vibrational mode of amide I band (1600-1700 cm^{-1}) has been an ideal probe for protein conformational studies.[26] In contrast, it is possible to have various side chain-based IR probes since the protein side chains are quite diverse and the incorporation of unnatural amino acids into protein is becoming more practical. Azides, nitriles, carbonyls, metal carbonyls, fluorocarbon, cysteine thiol, phosphate, and carbon deuterium are a few side chain-based IR probes. The focus of this work is on the development and applications of several of the mentioned probes, so in the following the basic spectroscopic properties of these probes are summarized.

1.1.1 Amide I mode

Amide vibrations of the polypeptide backbone have been widely used as a site-specific structural or environmental reporter in infrared studies of proteins,[26] for instance, by isotope editing method which replaces a $^{12}C{=}^{16}O$ group either with $^{13}C{=}^{16}O$ or $^{13}C{=}^{18}O$ into the protein backbone.[27,28]. Among amide vibrations, the amide I with the frequency range of 1600-1700 cm^{-1} is of particular interest, owing to its high extinction coefficient, distinct spectral signature of secondary structure, hydrogen bond network in peptides and proteins, ease of incorporation, and small perturbation to the protein structure. As illustrated in Figure 1.2, amide I vibrations include the C=O stretching and N-H bending vibrations of the backbone amide units with strong IR transition dipole moment. The coupling and physical interaction between different amide I vibrations lead to delocalized vibrations of the protein backbone. The side chain vibrations do not interact

strongly with amide I mode, except in the proline case. Thus, the amide units have similar vibrational frequencies due to the chemically identical units. As a result, local modes couple efficiently to form delocalized vibrations. In chapter three, the IR spectroscopy and dynamics of -CO labels are discussed in the wild type and mutant insulin monomer and dimer.

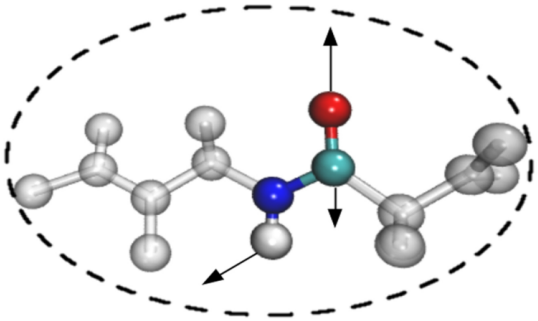


Figure 1.2: Displacement of amide I vibrations in a single amide unit of Insulin.

1.1.2 Azide Stretching Vibrations

The asymmetric stretching vibration of azide is in the range of 2000 to 2200 cm^{-1} depending on its environment and has a large extinction coefficient of $\sim$ 1000 $cm^{-1}M^{-1}$ which is ideal for low sample concentrations. Azide vibrational bands are accompanied by features arising from accidental Fermi resonance but this can be avoided using ^{15}N.[29] Moreover, several unnatural amino acids containing azide are available and as it has high usage in click chemistry, there are many methods available for incorporating azide into protein. The azide moiety can be attached to different amino acids, specifically at the terminus of aliphatic side chains. Additionally, azide has a relatively large dipole moment and this makes it an excellent IR chromophore.

Among the various unnatural amino acids, AHA with an infrared active azide side chain group has been shown to be an environment-sensitive infrared probe of local structure.[30] AHA may be the most versatile IR-active amino acid, owing to its relatively high extinction coefficient of 300-400 $M^{-1}cm^{-1}$ and easy incorporation into protein in virtually any position using known expression techniques.[23] Moreover, it is a minimally invasive probe, as evidenced, for example, by a small change of binding affinity after labeling a peptide ligand with AHA.[23] The capability of AHA label to investigate protein-ligand interactions,[31] water-specific

responses,[32] and sensitivity to local electrostatic environments[33] was previously explored and suggested that AHA is not only able to reveal large structural changes, e.g., protein folding and unfolding events, but also a very small variation of the electrostatic environment at the protein surface. Such studies confirm that AHA is a promising probe and worthwhile modification for the site-specific investigation of protein structure and dynamics. Chapter four discusses the vibrational spectroscopy of azide stretching in the gas phase and solvent and introduces an accurate computational model for the vibrational modes of azide which then will be applied in more complex systems such as protein using azidohomoalanine and/or azidoalanine ($AlaN_3$) label, see chapter five and six for more details. Figure 1.3 demonstrates a close representation of $AlaN_3$ when azide moiety is attached to one of the Alanine (Ala) residues in lysozyme protein.

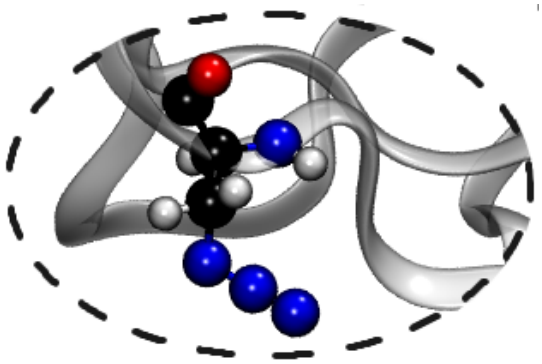


Figure 1.3: The close representation of azide attachment to one of the Ala residues in lysozyme protein which generates $AlaN_3$ label.

1.1.3 Fluorocarbon Stretching Vibrations

The fluorocarbon (-CF) stretching vibration is located at ~ 1200 cm^{-1} and has an extinction coefficient of ~ 700 cm^{-1}M^{-1}.[34] Fluorination is a common chemical modification for pharmaceuticals[35–37] owing to directionality of the interaction and high stability of CF bond which avoid metabolic transformation, for example in drugs interacting with P450.[38] Among the possible fluorinated compounds, those with phenyl rings have higher priority.[39] Moreover, fluorine is a popular NMR probe and fluorinated amino acids are often utilized to alter the pharmacokinetic and physico-chemical properties of designed peptides and proteins such as chemical reactivity, solubility, metabolic stability, enhanced membrane permeation, and biological activity compared to non-fluorinated analogs. Using halogen

atoms can also modulate the hydrophobicity around the modification site and alter the interactions with the environment.[36,40–45]

The importance of fluorine substitution in bioorganic and medicinal chemistry is highlighted by several studies such as changing drug metabolism[46] or enzyme substrate recognition,[47] mechanism-based inhibitors for various diseases and chemotherapeutic drugs,[48] and ligand binding affinity to enzyme active site.[49] As a result, a large number of drugs containing fluorine have been released for clinical research.[50] Recently, the solvatochromic and electrochromic properties of fluorine-containing aromatic compounds have been investigated which represented that the C-F vibration is sensitive to the environment.[51] Furthermore, it has been shown that the CF stretching vibration has a large Stark tuning rate which is an ideal reporter for the local electric field.[34] Fluorination has also been employed in the context of protein modifications, such as for insulin, to fine-tune thermodynamic stability and affinity to the insulin receptor.[52] These observations call for a more molecularly refined picture of the energetics and dynamics involving fluorinated model compounds. Therefore, to further explore the utility of fluorinated compounds, in chapter seven, the structural dynamics and spectroscopy of hydrated *para*-fluorophenol (F-PhOH) (see Figure 1.4) are characterized using linear infrared spectroscopy together with different computational approaches.

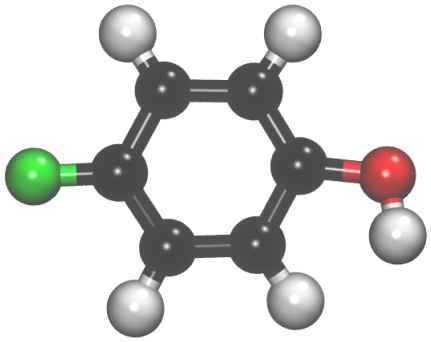


Figure 1.4: The structure of *para*-fluorophenol as an example of fluorinated compounds with phenyl rings. Fluorine atom is represented in green.

1.1.4 Nitrile/Thiocyanate Stretching Vibrations

The vibrational stretching frequency of nitrile (C≡N) is in the range of 2100-2400 cm^{-1} which is an uncongested region of protein IR spectrum.[6] Moreover, differ-

ent unnatural amino acids containing nitrile are available which makes it very common to be used in biological studies. The C≡N stretching vibration has a extinction coefficient of $\sim$ 50 $cm^{-1}M^{-1}$ in alkyl nitriles[25] while in the aromatic ones it is much larger[53] such as PheCN in water with extinction coefficient of $\sim$ 220 $cm^{-1}M^{-1}$. Therefore, PheCN has been widely used as a vibrational probe.[54,55]

Thiocyanate (SCN) is another common nitrile vibration probe owing to easy incorporation into protein.[56,57] When SCN moiety is incorporated into peptides or protein, it gives rise to absorption band owing to its relatively large extinction coefficient between 100 to 300 $M^{-1}cm^{-1}$. The IR stretching absorption of SCN is between 2140 and 2170 cm^{-1} and it is sensitive to its environment acting as a site-specific electric field probe for proteins.[58] Moreover, the lifetime of CN stretching mode in SCN label is so sensitive to the surrounding environment, owing to the insulating effect of heavy S atom which leads to domination of intra-molecular relaxation over inter-molecular vibrational relaxation in SCN .[59] Similar results were observed in a study of MeSCN in different solvents.[17] The application of SCN probe in lysozyme protein (Figure 1.5) will be further discussed in the outlook.

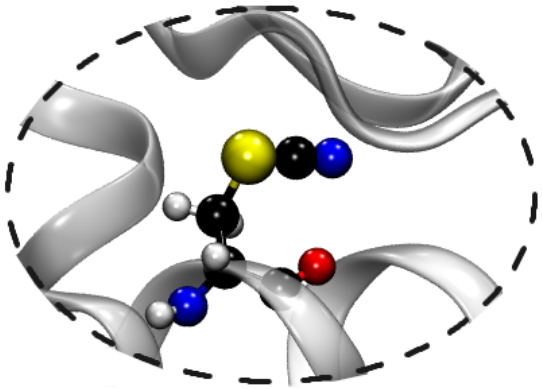


Figure 1.5: The close representation of SCN probe attached to one of the Ala residues in lysozyme protein.

Chapter 2

Theoretical Background

2.1 Potential Energy Surface

The Schrödinger equation (SE) is the fundamental concept of quantum mechanics. Wave functions is the solution to the Schrödinger equation which gives a complete description of any system. For many systems, it is adequate to solve the time-independent Schrödinger equation

$$\hat{H}|\Psi\rangle = E|\Psi\rangle \tag{2.1}$$

where $\hat{H}$ is the Hamiltonian operator for a system of nuclei and electrons and E is the energy eigenvalue. Unfortunately, even obtaining approximate solutions is computationally expensive and it is only possible for a few atoms. To overcome this problem, potential energy surfaces (PESs) are utilized to solve the SE by estimating the energy of the model system, evaluating an analytical function.

A PES is a multi-dimensional function that defines the potential energy of a system, especially a collection of atoms, in terms of nuclear coordinates. This originates from Born-Oppenheimer[60] approximation based on which the electronic and nuclear motions can be separated from each other. This approximation is based on the fact that nuclei are three orders of magnitude heavier than electrons, they move more slowly, thus, nuclei can be considered as being stationary. From this perspective, electrons always remain at the ground state energy and adjust instantaneously with nuclei movement. Therefore, the kinetic energy of nuclei is

neglected and repulsion among the nuclei is considered as constant. As a result, the electronic energy is determined by the nuclear potential which is, in turn, depends on the positions and nuclear charges and this leads to the concept of the potential energy surface.

The most common methods for solving SE are *ab initio* methods, semi-empirical methods, density functional theory, and force fields. Machine learning (ML) methods have become increasingly popular in recent years among which Kernel ridge regression (KRR) and artificial neural networks (NNs) are the promising methods that are mostly used in PES construction.[61]

2.2 *ab initio* Methods

Quantum mechanics is the essential method to describe electron distribution in detail by solving SE. When the solutions are obtained without referring to experimental data, then they are generally entitled as "*ab initio*" methods while those with an empirical correction from experimental results are known as semi-empirical methods. A brief explanation of some of the important *ab initio* methods used in the current study is mentioned in the following.

2.2.1 Hartree-Fock Theory

The Hartree-Fock (HF) approximation is a key concept in chemistry, not only for its own sake but as an initial step for more accurate methods. In HF theory, the correlation between electrons is disregarded, and instead, the average electron-electron interaction is considered. Thus, the total many-body wave functions can be defined as the product of individual electrons which is known as the Hartree product

$$\Psi_{HP}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = \chi_1(\mathbf{x}_1)\chi_2(\mathbf{x}_2) \cdots \chi_N(\mathbf{x}_N) \tag{2.2}$$

where $\chi_i(\mathbf{x}_i)$ is the spin orbital of electron i. Since electrons are fermions (spin of 1/2), the total electronic wave function must be antisymmetric if two electronic

coordinates interchange. Therefore, Eq. 2.2 does not satisfy the antisymmetry principle, and to achieve this the Slatter Determinants (SDs) can be used

$$\Psi_{SD} = |\chi_1\chi_2\dots\chi_N\rangle = \frac{1}{\sqrt{N!}}\begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \dots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \dots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \dots & \chi_N(\mathbf{x}_N) \end{vmatrix} \tag{2.3}$$

where the columns of the Slatter determinant refer to electron wave functions, orbitals, and the rows are the electron coordinates. Having selected a single-determinant as the wave function, according to variational principle, the best spin orbitals χ_i are those which minimize the electronic energy

$$E_0 = \langle\Psi_{SD}|\hat{H}|\Psi_{SD}\rangle \tag{2.4}$$

by using an iterative procedure called the self-consistent-field (SCF) method. The $\hat{H}$ is the Hamiltonian operator and the results lead to the Hartree-Fock equation

$$\hat{F}_i\chi_i = \epsilon_i\chi_i \tag{2.5}$$

where $\hat{F}_i$ is the Fock operator and ϵ_i is the energy of the i-th spin orbital χ_i. For a system of $2N$ electrons, the Fock operator has the expression

$$\hat{F}_i = \hat{h}_i + \sum_j (\hat{J}_j - \hat{K}_j) \tag{2.6}$$

and $\hat{J}_i$ and $\hat{K}_i$ describe electron-electron interaction known as Coulumb and exchange operator while $\hat{h}_i$ is the one-electron operator.[62,63]

2.2.2 Electron Correlation Methods

Molecular energies obtained by the Hartree-Fock method typically have 1% error which is very significant for describing chemical phenomena. The main drawback in the HF approximation is that the average electron-electron interaction is taken into account instead of the real one. This is an incorrect assumption since electron-electron interaction depends on their instantaneous positions which indicates the electron motions are correlated. The difference between the HF and

the lowest possible energy is known as electron correlation energy (EC). Thus, electron correlation should be considered in the wave function. To improve HF results, the wave function must contain more than one Slater determinant

$$\Psi = a_0\Psi_{HF} + \sum_{i=1} a_0\Psi \tag{2.7}$$

where a_i are the weights of the different SDs and a_0 is usually close to one which is defined by the normalization condition. Moreover, depending on how the coefficients are calculated, there are different electron correlation methods. Among the possible methods, Configuration Interaction (CI), Many-Body Perturbation Theory (MBPT), and Coupled Cluster (CC) are three main ones.[62,63]

2.2.2.1 Configuration Interaction

The CI method is analog to the HF method, using the variational principle. The wave function is represented as a linear combination of determinants together with expansion coefficients which are defined to have minimum energy.

$$\Psi_{CI} = a_0\Psi_{HF} + \sum_{S} a_S\Psi_S + \sum_{D} a_S\Psi_D + \sum_{T} a_T\Psi_T + \cdots = \sum_{i=0} a_i\Psi_i \tag{2.8}$$

Subscripts S, D, and T stand for Singly, Doubly, and Triply excited Slater determinants. The coefficients are optimized with SCF procedure and the molecular orbitals are taken from HF calculation and remain unchanged. If the basis is complete, full-CI can provide exact energy while this is computationally expensive and this calls for better approximations.[62]

In order to have a more tractable model, the number of excited determinants in Eq. 2.8 must be reduced. CI with singles (CIS) is similar to HF, while CI with doublets (CID) gives an improvement. The only CI method which can be applied for a large variety of systems is CISD which includes both singles and doubles. CIST is computationally feasible for medium-size molecules while with larger molecules it recovers less correlation energy. The multi-configuration self-consistent field (MCSCF) is another CI method in which both coefficients and molecular orbitals utilized for constructing the determinants are optimized. This

method typically converges faster than CI and to perform it cheaper, a complete active space SCF (CASSCF) approach can be used in which a limited number of electrons and orbitals contribute to the excitations. The above-mentioned CI methods have HF-type wave function while MCSCF wave function can also be selected as the reference. Given that, a CISD includes excitations of one or two electrons which defines the multireference configuration interaction (MRCI) method.[62]

2.2.2.2 Many-Body Perturbation Theory

The fundamental idea behind many-body perturbation theory is that the solution to a problem is approximately close to the already known solution. Thus, the Hamiltonian operator can be defined as

$$\hat{H} = \hat{H}_0 + \lambda \hat{H}' \tag{2.9}$$

where $\hat{H}_0$ is the reference and $\hat{H}'$ stands for perturbation. λ is a variable representing the strength of the perturbation. Perturbation methods can be used to add corrections to the solutions in quantum mechanics. Since λ varies between 0 and 1, the energy and wave function must alter continuously and can be expanded based on the perturbation parameter λ:

$$E = \lambda^0 E_0 + \lambda^1 E_1 + \lambda^2 E_2 + \cdots + \lambda^n E_n + \ldots \tag{2.10}$$

$$\Psi = \lambda^0 \Psi_0 + \lambda^1 \Psi_1 + \lambda^2 \Psi_2 + \cdots + \lambda^n \Psi_n + \ldots \tag{2.11}$$

When $\lambda = 0$, $\hat{H} = \hat{H}_0$, it refers to the unperturbed or zeroth-order wave function. Ψ_n and E_n correspond to the n-th order corrections to the wave function and energy, respectively. If $\lambda = 1$, then the n-th order energy or wave function refers to the sum of all terms up to order n. After normalizing the wave function, then the zero-, first-, second- and n-th order perturbation equations can be written as follow:

$$\lambda^0 : \ \hat{H}_0\Psi_0 = E_0\Psi_0, \tag{2.12}$$

$$\lambda^1 : \ \hat{H}_0\Psi_1 + \hat{H}'\Psi_0 = E_0\Psi_1 + E_1\Psi_0, \tag{2.13}$$

$$\lambda^2 : \ \hat{H}_0\Psi_2 + \hat{H}'\Psi_1 = E_0\Psi_2 + E_1\Psi_1 + E_2\Psi_0, \tag{2.14}$$

$$\lambda^n : \ \hat{H}_0\Psi_n + \hat{H}'\Psi_{n-1} = \sum_{i=0}^{n} E_i\Psi_{n-i} \tag{2.15}$$

Zeroth- and first-order perturbations yield the HF energy while adding a correction gives the MP2 energy ($\hat{H}_0$ is the Fock operator and $\hat{H}'$ accounts for the difference between the Fock operator and the exact Hamiltonian). Higher-order corrections give MP3, MP4, MP5, etc. energy.[62]

2.2.2.3 Coupled Cluster

In perturbation methods, all types of correction (S, D, T, ...) to a given order (2,3,4, ...) is considered while based on Coupled Cluster methods all corrections of a given type to infinite order can be included. Thus, an excitation operator can be defined

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \cdots + \hat{T}_{N_{elec}} \tag{2.16}$$

where N_{elec} is the total number of electrons and by acting $\hat{T}_i$ operator on HF reference wave function Ψ_0, all ith excited Slater determinants can be generated.

$$\hat{T}_1\Psi_0 = \sum_{i}^{occ.}\sum_{a}^{vir.} t_i^a\Psi_i^a \tag{2.17}$$

$$\hat{T}_2\Psi_0 = \sum_{i<j}^{occ.}\sum_{a<b}^{vir.} t_{ij}^{ab}\Psi_{ij}^{ab} \tag{2.18}$$

Ψ_i^a and Ψ_{ij}^{ab} are singly and doubly excited Slater determinants where i and j represent the occupied orbitals and a and b are the virtual (unoccupied) ones. t_i^a and t_{ij}^{ab} are the amplitudes which can be determined using variational principle.

Thus, with generating all possible excited determinants, the coupled cluster wave function is equivalent to full CI and can be defined as

$$\Psi_{CC} = e^{\hat{T}}\Psi_0 \tag{2.19}$$

which is impossible for all but the smallest systems. Therefore, there is a need for truncation of the cluster operator. The lowest level of approximations is coupled clustered doubles (CCD) when $\hat{T} = \hat{T}_2$. Using $\hat{T} = \hat{T}_1 + \hat{T}_2$ gives CCSD which is more complete and the only generally applicable model. The higher levels such as CCSDT and CCSDTQ are computationally demanding and consequently can only be used for small systems.[62]

2.3 Basis Sets

The objective of all *ab initio* methods is to derive information by solving the Schrödinger equation. One of the essential approximations to all *ab initio* methods is the introduction of a basis set which is an expansion to define an unknown function such as molecular orbital (MO) in terms of a set of known functions. Considering a set of known M_{basis} basis functions $\{\chi_\alpha \mid \alpha = 1, 2, \ldots, M_{basis}\}$, the unknown molecular orbital can be expanded in the linear expansion

$$\psi_i = \sum_{\alpha}^{M_{basis}} c_{\alpha i}\chi_\alpha \tag{2.20}$$

which is the matrix of one particle wave function. If the basis set χ_α is complete, then the molecular orbital is not an approximation anymore. However, for a complete basis function, an infinite number of functions must be utilized which is impossible in practice. Thus, the size of the basis set and the type of the basis function are important factors that affect the accuracy of the results. The better a single basis function, the fewer basis set is needed to achieve a given level of accuracy.[62]

In the most general sense, the basis set is a collection of basis functions including a specific set of parameters while a basis function is a specific type of mathematical function. There are two types of basis functions: Slater-Type Orbitals (STOs)

and Gaussian-Type Orbitals (GTOs). The functional form of Slater-type orbitals is defined as

$$\chi_{\zeta,n,l,m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi)r^{n-1}e^{-\zeta r} \tag{2.21}$$

where $Y_{l,m}$ are spherical harmonic functions and N is the normalization constant. Moreover, ζ represents the orbital exponent and n, l, m define the angular momentum quantum numbers $L = n + l + m$ which determines the type of orbital (e.g. $L = 0, 1, 2, \dots$ stands for $s, p, d, \dots$). On the other hand, Gaussian-type orbital in terms of polar coordinates can be written as follow:

$$\chi_{\zeta,n,l,m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi)r^{2n-2-l}e^{-\zeta r} \tag{2.22}$$

GTOs are computationally more efficient compared to Slater-type functions. However, the GTOs do not have a cusp at the nuclear position and decay rapidly at large distances in contrast with Slater type orbital which has a slow decay. Therefore, a new approach was suggested in which a linear combination of Gaussian functions (also called primitive function) is used to construct an approximate Slater type function which is known as a contracted Gaussian function. This allows having different signs for atomic orbitals in different parts of space. Moreover, since the inner shell of atoms has fewer contributions to the total energy, based on the split valence basis set, a single STO can be used for the inner shell to have a minimal basis set, a double-ζ for two, and triple-ζ for three basic functions per atomic orbital and so on. Pople basis sets such as 6-31G are among the most used of this type. This can be further improved by adding polarization effect using the so-called polarized (p) basis functions which add higher orbital angular momentum to the basis set and are represented by ($*$), e.g. 6-31G*.

Diffuse basis function is another type of function that is mostly used for the description of anions and van der Waals complexes. This function has a slow decay at large distances and also has small exponents. Basis sets with this function are known as augmented basis sets and are denoted by the '+' sign. Furthermore, correlation consistent (CC) is the dunning type basis set that accounts for optimizing the basis sets using the methods that include electron correlations. As an

example, aug-cc-pVTZ basis set indicates an augmented, correlation-consistent, polarized, valence, triple-ζ basis set.[62]

2.4 Force Field

Computer simulations are a powerful method to investigate the structure and dynamics of biological systems using basic principles of quantum mechanics. However, there are numerous problems that cannot be tackled with the quantum mechanics approach due to high computational costs. Therefore, this calls for a more simplified approach to understand the structure, dynamics, and function of biological macromolecular systems.[64] The force field (FF) or molecular mechanics (MM) method is based on a ball-and-spring model for the bonded and non-bonded interactions, considering the nuclei as balls that are held together with springs. In the FF model, only the motions of the nuclei are computed, ignoring the electronic degrees of freedom. This methodology is based on the Born–Oppenheimer approximation, allowing nuclear and electronic motions to be separated. The potential energy of a system as a function of nuclear coordinate, $U(\vec{R})$, is usually split into bonded and non-bonded contributions:

$$U(\vec{R}) = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{2.23}$$

One of the advantages of modeling PES with empirical FFs is that they are computationally efficient, owing to analytical formulas of the potential energy terms, Eq. 2.23. As a result, FFs can be applied to large systems with thousands of atoms, for which it is not possible to solve the Schrödinger equation. Moreover, FF can be used for all chemical systems with the same types of atoms. Commonly applied force fields are. e.g., CHARMM,[64] Amber,[65] and OPLS[66]. The most common general form of the potential energy used in CHARMM for macromolecular simulations is shown in Eq. 2.24. Generally, the bonded terms include

the energy function for bonds, valence angles, dihedrals, and improper dihedrals while electrostatic and van der Waals potential is defined in non-bonded terms.

$$
\begin{aligned}
U(\vec{R}) = & \sum_{bonds} k_b(b - b_0)^2 + \\
& \sum_{angles} k_\theta(\theta - \theta_0)^2 + \\
& \sum_{dihedral} k_\varphi(1 + \cos(n\varphi - \delta)) + \\
& \sum_{impropers} k_\omega(\omega - \omega_0)^2 + \\
& \sum_{non-bonded\ pairs} \left(\epsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right)
\end{aligned}
\tag{2.24}
$$

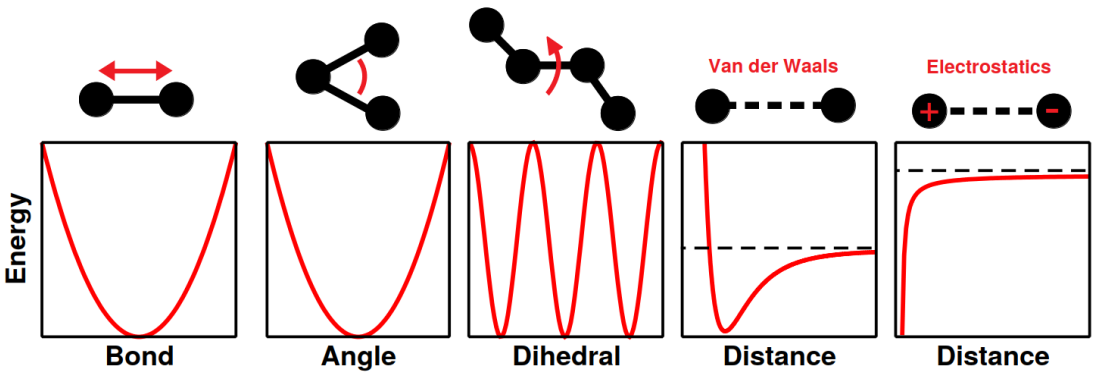


Figure 2.1: Schematic representation of the force fields together with their potentials: bond stretching, bending, rotation and non-bonded interactions.

The potential energy, $U(\vec{R})$, is a sum over individual terms representing the internal and non-bonded contributions as a function of the atomic coordinates. The parameters k_b, k_θ and k_ω are the respective force constants describing bonds, angles, and improper dihedrals, respectively. Variables with 0 subscriptions are the equilibrium values. All internal terms are considered as harmonic except for dihedral angle with a sinusoidal expression where k_φ is the force constant, n the multiplicity or periodicity, φ the dihedral angle, and δ the phase shift. For the van der Waals term, the potential energy is defined as a Lennard-Jones (LJ) potential with well depth $\epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$ and range $R_{min,ij} = (R_{min,i} - R_{min,j})/2$ which is the distance at which the LJ potential is minimum. Moreover, r_{ij} is the distance between two atoms, where i and j are the indices of the interacting atoms. In contrast to the r^{-6} dependence of the attractive part, the repulsive part of the

LJ interactions (r^{-12}) is just chosen for mathematical simplicity with no physical meaning. For the electrostatic part, q_i and q_j are the partial charges of atoms i and j involved and ϵ_0 is the effective dielectric constant.[64]

The terms mentioned in Eq. 2.24 are the standard terms of molecular mechanics force fields, however, in some force fields, two extra terms are also might be used. The first term is Urey-Bradley $\sum_{Urey-Bradley} k_{UB}(S - S_0)^2$ which is a harmonic term in the distance (S) between A and C atoms for three bonded atoms A-B-C, where k_{UB} is the force constant and S_0 is the equilibrium distance. This term turned out to be important for the in-plane deformation as well as symmetric and asymmetric bond stretching modes separation.[67] The second term relates to the correction map (CMAP) procedure to treat the conformational properties of protein backbones. The CMAP term $\sum_{residues} U_{CMAP}(\varphi, \psi)$ is a cross-term for the φ, ψ (backbone dihedral angle) values, defined by grid-based energy correction map which can be applied to any pair of dihedral angle.[68]

Note that in most FFs, bonds cannot break and form as they are defined as harmonic potential. Moreover, due to equal spacing of energy, all transitions occur at the same frequency. To alleviate this problem, Morse potential can be used to describe the bond terms which is a powerful approach to account for the anharmonicity of real bonds and a better approximation for the vibrational structure of the molecule,

$$U_{\text{Morse}} = D_e[1 - e^{-\beta(r-r_0)}]^2, \tag{2.25}$$

where D_e is the dissociation energy or depth of the potential as shown in Figure 2.2, β controls the width of the potential and r_0 is the equilibrium bond length. In the case of bond dissociation, the bonded terms of the FF for two atoms need to be redefined.

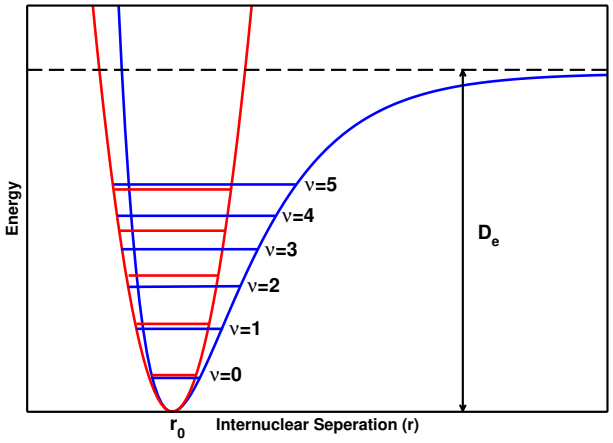


Figure 2.2: The harmonic oscillator (red) and Morse (blue) potential. The space between Morse potential levels decreases as the energy approaches the dissociation energy (D_e), in contrast to the harmonic oscillator in which the energy levels are evenly spaced.

2.4.1 Mutipolar Force Field

The electrostatic model of FF which is based on partial charges, cannot obtain certain characteristics of the electrostatic potential (ESP) such as lone pairs, hydrogen bonding, π-electron density or σ-holes (halogens).[69] Multipoles (MTPs) are usually obtained using *ab initio* Electrostatic potential (ESP). Considering $\phi(q)$ as an electron density function of coordinate q, the values of $\phi(q)$ can be discretised on a 3-dimensional (3D) grid with (r_k) coordinate points

$$\phi(q_i) \approx \Phi(r_k) \tag{2.26}$$

where $k = k^x, k^y, k^z$ refers to the coordinate where q_i is at the closest distance. Given that, the ESP at any grid point r can be approximated using MTPs up to quadrupoles[70,71]:

$$\begin{aligned}
\Phi(\mathbf{r}) = &\sum_i \sum_j Q_j^{(i)} f_j^{(i)}(\mathbf{r}) \\
&\approx \sum_i Q_{00}^{(i)} r^{-1} + Q_{10}^{(i)} r^{-2} \hat{r}_z + Q_{11c}^{(i)} r^{-2} \hat{r}_x + Q_{11s}^{(i)} r^{-2} \hat{r}_y \\
&+ Q_{20}^{(i)} r^{-3} (3\hat{r}_z^2 - 1)/2 + Q_{21c}^{(i)} r^{-3} \sqrt{3} \hat{r}_x \hat{r}_z \\
&+ Q_{21s}^{(i)} r^{-3} \sqrt{3} \hat{r}_y \hat{r}_z + Q_{22c}^{(i)} r^{-3} \sqrt{3} (\hat{r}_x^2 - \hat{r}_y^2)/2 \\
&+ Q_{22s}^{(i)} r^{-3} \sqrt{3} \hat{r}_x \hat{r}_y
\end{aligned} \tag{2.27}$$

where i repeats for all atoms and j for all MTP coefficients. $f_j^{(i)}(\mathbf{r})$ are geometrical factors together with angular- and distance-dependent terms for the MTP moment $Q_j^{(i)}$ at point $\mathbf{r}$. $r = \| \mathbf{r} \|$ is the norm of vector r and $\hat{r}_a = \mathbf{r} \cdot \hat{a}/r$ is the norm of projection of vector $\mathbf{r}$ on one of the three vectors x, y, or z. Q_{kl} is the lth MTP moment for rank k in spherical coordinates. As a next step, in order to minimize the difference between *ab initio* and MTP estimated ESPs, a linear fit is used to optimize the following function:

$$\chi^2 = \min \sum_k (\Phi_{ab\ initio}(r_k) - \Phi_{\mathrm{MTP}}(r_k)) \tag{2.28}$$

2.5 Reproducing Kernel Hilbert Space

Linear regression is one of the simplest methods to model underlying trends in data. Considering a data set $\{(y_i; x_i)\}_{i=1}^N$ of N points, the representer theorem states that any data set can always be approximated as a linear combination

$$\widetilde{f}(x) = \widetilde{y} = \sum_{i=1}^{N} \alpha_i K(\mathrm{x}, \mathrm{x}_i) \tag{2.29}$$

where $K(\mathrm{x}, \mathrm{x}_i)$ is a kernel function and α_i are the coefficients which can be defined by satisfying the linear relation

$$y_i = \sum_{i=1}^{N} \alpha_i K_{ij} \tag{2.30}$$

applying e.g., Cholesky decomposition,[72] where $K_{ij} = K(\mathrm{x}_i, \mathrm{x}_j)$ is the symmetric, positive-definite kernel matrix which can be calculated based on known values of y_i in the training set using linear equations as mentioned in Eq. 2.31.[73,74]

$$\begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1N} \\ k_{21} & k_{22} & \cdots & k_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ k_{N1} & k_{N2} & \cdots & k_{NN} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{2.31}$$

Once the coefficients α_i have been defined, for any arbitrary point x the function value can be computed using Eq. 2.29. This method is known as kernel ridge

regression (KRR) or reproducing kernel Hilbert space (RKHS).[74]

Forces are needed for the application of PES in the molecular dynamics simulations and fortunately, the derivatives of $\widetilde{f}(x)$ can also be determined analytically replacing $K(\mathrm{x}, \mathrm{x}^{'})$ in Eq. 2.29 with its corresponding derivatives. If the quality of approximation for $f(x)$ is good, then it can be expected to have also a good approximation for the derivatives of $f(x)$. In cases that KRR faces overfitting or ill-conditioned kernel matrix $\mathbf{K}$, a regularized solution can be used by adding a small positive constant λ to the diagonal of $\mathbf{K}$, such that

$$y_i = \sum_{i=1}^{N} \alpha_i (K_{ij} + \lambda \delta_{ij}) \tag{2.32}$$

is solved instead of Eq. 2.30. Here, δ_{ij} is the Kronecker delta which is $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$, when $i \neq j$. However, if the training data is noise-free and has high quality, no regularization is needed and it is best to use KRR for the reproducing of the reference values.

Molecular dynamics simulations normally need many thousands of trajectories during the dynamics. This calls for a suitable analytical representation of PES according to *ab intio* calculations which should also be computationally efficient. However, the implementation mentioned above confronts two drawbacks when the size of the training set N is large ($N \gg 10^5$). One is the inefficiency to calculate the coefficients α_i and the other occurs when evaluating the model function which needs to sum over all training samples. To resolve these drawbacks, a toolkit was introduced to construct the multi-dimensional PES using grid-formatted *ab inito* points based on reproducing kernel Hilber space.[74]

If the training set consists of different $N^{(d)}$, then total size N is

$$N = \prod_{d=1}^{D} N^{(d)} \tag{2.33}$$

and for D-dimentional kernel, it is possible to construct a kernel function as a tensor products of 1-dimentional kernels $K(\mathrm{x}, \mathrm{x}^{'})$

$$K(\mathrm{x}, \mathrm{x}_i) = \prod_{d=1}^{D} k^{(d)}(x^{(d)}, x_i^{(d)}) \tag{2.34}$$

while the coefficients α_i can be calculated for D 1-dimensional matrices $k^{(d)}$ at the cost of a matrix vector multiplication. On the other hand, for some of the 1-dimensional kernel functions, the function can be decomposed to different x and $x^{'}$ contributions

$$K(\mathrm{x}, \mathrm{x}^{'}) = \sum_{i=1}^{M_2} p_{2k} f_{2k}(x_<) f_{3k}(x_>) \tag{2.35}$$

where $M_2 = \widetilde{M}_1 + \widetilde{M}_2$ and is usually between 2 and 5. Moreover, $f_{1k} = f_{2k} = f_{3k}$ and $p_{1k} = p_{2k}$. The resulting representation of RKHS leads to energy and force evaluation at *ab intio* quality without tuning of parameters.

2.6 Neural Network

Artificial neural networks are among a specific category of ML algorithms confirmed to be general function approximators for the construction of a PES.[75] The idea behind NNs originated from networks formed by neurons in the nervous system. NNs rely on neuron layers that are connected to each other, allowing the network to learn and predict a specific property. Another approach of NNs is known as a high-dimensional neural network (HDNN) based on which a ML model can be predicted for large molecules by training smaller molecules that have similar structures, the so-called "amons".[76] Amon stands for atoms in molecules and the suffix shows its application as a building-block dictionary. Considering the fact that proteins consist of only 20 different amino acids with many bond patterns, a small number of amons is enough to support all the possible graphs. Therefore, the total energy of the system is decomposed based on the atomic contribution and one single NN can be used to predict the total energy of the system. Given that, NNs are promising alternatives to construct the PESs.

The fundamental transformation of every fully connected NN is a liner regression[77]

$$\mathbf{y} = \mathbf{W}\ \mathbf{x} + \mathbf{b} \tag{2.36}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters and $\mathbf{x}$ and $\mathbf{y}$ are the input and output vectors, respectively. A single layer can only demonstrate a linear relation. To approximate the nonlinear relationship between input and output, a combination of at least two so-called dense layers are needed together with a nonlinear activation function (σ) which gives Eq. (2.37).

$$\mathbf{y} = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2 \tag{2.37}$$

In general, two dense layers (known as *shallow* neural networks) are already capable to approximate arbitrary functions considering that the first layer contains enough neurons, and an activation function is also used. However, a deep neural network which composed of more than two layers was shown to approximate the parameters more efficiently.[77] The PhysNet architecture which is a high-dimensional of the message passing type[77] is used in chapter seven. In this model, the charges and Cartesian coordinates are utilized as input and a deep neural network (DNN) is representative of information exchange between individual atoms. The HDNNs learn in an iterative fashion to build features in a data-driven form that encodes the local chemical environment of each atom to predict the total energy, atomic forces, and molecular dipole moments. In order to capture the best parameters compared to the reference energies, forces, and dipole moments, the PhysNet parameters are optimized, minimizing a loss function[77]

$$\begin{aligned}\mathcal{L} = w_E|E - E^{\text{ref}}| + \frac{w_F}{3N}\sum_{i=1}^{N}\sum_{\alpha=1}^{3}\left|-\frac{\partial E}{\partial r_{i,\alpha}} - F_{i,\alpha}^{\text{ref}}\right| \\ + w_Q\left|\sum_{i=1}^{N} q_i - Q^{\text{ref}}\right| + \frac{w_P}{3}\sum_{\alpha=1}^{3}\left|\sum_{i=1}^{N} q_i r_{i,\alpha} - p_\alpha^{\text{ref}}\right| + \mathcal{L}_{nh}\end{aligned} \tag{2.38}$$

which depends on the reference energy (E^{ref}), force (F^{ref}), charge (Q^{ref}), and dipole moments. The weights w_E, w_F, w_Q and w_p control the relative contributions of individual errors to the total loss function. $\{\alpha\}$ defines the Cartesian coordinates and q_i stands for a partial charge of atom $\{i\}$. Moreover, $\mathcal{L}_{nh}$ is a regularization term that penalizes when the predictions of individual models decay slowly. Thus, the neural network learns to carry out a smooth decomposition using a data-driven approach.

The total energy of a molecule including long-range electrostatic and dispersion interactions for an arbitrary geometry is defined as

$$E = \sum_{i=1}^{N} E_i + k_e \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{q_i q_j}{r_{ij}} + E_{D_3} \tag{2.39}$$

where E_i is the atomic energy contributions and q_i and q_j refers to partial charges which are corrected to ensure energy conservation. N is the total number of atoms, r_{ij} is the distance between atoms i and j, k_e is the Coulomb constant, and E_{D_3} defines the dispersion correction.[78] Given that, PhysNet is capable of predicting energies and forces on various structures with chemical and conformational changes and diverse data sets.[77]

2.7 Molecular Dynamics Simulation

Molecular dynamics (MD) simulations are powerful tools to provide valuable insights into the physical basis of the structure and function of biological macromolecules and their impact has extensively developed in recent years.[79] Simulations can provide information about individual particle motions as a function of time, capturing various biomolecular processes such as conformational change, protein folding, and ligand binding. Importantly, they can also predict the atomic level responses to a perturbation such as mutation, protonation, phosphorylation, or ligand removal or attachment.[79] The simulations are potent for various reasons. First, they can provide information about individual particle motions as a function of time. Therefore, they can be utilized to address specific questions about the model system characterizations, which is very problematic with experimental techniques. Another important aspect is that conditions of simulation are accurately known and thoroughly under the control of the user so that by

changing specific contributions and comparing simulations with different setups, the effects of various molecular perturbations can be determined.[80]

MD simulation has a straightforward idea behind it. Given the positions of all atoms in a model system, the exerted force on each individual atom can be calculated using Newton's laws of motion. Therefore, the spectral position of each atom can be determined as a function of time. By integrating Newton's laws of motion, the successive configurations of motions are generated called trajectory which is a 3-dimensional movie, describing the configuration of the system at every point with respect to time.[81] According to Newton's second law:

$$\mathbf{f}_i = m_i \mathbf{a}_i \ = m_i \frac{d^2 \mathbf{r}_i}{dt^2} \tag{2.40}$$

Where f_i, m_i, r_i, a_i are the exerted force, mass, position and acceleration of atom i at time t. For a given potential energy surface (PES), $V(r^N)$, where $r^N = (r_1, r_2, r_3, ..., r_N)$ are the 3N spatial coordinates of atom i, the force on each particle is calculated as

$$\mathbf{f}_i = -\frac{\partial V}{\partial \mathbf{r}_i} \tag{2.41}$$

To access the trajectory, first Eq. 2.40 required to be integrated. There are various algorithms to integrate the equations of motion, among which the velocity Verlet[82] and Leapfrog algorithm[83] are probably the most common integrator in MD simulation which are also used in the current study. Using the positions and accelerations at time t, the new positions, $r(t + \delta t)$, is calculated at $t + \delta t$. The relationship between these three quantities and velocity at time t, $v(t)$, can be considered as follow:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 \tag{2.42}$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2}\left[\mathbf{a}(t) + \mathbf{a}(t + \delta t)\right]\delta t \tag{2.43}$$

Using equations (2.42) and (2.43), propagating coordinates and velocities, a trajectory can be recorded. One positive aspect of the velocity Verlet algorithm is the synchronized calculation of positions and velocities which is not the case for the leapfrog algorithm (half a time step apart). This means when the positions are defined, it is impossible to calculate the kinetic energy contribution to the total energy.

In leap-frog algorithm, using velocities at time $t - \frac{1}{2}\delta t$, and the acceleration at time t, velocities at time $t + \frac{1}{2}\delta t$ are first calculated which is then utilized to get positions at time $t + \delta t$ according to the following relationships:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t + \frac{1}{2}\delta t)\delta t \tag{2.44}$$

$$\mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t - \frac{1}{2}\delta t) + \mathbf{a}(t)\delta t \tag{2.45}$$

The leap-frog algorithm does not require the calculation of the differences of large numbers and explicitly includes the velocity which is the advantage over the velocity Verlet algorithm.

2.8 Quantum Mechanics/Molecular Mechanics

Quantum mechanical methods provide a more accurate description of the system compared to classical mechanics. Particularly, they are able to describe the details of bond breaking/forming or electron transfer reactions which is not the case for force fields in molecular dynamics simulations. However, the size of the system in the QM approach is limited to small systems with low complexity. Therefore, if the system is too large to calculate the electronic structure, an approximation can be used in terms of the combined quantum mechanical and molecular mechanical (QM/MM) model. In QM/MM approach, the active part where chemical reactions occur is treated with QM methods and the rest of the system is described by a classical MM force field. This approach is appealing as the computational effort can be focused on the region of chemical reaction, e.g. enzyme active site,

while taking advantage of classical force field describing the environment far from the active site. The partition in this hybrid method can be written as

$$\hat{H}_{total} = \hat{H}_{QM} + \hat{H}_{MM} + \hat{H}_{QM/MM} \tag{2.46}$$

where $\hat{H}_{QM}$ is the quantum Hamiltonian and $\hat{H}_{QM}$ is the Hamiltonian describing the classical part. Moreover, $\hat{H}_{QM/MM}$ defines the coupling between quantum mechanical particles and the classical region. Using this representation, the total energy of the mixed quantum/classical system can be calculated using the lowest eigenvalue of Eq. 2.47.

$$E_{total} = E_{QM} + E_{MM} + E_{QM/MM} \tag{2.47}$$

One of QM/MM methods is the Car-Parinello molecular dynamics (CPMD) which explicitly introduces electronic degrees of freedom in addition to the nuclear ones in contrast to Born-Oppenheimer approximation. As a result, equations of motion for nuclei and electrons are coupled and electronic minimization is only done initially. The fictitious dynamics in any subsequent MD simulation control the electrons and keep them in the electronic ground state.[84]

2.9 Normal Mode Analysis

Normal mode analysis (NMA) is one of the main simulation techniques utilized to probe dynamical features of the biological systems, especially for characterizing large-scale conformational changes. The main idea behind NMA is that normal modes with the largest fluctuations (lowest frequency modes) are the ones that are functionally relevant, which arises from comparison with experimental data. NMA is a harmonic analysis that uses the same force field as used in molecular dynamics simulation. To do a normal mode analysis one requires a set of coordinates, a force field to describe the interaction between atoms in the system, and software to carry out the necessary calculations. NMA needs three main calculations to perform in Cartesian coordinates. First, conformational potential energy should be minimized as a function of atomic coordinates. Second, the second derivative of potential energy with is so-called “Hessian” matrix should be calculated with respect to the mass-weighted atomic coordinates. The final

step is the diagonalization of the Hessian matrix to yield the normal modes which are the eigenvalues and eigenvectors. Depending on the size of the molecule, each of these three steps can be computationally demanding.[85]

For minimization, the potential energy function V can be written as a Taylor series based on mass-weighted coordinates $q_i = \sqrt{m_i}\Delta x_i$ where m_i is the mass of corresponding atom and Δx_i is the displacement of the ith coordinate from the energy minimization. Given that, the potential energy for $3N$ Cartesian coordinates is

$$V = \frac{1}{2}\sum_{i,j=1}^{3N} \frac{\partial^2 V}{\partial qi \partial qj}\Bigg|_0 q_i q_j \tag{2.48}$$

with setting the energy at minimum to zero, the first term in the expansion. Other linear terms and also the first derivative of energy (the force) are also zero at minimum. In NMA, the higher-order terms are neglected and the energy surface is approximated by a parabola which is defined by second derivatives at the minimum. The second derivatives are defined as the Hessian, H, matrix and are used to determine the eigenvalues and eigenvectors.[85]

$$Hw_j = \omega_j^2 w_j \tag{2.49}$$

w_j and ω_j^2 accounts for the jth eigenvector and eigenvalue, respectively. 3N such eigenvectors are available, defining normal mode coordinates

$$Q_j = \sum_{i=1}^{3N} w_{ij} q_i \tag{2.50}$$

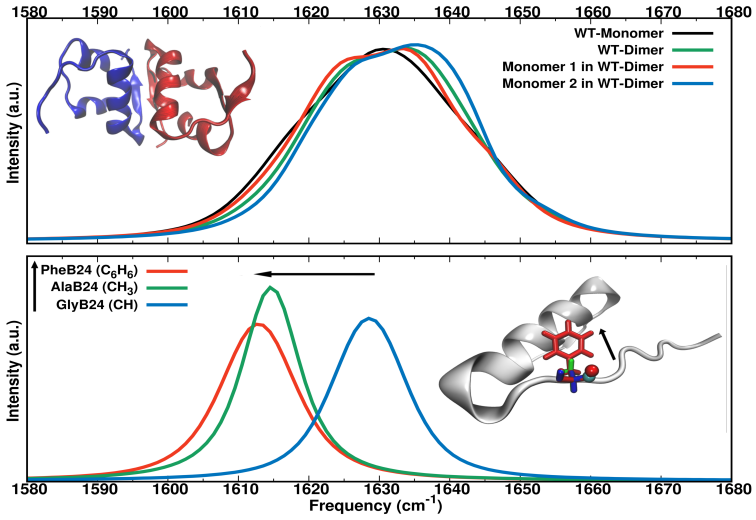
which oscillate harmonically and independently of each other

$$Q_j = A_j cos(\omega_j t + \varepsilon_j) \tag{2.51}$$

where ω_j is the angular frequency, A_j is the amplitude, and ε_j represents the phase. Note that in Eq. 2.50, $|w_j| = 1$ and the sum is done over all elements of w_j. For normal mode analysis in the present study, the "vibran" facility in CHARMM[64] is used.

Chapter 3

The Dynamics and Infrared Spectroscopy of Monomeric and Dimeric Wild Type and Mutant Insulin



The results presented in this chapter have been previously published:

J. Phys. Chem. B. 2020, 124, 11882-11894.

doi:10.1021/acs.jpcb.0c08048

***

Dr. Debasish Koner from the University of Basel contributed to this work as a second author.

3.1 abstract

The infrared spectroscopy and dynamics of -CO labels in wild type and mutant insulin monomer and dimer are characterized from molecular dynamics simulations using validated force fields. It is found that the spectroscopy of monomeric and dimeric forms in the region of the amide-I vibration differs for residues B24-B26 and D24-D26, which are involved in dimerization of the hormone. Also, the spectroscopic signatures change for mutations at position B24 from phenylalanine - which is conserved in many organisms and known to play a central role in insulin aggregation - to alanine or glycine. Using three different methods to determine the frequency trajectories - solving the nuclear Schrödinger equation on an effective 1-dimensional (1D) potential energy curve, instantaneous normal modes, and using parametrized frequency maps - lead to the same overall conclusions. The spectroscopic response of monomeric WT and mutant insulin differs from that of their respective dimers and the spectroscopy of the two monomers in the dimer is also not identical. For the WT and F24A and F24G monomers spectroscopic shifts are found to be ~ 20 cm^{-1} for residues (B24 to B26) located at the dimerization interface. Although the crystal structure of the dimer is that of a symmetric homodimer, dynamically the two monomers are not equivalent on the nanosecond time scale. Together with earlier work on the thermodynamic stability of the WT and the same mutants it is concluded that combining computational and experimental infrared spectroscopy provides a potentially powerful way to characterize the aggregation state and dimerization energy of modified insulins.

3.2 Introduction

Insulin is a small, aggregating protein with an essential role in regulating glucose uptake in cells. Physiologically, it binds to the insulin receptor (IR) in its monomeric form but thermodynamically the dimer is more stable for the wild type (WT) protein.[86–88] The storage form is that of a zinc-bound hexamer with either two or four Zn atoms.[89] Hence, to arrive at the functionally relevant monomeric stage, insulin has to cycle through at least two dissociation steps: from the hexamer to three dimers and from the dimer to the monomer.

For pharmacological applications the dimer↔monomer equilibrium is particularly relevant because for safe insulin administration this equilibrium needs to be tightly controlled. However, reliable experimental physico-chemical information about the relative stabilization of insulin monomer and dimer, which is −7.2 kcal/mol in favour of the dimer,[86] is only available for the WT and the barrier between the two states is unknown. For mutant insulins, there is no such quantitative information from experiments. On the other hand, insulin has become a paradigm for studying coupled folding and binding,[90] whether or not association proceeds along one or multiple pathways,[91,92] and for the role of water in protein association.[93–95] Most of these studies were based on atomistic molecular dynamics (MD) simulations and provided remarkable insight into functionally relevant processes for this important system.

Infrared spectroscopy has been proposed[96] and recently demonstrated[97] to provide a way to quantify protein-ligand binding strengths through observation of spectroscopic shifts. The physical foundation for this is the Stark effect which is based on the electrostatic interaction between a local reporter and the electric field generated by its environment. Using accurate multipolar force fields[98] it was possible to assign the structural substates in photodissociated CO from Myoglobin[99] whereas more standard, point charge-based force fields are not suitable for such investigations.[100]

The frequency trajectory of a local reporter can be followed in different ways. One of them uses so-called parametrized "frequency maps" which are precomputed for a given reporter from a large number of ab initio calculations.[27,101–103] Alternatively, the sampling of the configurations and computing frequencies for given snapshots can also be done using the same energy function ("scan"). In this approach, the MD simulations are carried out with the same energy function that is also used for the analysis, which is typically a multipolar representation for the electrostatics around the spectroscopic probe and an anharmonic (Morse) for the bonded terms.[104,105] On each snapshot, the local frequency is determined from either an instantaneous normal mode (INM) calculation or by solving the 1D or 3-dimensional nuclear Schrödinger equation.[106]

Here, the WT proteins and two mutants at position B24 (Phe) are considered. Phenylalanine B24 is located at the dimerization interface and invariant among insulin sequences.[107] Compared with the WT, the SerB24,[108,109] LeuB24,[110] and HisB24[111] analogues show reduced binding potency towards the receptor. On the other hand, substitutions such as GlyB24, D-AlaB24, or D-HisB24 are well tolerated as judged from their binding affinity. Nevertheless, substitutions such as GlyB24 (F24G) or AlaB24 (F24A) were found to have reduced stability of the modified insulin dimer, both from simulations and experiment,[93,112,113] and these are the variants considered in the present work.

In the present work the infrared spectrum in the amide-I stretch region is studied for wild type (WT) and two mutant insulins in their monomeric and dimeric states using accurate multipolar force fields. The IR lineshapes are calculated from frequency trajectories calculated by using a normal mode analysis, solving the Schrödinger equation from a 1-d scan along the amide-I normal mode and using previously parametrized maps. First, the methods are presented. Then, results for IR lineshapes and frequency correlation functions from scanning along the amide-I normal mode are presented and discussed and compared with the two other approaches. Finally, conclusions are drawn.

3.3 Methods

3.3.1 Molecular Dynamics Simulations

All molecular dynamics (MD) simulations were carried out using the CHARMM[64] package together with CHARMM36[67] force field including the CMAP correction[114,115] and multipoles up to quadrupole on the [CONH]-part of the backbone.[105,116] The X-ray crystal structure of the insulin dimer was solvated in a cubic box (75^3 Å^3) of TIP3P[117] water molecules, which leads to a total system size of 40054 atoms. For the monomer simulations, chains A and B were retained and also solvated in a water box (75^3 Å^3), the same box size as the dimer. In these simulations the multipolar[98,105,116,118] force field is used for the entire amide groups and all CO bonds are treated with a Morse potential $V(r) = D_e(1 - \exp(-\beta(r - r_e)))^2$. The parameters are $D_e = 141.666$ kcal/mol, $\beta = 2.112$ Å^{-1} and $r_0 = 1.231$ Å.

Hydrogen atoms were included and the structures of all systems were minimized using 2000 steps of steepest descent (SD) and 200 steps of Newton Raphson (ABNR) followed by 20 ps of equilibration MD at 300 K. A Velocity Verlet integrator[82] and Nosé-Hoover thermostat[119,120] were employed in the NVT simulations. Then production runs (1 ns or 5 ns) were carried out in the NpT ensemble, with coordinates saved every 10 fs for subsequent analysis. For the NpT simulations an Andersen and Nosé-Hoover constant pressure and temperature algorithm was used[120–122] together with a leapfrog integrator.[83] a coupling strength for the thermostat of 5 ps and a damping coefficient of 5 ps^{-1}. All bonds involving hydrogen atoms were constrained using SHAKE[123]. Nonbonded interactions were treated with a switching function[124] between 10 and 14 Å and for the electrostatic interactions, the Particle Mesh Ewald (PME) method was used with grid size spacing of 1 Å, characteristic reciprocal length $\kappa = 0.32$ $Å^{-1}$, and interpolation order 4.[125] Figure 3.1A shows the insulin dimer highlighting some of the CO labels studied in the current work with particular attention to the -CO labels at the protein-protein interface (B24-B26) and (D24-D26).

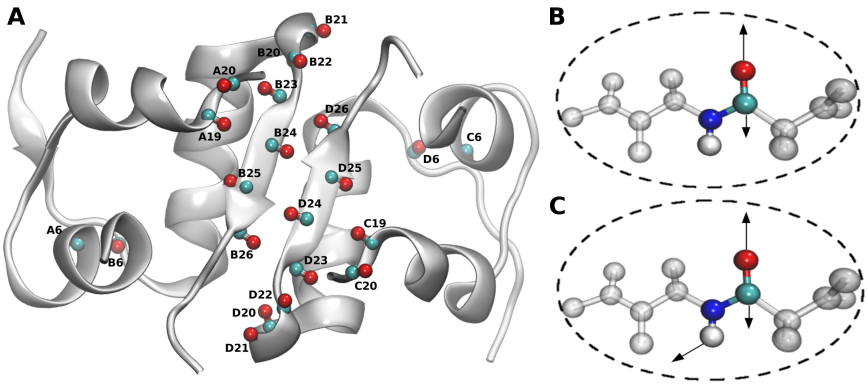


Figure 3.1: Panel A: Structure of wild type insulin dimer with the -CO labels that are specifically probed in the present work. The dimerization interface involves residues B24-B26 and D24-D26. Panels B and C show the displacement vectors for the two scan approaches considered to construct 1D potentials along the CO and CONH directions, respectively.

3.3.2 Frequencies from Solving the 1D Schrödinger Equation: Scan

Anharmonic transition frequencies can be determined from calculating the 1-d potential energy along the CO or amide-I normal mode (from a normal mode analysis on N-methyl acetamide (NMA) in the gas phase) and solving the nuclear

Schrödinger equation (SE) for each snapshot using a discrete variable representation (DVR) approach[126]. It was shown previously for NMA[127] that frequency trajectories obtained from solving the SE on the 1-d PES scanned along either the CONH (amide-I) or the CO mode (see Figure 3.1B and C) result in similar decay times with frequencies shifted by some ~ 15 cm^{-1}. Here, scans were performed for each snapshot for 61 points along the CO normal mode vector around the minimum energy structure using the same energy function as that used for the MD simulations, i.e. a multipolar representation of the electrostatics and an anharmonic Morse potential for the CO-bond. An RKHS representation of the 1-d PES is then constructed from these energies and the SE is solved on a grid (-0.53 Å $< r < 0.53$ Å) using a reduced mass of 1 amu.[127] For direct comparison, scans along the amide-I mode were also carried out for selected residues.

3.3.3 Instantaneous Normal Mode

The instantaneous (harmonic) frequencies for each snapshot of the trajectory from the NPT simulation were calculated for the same snapshots for which the scan along the CO normal mode was carried out, see above. Such instantaneous normal modes (INM) are determined by minimizing CO or [CONH] while keeping the environment (protein plus solvent) fixed. Next, normal modes were calculated from the "vibran" facility in CHARMM.

3.3.4 The Amide I Frequency Maps

The frequency map used in the present work is that parametrized by Tokmakoff and coworkers.[103] It requires MD simulations to be run with fixed CO bond length and is based on the expression

$$\omega_i = \omega_0 + aE_{C_i} + E_{N_i} \tag{3.1}$$

where ω_i is the instantaneous frequency for the ith vibrational label, E_{C_i} is the electric field on the C atom in the ith label along the C=O bond direction, and E_{N_i} is that on the N atom. Parameters ω_0, a, and b were fitted such that they optimally reproduce the experimental IR absorption spectra of NMAD. The optimized backbone map is[103]

$$\omega_i = 1677.9 + 2557.8E_{\mathrm{C_i}} - 1099.5E_{N_i} \tag{3.2}$$

In this equation, ω_i is in cm^{-1} and E_{C_i} and E_{N_i} are in atomic units. As a separate evaluation, a different map[27] is also used in which the frequency shift due to the dihedral angles (ϕ, ψ) between neighboring peptide units are included. Here the map parametrization is

$$\omega_i = 1684 + 7729E_{C_i} + 3576E_{N_i} \tag{3.3}$$

and the local frequency is

$$\omega_i^b = \omega_i + \Delta\omega_N(\phi_{i-1}, \psi_{i-1}) + \Delta\omega_C(\phi_{i+1}, \psi_{i+1}) \tag{3.4}$$

Based on the (ϕ, ψ) angles for ith chromophore, $\Delta\omega_N$ and $\Delta\omega_C$ are the contributions from $(i-1)$th and $(i+1)$th residues.

3.3.5 Frequency Fluctuation Correlation Function and Lineshape

From the harmonic or anharmonic frequency trajectory $\omega_i(t)$ or $\nu_i(t)$ for label i its frequency fluctuation correlation function, $\langle\delta\omega(0)\delta\omega(t)\rangle$ is computed. Here, $\delta\omega(t) = \omega(t) - \langle\omega(t)\rangle$ and $\langle\omega(t)\rangle$ is the ensemble average of the transition frequency. From the FFCF the line shape function

$$g(t) = \int_0^t \int_0^{\tau'} \langle\delta\omega(\tau'')\delta\omega(0)\rangle d\tau'' d\tau'. \tag{3.5}$$

is determined within the cumulant approximation. To compute $g(t)$, the FFCF is numerically integrated using the trapezoidal rule and the 1D-IR spectrum is calculated according to[128]

$$I(\omega) = 2\Re \int_0^\infty e^{i(\omega - \langle\omega\rangle)t} e^{-g(t)} e^{-\frac{t\alpha}{2T_1}} dt \tag{3.6}$$

where $\langle\omega\rangle$ is the average transition frequency obtained from the distribution, $T_1 = 0.45$ ps is the vibrational relaxation time and $\alpha = 0.5$ is a phenomenological factor to account for lifetime broadening.[128]

For extracting time information from the FFCF, $\langle \delta\omega(t)\delta\omega(0)\rangle$ is fitted to an empirical expression[129]

$$\langle \delta\omega(t)\delta\omega(0)\rangle = a_1 \cos(\gamma t) e^{-t/\tau_1} + \sum_{i=2}^{n} a_i e^{-t/\tau_i} + \Delta_0 \tag{3.7}$$

where a_i are amplitudes, τ_i are decay times and Δ_0 is an offset for long correlation times. The cos −term allows to capture a short-time recurrence (anticorrelation) that may or may not be present in the correlation function. This minimum at very short time ($t \sim 0.1$ ps) is known from previous simulations[28] and can be related to the strength of the interaction between solute and solvent[104–106,129] or between the spectroscopic probe and its environment (as in the present case). The decay times τ_i of the frequency fluctuation correlation function reflect the characteristic time-scale of the solvent fluctuations to which the solute degrees of freedom are coupled. In most cases the FFCFs were fitted to an expression containing two decay times using an automated curve fitting tool from the SciPy library.[130] Only if the quality of the resulting fit was evidently insufficient, a third decay time was included.

3.4 Results

The results section is structured as follows. First, a brief account is given of representative structures along the trajectories for the different simulation conditions used. Next, the amide-I spectroscopy for the WT monomer and dimer using the "scan" approach is given. This is followed by the spectroscopy for the mutant monomer and dimer compared with the WT systems. Then, a comparative discussion of the results for WT and mutant monomer and dimer is given for the three methods to determine the frequency trajectories ("scan", "INM" and "map") and finally, the FFCFs from the "scan" and "INM" frequency trajectories are discussed.

3.4.1 Structural Characterization

The root mean squared deviation between the reference X-ray structure and those of the monomer and dimer structure of the WT protein in solution is reported in Figure 3.2. Typically, the RMSD is around 1.5 Å which is indicative of a

stable simulation on the nanosecond time scale. Such RMSD values have also been reported from simulations in smaller water boxes.[88,131]

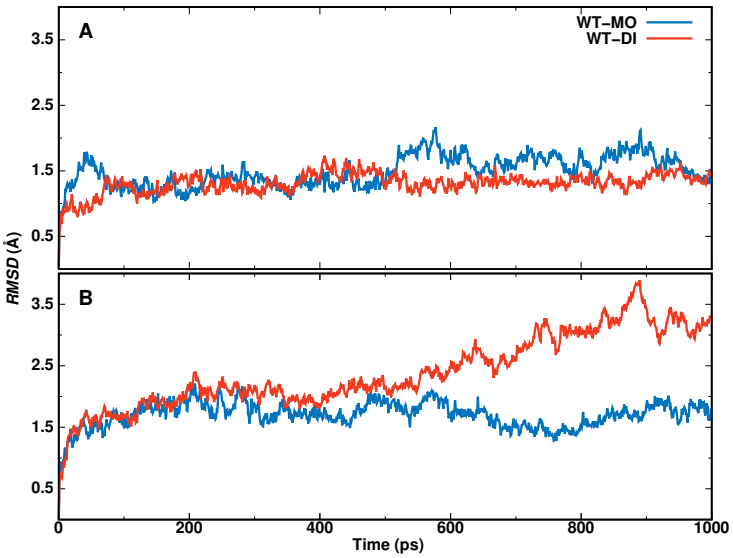


Figure 3.2: The structural RMSD between the reference X-ray structure and the Wild type monomer and dimer insulin for A) flexible and B) constrained CO.

With constrained CO (as is required for using the frequency maps) the structure of the monomer is equally well maintained whereas for the dimer it starts to deviate from the reference structure by $\sim$ 3 Å after 0.8 ns. This is indicative of structural changes which involve separation of the terminal of chain B (PheB1 and AlaB30) from each other. A similar but less pronounced effect was also observed for chain D between PheD1 and AlaD30.

3.4.2 Amide-I Spectroscopy Using Scan for WT and Mutant Monomer and Dimer

To set the stage, the Amide-I spectroscopy for the WT monomer and dimer is discussed from frequency trajectories obtained by scanning along the CO normal mode for each snapshot. Figure 3.3A reports the lineshapes for all CO-labels for the WT monomer. Lineshapes for chain A are solid lines and those for chain B are dashed. The overall lineshape for the monomer (black solid line) is centered at 1630.5 cm^{-1} and has a full width at half maximum of $\sim$ 30 cm^{-1}, compared with a center frequency of $\sim$ 1650 cm^{-1} and a FWHM of $\sim$ 30 cm^{-1} from experiments.[132,133] When comparing the position of the frequency maximum it should be noted that the present parametrization is for NMA and slight readjustments of

the Morse parameters could be made to yield quantitative agreement. However, for the present purpose such a step was deemed unnecessary.

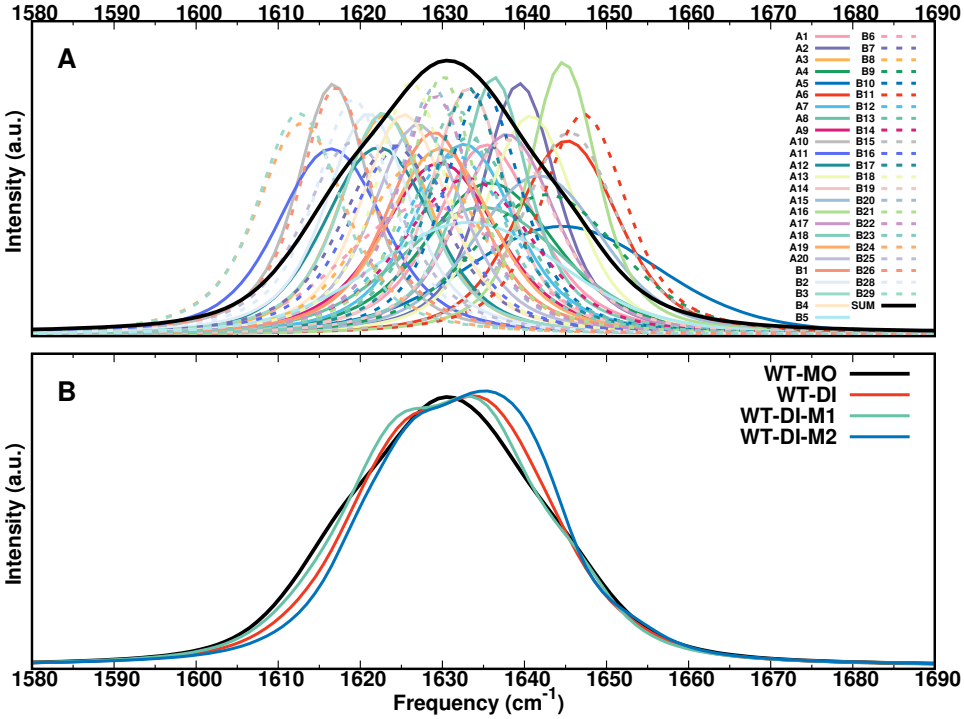


Figure 3.3: Panel A: 1D-IR spectra for all residues in WT monomer based on "scan" for the frequency calculation. The labels for the individual line shapes are given in the panel and the overall sum is the solid black line. Panel B: The total lineshape for all CO probes of the monomer (black) compared with that of M1 (green) and M2 (blue) within the dimer and with the dimer itself (red). All lineshapes are scaled to the same maximum intensity. The line shapes are determined from 1 ns simulations and the snapshots analyzed are separated by 10 fs.

On the other hand, scanning the 1D potential along the amide-I normal mode shifts the frequencies by about 30 cm^{-1} to the blue (see Figure 3.4A). The correlation between scanning along the CO and amide-I normal modes is high, as Figure 3.4C shows. In addition, the full 1D infrared spectrum was also calculated from scanning along the amide-I normal mode (Figure 3.4C) and confirms the overall shift to the blue by 25 cm^{-1} while maintaining the shape and width of the total lineshape from scanning along the CO normal mode.

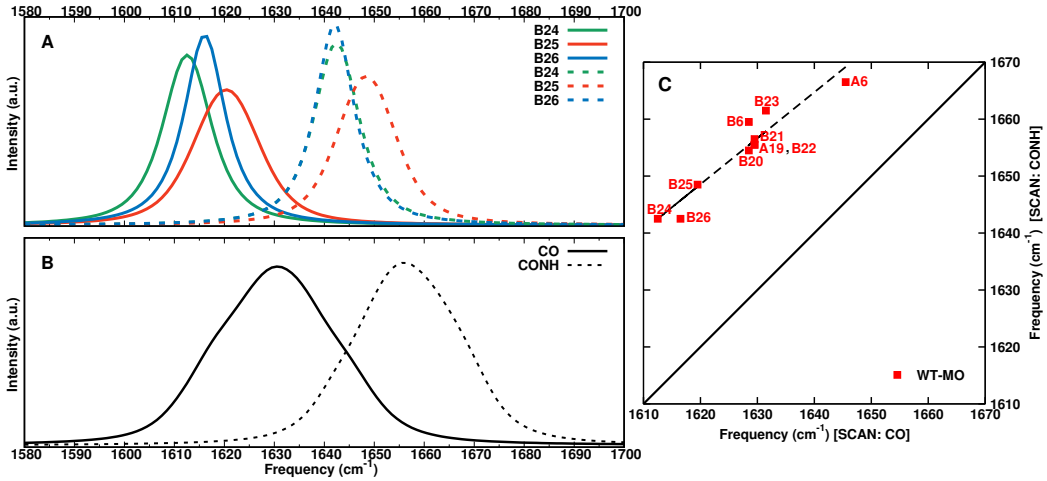


Figure 3.4: Comparison for scanning along the CO (solid line) and CONH (amide-I, dashed line) normal modes for "scan" for the insulin monomer. Panel A: 1D-IR spectra for residues (B24-B26), panel B: the sum frequency of all the residues and panel C: Comparison of the maximum frequency of the 1D-IR spectra for the selected residues (A6, A19, B6, B20-B26). The black dashed line shows the linear regression with regression coefficient (slope) of 0.81 and correlation coefficient of 0.95. The analysis is done for 1 ns simulation and the snapshots analyzed are separated by 10 fs. The frequency maxima from scanning along the [CONH] INM are shifted to the blue, in accord with the experimental observations.[132,133]

Most notably, the center frequencies for each of the labels cover a range from 1612.5 cm^{-1} (residues B24, B29) to 1647.5 cm^{-1} (residue B11) although the bonded potential (Morse) for the CO stretch is the same for all 51 labels. Hence, the multipolar charge distribution used for the electrostatics and its interaction with the environment leads to the displacements of the center frequencies. The linewidths also vary for the -CO probes at the different locations along the polypeptide chain and cover a range from 10 cm^{-1} (Residues A10, A16, A18, B18, B21) to 28 cm $^{-1}$ (Residue A5).

Selected lineshapes for the monomer and each of the two monomers within insulin dimer from scanning along the CO normal mode are reported in Figure 3.5. For the dimer it is noted that some probes at symmetry related positions within the dimer structure typically have their maxima at different frequencies. In other words, structurally related -CO probes sample different environments in the hydrated system at room temperature. The overall lineshapes of M1 and M2 are directly compared with that of the isolated monomer and the dimer in Figure 3.3B. The lineshape of M1 and M2 differ which confirms the asymmetry

noted earlier from X-ray experiments.[89,134] Also, the spectroscopy of the isolated monomer differs from that of M1 and M2 within the dimer. Notably, the -CO groups involved in the hydrogen bonding motif of the insulin dimer (B24 to B26 and D26 to D24) display frequency maxima that differ by ~ 10 cm^{-1}. Other -CO reporters, such as B20 and D20, have their maxima only ~ 5 cm^{-1} apart.

It is also observed that the absolute frequency maximum of the same reporter in the monomer and in the dimer can differ. For example, while the maximum frequency of -CO at position B24 in the monomer is at 1612.5 cm^{-1} the maxima for B24 and D24 in the dimer are at 1625.5 cm^{-1} and 1620.5 cm^{-1}. Hence, in addition to a splitting in the dimer spectrum also an overall shift of the frequencies compared with the monomer is found. Again, these effects are largest for the dimerization motif and for residues A/C6, see Figure 3.5C.

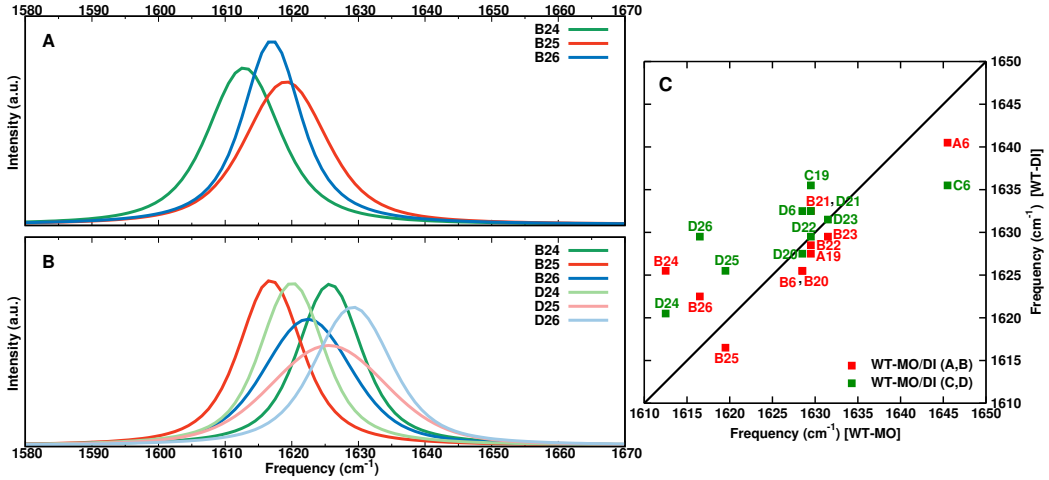


Figure 3.5: 1D-IR spectra for WT monomer (panel A) and dimer (panel B) for residues at the dimerization interface (B24-B26) and (B24-B26, D24-D26), respectively, based on "scan" for frequency calculation. Panel C compares the maximum frequency of the 1D-IR spectra for the selected residues (A6, A19, B6, B20-B26, C6, C19, D6, D20-D26) between WT monomer and dimer.

The close agreement of the computed overall spectrum with the experimentally measured one (see above) and the fact that the same computational model was successful in describing the spectroscopy and dynamics of hydrated NMA[105,135] provides a meaningful validation of the present approach.

Amide-I Spectroscopy of Wild Type and Mutant Monomers: Mutation at position B24 considerably influences the dimerization behaviour of the hormone.[135] Hence, the dynamics of the hydrated F24A and F24G monomers was first considered. The infrared lineshapes for residues along the dimerization interface and the same selected -CO probes for the WT monomer are reported in Figure 3.6C. For the two mutant monomers (Figure 3.6A for F24A and Figure 3.6B for F24G) the frequency maximum for -CO at position B24 is shifted from 1612.5 cm^{-1} (WT) to 1614.5 cm^{-1} (F24A) and 1628.5 cm^{-1} (F24G), respectively. The amide-I band maxima at positions B25 and B26 show differences for the the F24A mutant but not for F24G and for position A19 the frequency maxima shift to the blue (7 cm^{-1}) for F24A and to the red (6 cm^{-1}) for F24G compared to WT. For all other -CO labels in the monomer the differences between F24A and F24G are less than 14 cm^{-1}. The most pronounced differences in the maximum absorbances occur around the mutation site whereas away from it they are minor, except for -CO at position A19. Interestingly, residue TyrA19 is structurally close to PheB24 (see Figure 3.1A) which explains the dynamical coupling between the two sites that leads to a shift of $\sim \pm 7$ cm^{-1} and is also consistent with recent work on the stability of B24-mutated insulin.[93]

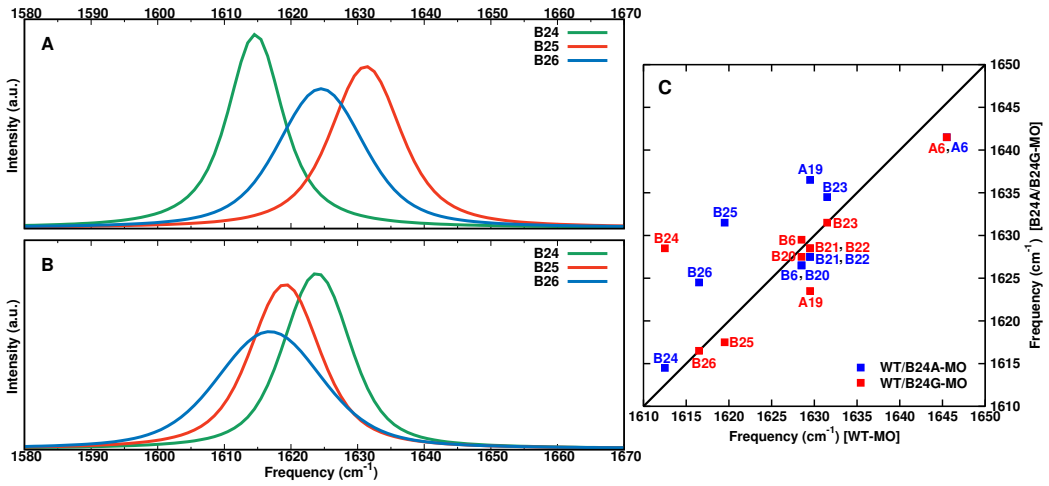


Figure 3.6: 1D-IR spectra for monomeric mutants at position B24. Panels A and B report spectra for F24A (panel A) and F24G (panel B) for residues (B24-B26) at the dimerization interface, based on "scan" for frequency calculations. Panel C compares the maximum frequency of the 1D-IR spectra for selected residues (A6, A19, B6, B20-B26) between monomeric WT and mutants F24A and F24G.

Amide-I Spectroscopy of Wild Type and Mutant Dimers: The peak frequencies for residues at the dimerization interface for the WT and the F24A mutant are

reported in Figures 3.7A and B and directly compared for a larger number of residues, see Figure 3.7C. As for the monomer, there are specific differences such as for TyrA19, PheB25, and PheD25 which shift by up to 15 cm^{-1} between the two systems. For other residues the differences are considerably smaller. For the F24G mutant differences persist, but are in general smaller, see Figure 3.8. What is found from simulations for both mutants is that residues are not necessarily symetrically affected, in particular for those along the dimerization interface. Also, depending on the modification at position B24 the effects differ and may allow to distinguish between the different insulin variants.

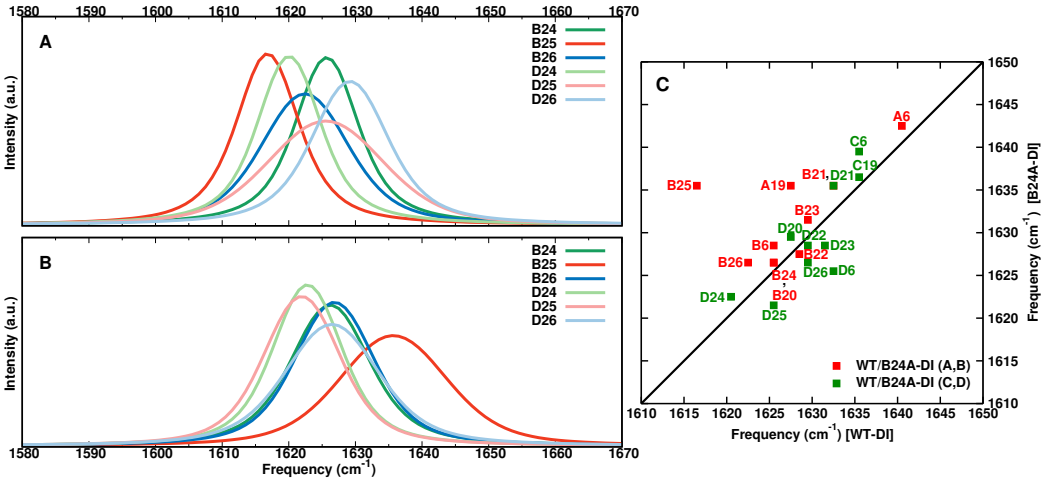


Figure 3.7: 1D-IR spectra for WT (panel A) and the F24A (panel B) dimer for residues at the dimerization interface (B24-B26, D24-D26), based on "scan" for frequency calculation. Panel C compares the maximum frequency of the 1D-IR spectra for selected residues (A6, A19, B6, B20-B26, C6, C19, D6, D20-D26) between the WT and F24A mutant dimer.

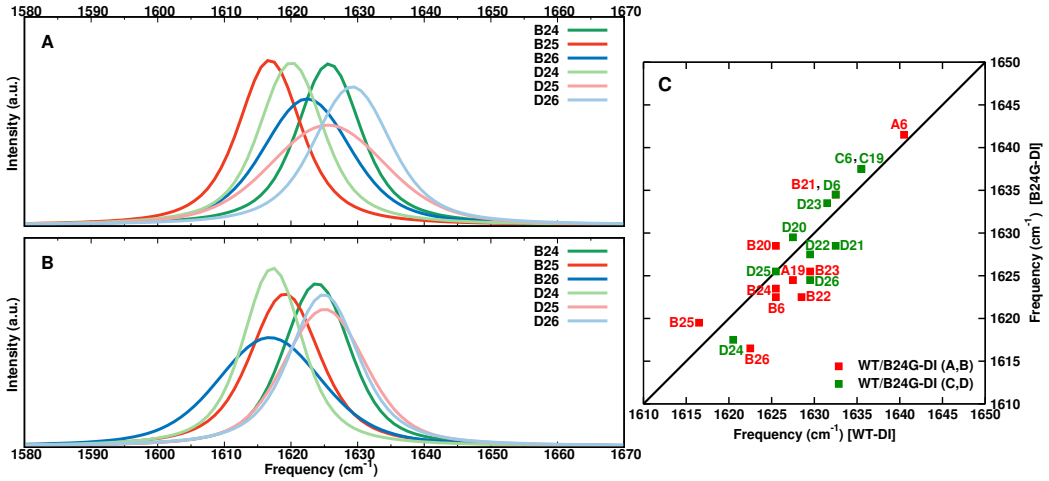


Figure 3.8: 1D-IR spectra for WT (panel A) and the F24G (panel B) dimer for residues at the dimerization interface (B24-B26, D24-D26), based on "scan" for frequency calculation. Panel C compares the maximum frequency of the 1D-IR spectra for selected residues (A6, A19, B6, B20-B26, C6, C19, D6, D20-D26) between the WT and F24G mutant dimer.

3.4.3 Comparison of Amide-I Spectroscopy from Scan, Normal Mode and Map Analyses

The three approaches to determine frequency trajectories considered here ("scan", "INM", and "map") differ considerably in terms of computational expense and the formal approximations in applying them. Scanning along the CO or amide-I normal mode for every snapshot is computationally expensive as it requires for every snapshot to carry out a 1D scan of the PES, representing it as a RKHS, and solving the nuclear Schrödinger equation. As this needs to be done for $\sim 10^5$ snapshots per nanosecond, such an approach does not scale arbitrarily to larger systems and long time scales (μs or longer). Compared to "scan", determining instantaneous normal modes is computationally less demanding and the "map" approach is also computationally efficient. In the following, the lineshapes from the frequency trajectory for the WT monomer using the three methods are compared.

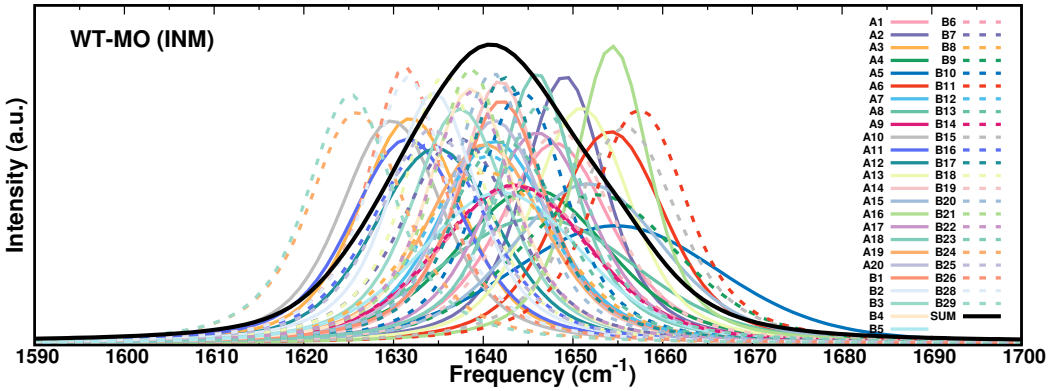


Figure 3.9: 1D-IR spectra for all residues for the WT monomer from INM for the frequency calculations. The black line shows the superposition of all CO spectra compared with other single CO spectrum.

Figure 3.9 reports the 1D lineshapes for all residues of the WT monomer from INM. As for "scan" the maxima of the individual line shapes cover a range between 1625.5 cm^{-1} and 1657.5 cm^{-1} and the average spectra over all individual lineshapes is centered at 1640.5 cm^{-1} with a FWHM of 26 cm^{-1}, compared with 1630.5 cm^{-1} and a FWHM of $\sim$ 30 cm^{-1} from "scan", see Figure 3.3. A direct comparison of the frequency maxima for the WT monomer from "scan" and INM is reported in Figure 3.10A.

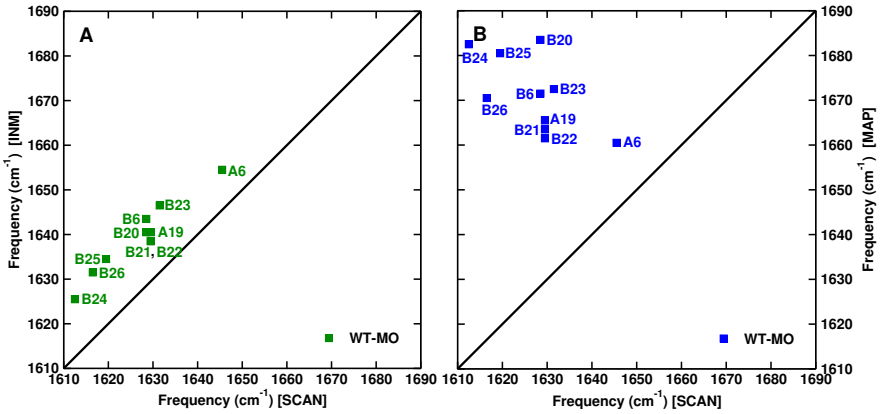


Figure 3.10: Comparison of the maximum frequency of the 1D-IR spectra between "scan" and "INM" (panel A) and "scan" and "map" (panel B) for the selected residues (A6, A19, B6, B20-B26, C6, C19, D6, D20-D26) for WT monomer. The CO probes are flexible in the simulations analyzed with "scan" and "INM" and constrained for the one using "map".

The individual and total lineshapes from using the "map" frequencies are reported in Figure 3.11. Again, the individual frequency maxima span a range of $\sim$ 50

cm^{-1} and the FWHM differ for the residues. Contrary to the overall line shape for the monomer from "scan" and "INM", using the frequency map leads to an infrared spectrum with two peaks. This shape is not consistent with the experimentally observed IR spectrum.[132,133] Also, the frequency maxima are somewhat displaced to higher frequencies and do not correlate particularly well with the frequency maxima from "scan" (see Figure 3.10B). One possibility for these differences may be the fact that for using "map" simulations with constrained -CO are required. Also, the map used in the present work was parametrized with respect to experiments and using a point charge-based force field whereas the simulations in the present work used multipoles.

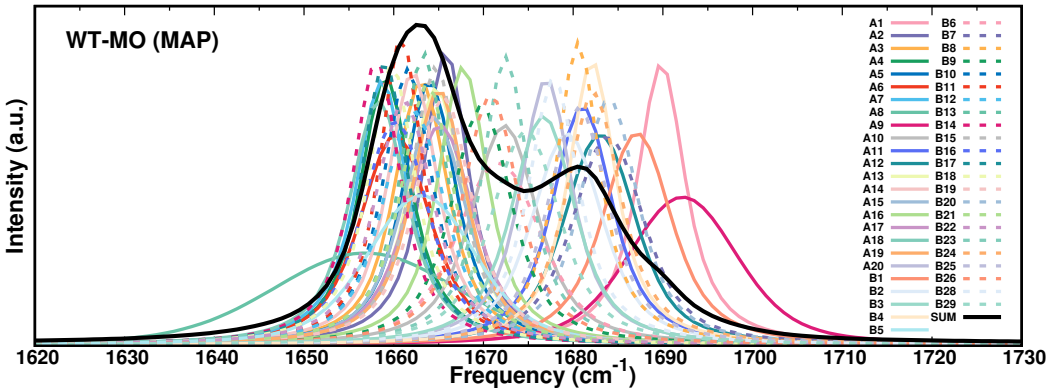


Figure 3.11: 1D-IR spectra for all residues in WT monomer based on "map" for the frequency calculation. The labels for the individual line shapes are given in the panel and the overall sum is the solid black line. The line shapes are determined from 1 ns simulations and the snapshots analyzed are separated by 10 fs.

Next, the lineshapes for the residues involved in the dimerization interface and the selection of other residues already considered until now are analyzed for WT monomer and dimer for INM and "map", see Figures 3.12, 3.13. When using INM it is again found that for the residues at the dimerization interface the location of the frequency maxima in the two monomers differ and also change compared with the isolated monomer (see Figure 3.12C). These effects are not only observed for residues at the interface but also away from it. Splitting for B/D24, B/D25, and B/D26 are comparable or larger than with "scan" and blue/red shifts are consistent for the two methods.

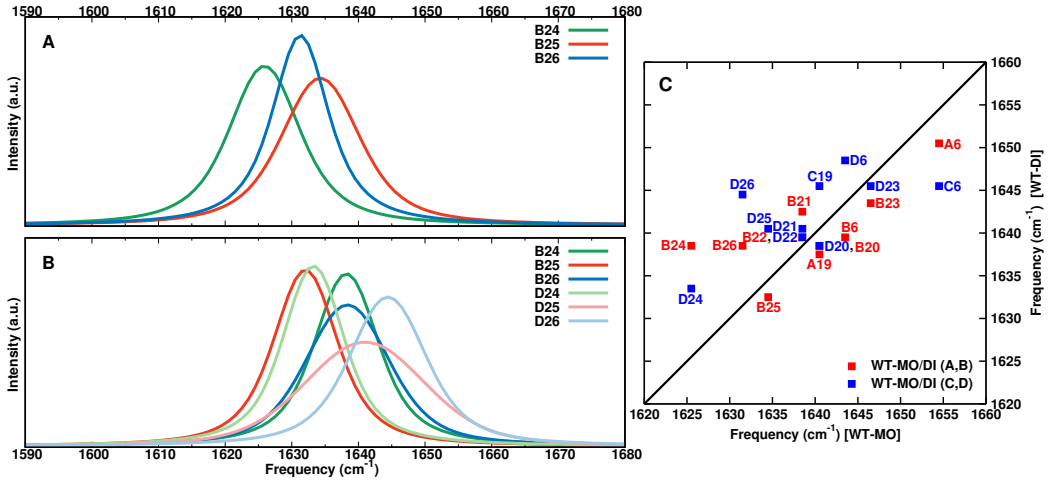


Figure 3.12: 1D-IR spectra from INM for residues (B24-B26) and (B24-B26, D24-D26) at the dimerization interface for WT monomer (panel A) and WT dimer (panel B). Panel C compares the maximum frequency of the 1D-IR spectra for the residues (A6, A19, B6, B20-B26, C6, C19, D6, D20-D26) between WT monomer and dimer.

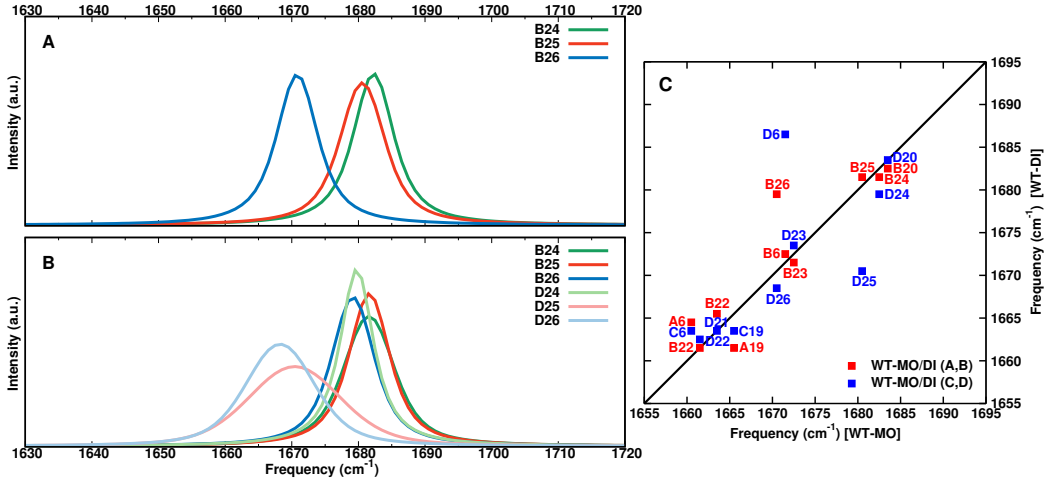


Figure 3.13: 1D-IR spectra from "map" for residues (B24-B26) and (B24-B26, D24-D26) at the dimerization interface for WT monomer (panel A) and WT dimer (panel B). Panel C compares the maximum frequency of the 1D-IR spectra for the residues (A6, A19, B6, B20-B26, C6, C19, D6, D20-D26) between WT monomer and dimer. The CO bond length is constrained in the MD simulations.

For the analysis using “map” in Figure 3.13 it is important to note that they do not use the same structures for analysis as for “scan” and INM because the -CO bond lengths were constrained. As for the other two methods the frequency maxima for B24 to B26 do not coincide for the monomer (Figure 3.13A) and the -CO labels in the two monomers have their maxima at different frequencies in the dimer (Figure 3.13B). However, the actual frequency maxima between the three methods differ. For a comparison of the maximum frequencies for the

Residue	Scan	INM	Map
B24	1612.5	1625.5	1682.5
B25	1619.5	1634.5	1680.5
B26	1616.5	1631.5	1670.5

Table 3.1: Position of the frequency maxima of the 1D-IR spectra for WT monomer using the three different approaches ("scan", "INM", and "map"). For "scan" and INM the CO probes are flexible while for "map" the structures were those from a simulation with constrained CO bond length.

three methods for B24 to B26 and D24 to D26 for direct numerical comparison, see Table 3.1. Figure 3.14 reports a comparison of the map used here and an alternative parametrization.[27] Consistent with earlier work that compared the performance of different maps,[136] it is found that the two correlate quite well (within a few cm^{-1}) except for residue B20 for which they differ by ~ 25 cm^{-1}. It is noteworthy that for both, scanning along the [CONH] normal mode (Figure 3.4) and for using "map" (Figure 3.10) compared with scanning along the CO mode, the frequency maxima are shifted towards the blue, in accord with experiment (frequency maximum ~ 1650 cm^{-1}).[132,133]

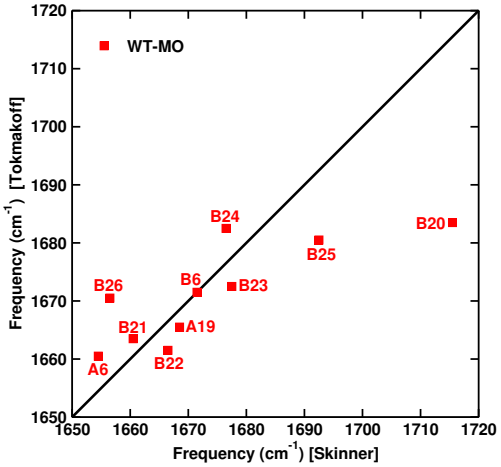


Figure 3.14: Comparison between maximum frequency of 1D-IR spectra for residues (A6, A19, B6, B20-B26, C6, C19, D6, D20-D26) based on two different maps[27,103] for WT monomer. Snapshots from the same trajectory, run with constrained CO, were analyzed.

Using "map" the labels at B/D25 and B/D26 show splittings comparable to those from "scan" and INM whereas for B/D24 the splitting is only 1 to 3 cm^{-1} which is considerably smaller than for the two other methods. Nevertheless, the results from "map" also indicate that the spectroscopic signatures of the residues at the dimerization interface are not identical and differ from the monomer whereas for

the other residues considered the differences between monomer and dimer and the two monomers within the dimer are smaller.

In summary, all three methods agree in that a) the individual labels have their frequency maxima at different frequencies and b) in going from the WT monomer to the dimer the IR spectra of the labels involved in dimerization split and shift. The magnitude of the splitting and shifting differs between the methods which is not surprising given their very different methodologies. For the two mutants F24A and F24G the IR lineshapes using "scan" were determined for the residues involved in the dimerization interface and a selection of other residues, see Figure 3.1A. Compared with the WT monomer and dimer, characteristic shifts were found.

3.4.4 Frequency Fluctuation Correlation Functions

The frequency fluctuation correlation functions that can be computed from the frequency time series contain valuable information about the dynamics around a particular site considered, here the -CO groups of every residue. Specifically, FFCFs were analyzed for labels along the dimerization interface, for WT and the two mutant monomers and dimers, from using frequencies determined from "scan" and INM. Before discussing the FFCFs their convergence with simulation time is considered as it has been observed that an extensive amount of data is required.[106]

For this, the first 1 ns and the entire 5 ns run for WT insulin monomer was analyzed using "scan". For the 1 ns simulation snapshots every 10 fs and every 2 fs were analyzed (see Figure 3.15 top and middle row) and every 10 fs for the 5 ns simulations (Figure 3.15 bottom row). The computational resources required for such an analysis are considerable. Using 8 processors, the analysis of the 1 ns simulation for 10^5 snapshots (saved every 10 fs) takes 400 hours for a single spectroscopic probe. Figure 3.15 shows that except for one feature at ~ 3 ps for residue B26 the FFCFs from the 1 ns simulation with saving every 10 fs and every 2 fs are very similar. On the contrary, using snapshots from the 5 ns simulation leads to reducing the fluctuations in the FFCFs and determinants such as the static component (the value at a correlation time of 4 ps) are higher from the

longer simulation. A quantitative comparison for the time scales, amplitudes and static component (see Eq. 3.7) is provided in Table 3.2. The amplitudes and short decay times of all fits are within a few percent. The picosecond time scale (τ_2) can differ by up to 30 % (B26) and the offset Δ_0 can differ by a factor of two or more. To balance computational expense and quality of data, the remaining analysis was carried out with data from the 1 ns simulation with snapshots recorded every 10 fs.

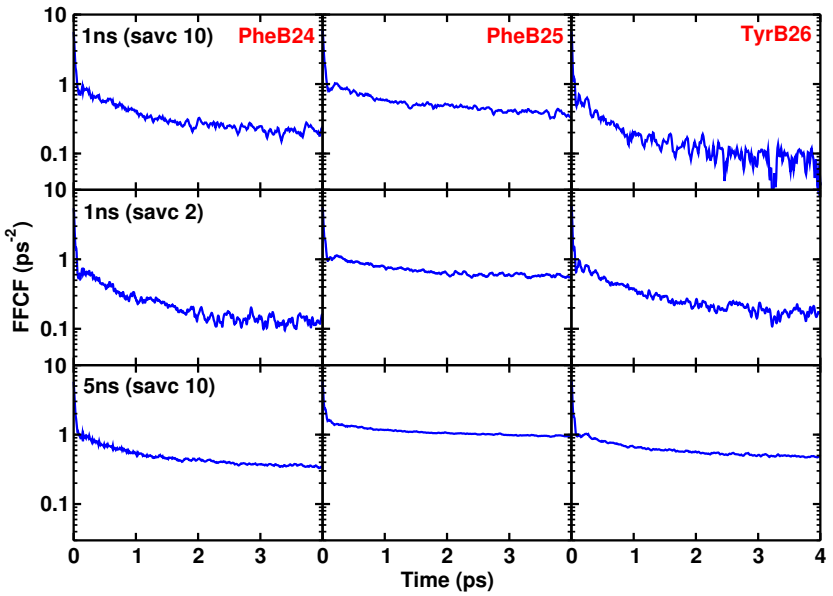


Figure 3.15: FFCF for residues at the dimerization interface (B24-B26) from frequency trajectories based on "scan" for WT monomer. The FFCF is shown based on different simulation lengths (1 ns and 5 ns) and computing frequencies from snapshots saved every 2 or 10 fs (savc2 and savc10). The overall shape of the FFCFs changes little whereas the noise level decreases especially for longer simulation times. Also, the magnitude of the static component increases for longer simulation times.

The FFCFs for B24 to B26 of the insulin monomer and the two monomers within the dimer are reported in Figure 3.16 together with the fits to Eq. 3.7. For the three labels from the monomer simulations the FFCFs differ in the longest decay time and the offset Δ_0. As for the infrared spectra, the three -CO labels exhibit different environmental dynamics. When compared with the two monomers in the insulin dimer these differences are even more pronounced. In general, all decay times increase to between 1 ps and $\sim$ 13 ps and the offset can be up to 5 times larger than for the monomer. This is owed to the considerably restrained dynamics of the residues at the dimerization interface compared with the free monomer.

	a_1	γ	τ_1	a_2	τ_2	Δ_0
1ns (savc10)						
B24	4.64	25.44	0.025	0.75	0.74	0.21
B25	4.94	22.19	0.028	0.66	0.98	0.37
B26	4.97	21.82	0.019	0.62	0.62	0.07
1ns (savc2)						
B24	4.64	25.77	0.023	0.65	0.68	0.13
B25	4.40	20.96	0.025	0.62	1.05	0.54
B26	4.89	25.63	0.020	0.77	0.70	0.17
5ns (savc10)						
B24	4.74	17.46	0.023	0.69	0.83	0.35
B25	4.24	0.00	0.022	0.60	1.07	0.94
B26	4.96	16.09	0.021	0.59	0.92	0.47

Table 3.2: Parameters obtained from fitting the FFCF to Eq. 3.7 from "scan" frequencies for residues (B24-B26) for WT monomer based on different simulation length and different time separations between coordinates analyzed (every 2 or 10 fs - nsavc2 and nsavc10). The amplitudes a_1 to a_3 are in ps^{-2}, the decay times τ_1 to τ_3 in ps, the parameter γ in ps^{-1}, and the offset Δ_0 in ps^{-2}.

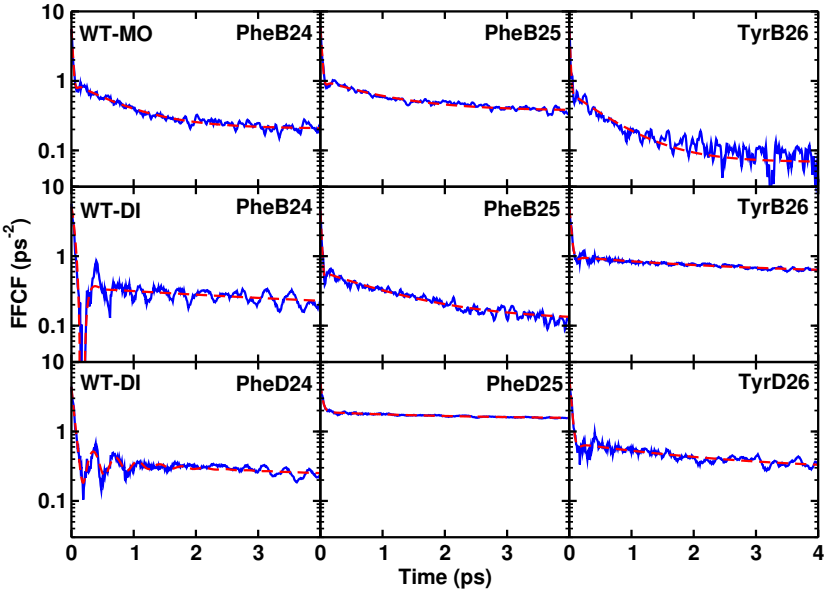


Figure 3.16: Comparison of the FFCFs for WT monomer and dimer for residues B24 to B26 at the dimerization interface. The frequencies are based on "scan" and snapshots from the 1 ns simulation, saved every 10 fs were analyzed.

Comparing the two monomer mutants with the WT it is found that the picosecond component is comparable whereas Δ_0 is similar (for F24A) or somewhat larger (for F24G), see Table 3.3. When moving to the mutant dimers, the differences with their monomeric counterparts are considerably smaller than for the WT system. This is likely to be related to a weakening of the F24A and F24G dimers

	a_1	γ	τ_1	a_2	τ_2	Δ_0	a_3	τ_3
WT monomer								
B24	4.64	25.44	0.025	0.75	0.74	0.21		
B25	4.94	22.19	0.028	0.66	0.98	0.37		
B26	4.97	21.82	0.019	0.62	0.62	0.07		
WT dimer M1								
B24	4.80	14.50	0.080	0.23	4.72	0.13		
B25	3.92	27.74	0.023	0.49	1.15	0.12		
B26	4.12	16.36	0.038	0.44	2.51	0.55		
WT dimer M2								
D24	0.30	17.59	0.56	3.68	0.039	0.18	0.19	4.08
D25	3.17		0.033	0.41	2.32	1.50		
D26	5.32	13.49	0.040	0.42	2.10	0.27		
F24A monomer								
B24	4.94	29.51	0.020	0.51	0.61	0.07		
B25	4.37	13.45	0.020	0.64	0.79	0.21		
B26	4.11	25.05	0.027	0.63	1.38	0.45		
F24A dimer M1								
B24	4.90	14.61	0.046	0.33	1.68	0.41		
B25	3.19		0.028	0.62	1.81	1.08		
B26	2.15		0.040	0.34	1.89	0.48		
F24A dimer M2								
D24	1.27		0.043	0.31	1.17	0.36		
D25	1.39		0.031	0.32	1.10	0.51		
D26	4.91	13.72	0.039	0.60	1.40	0.63		
F24G monomer								
B24	4.72	29.74	0.032	0.43	1.24	0.26		
B25	4.57	16.46	0.019	0.58	0.81	0.24		
B26	4.60	25.73	0.022	0.59	0.54	0.04	0.51	7.89
F24G dimer M1								
B24	1.42		0.028	0.21	1.02	0.37		
B25	3.70		0.018	0.48	1.18	0.24		
B26	3.87		0.029	0.68	1.90	0.88		
F24G dimer M2								
D24	2.50	38.76	0.016	0.30	1.29	0.17		
D25	1.53		0.030	0.25	1.70	0.65		
D26	3.94	5.32	0.042	0.27	2.14	0.23		

Table 3.3: Parameters from fitting the FFCF to Eq. 3.7 for frequencies from "scan" for the selected residues (B24-B26 and D24-D26). The amplitudes a_1 to a_3 in ps^{-2}, the decay times τ_1 to τ_3 in ps, the parameter γ in ps^{-1}, and the offset Δ_0 in ps^{-2}. For residues D24 in monomer M2 from the WT dimer and B26 in the F24G monomer the third time scale is required for a good fit.

which also allows water to penetrate more or less deeply into the dimer interface.[93] Overall, the dynamics still is slowed down in the mutant dimers by up to a factor of two compared with the mutant monomer but the effects are considerably less pronounced than for the WT systems.

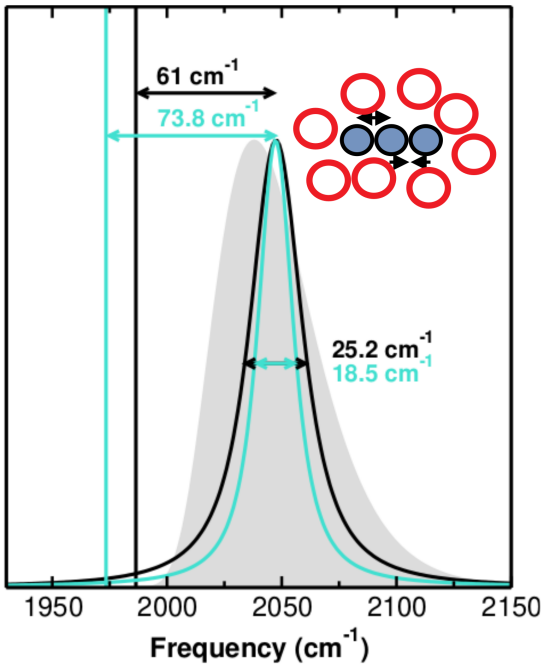
3.5 Conclusion

The present work demonstrates that WT insulin monomer and dimer and mutant monomers and mutant dimers lead to different spectroscopic and dynamical signatures for residues along the dimerization interface. This is found - to different extent - for all three approaches used for computing the frequency trajectory ("scan", INM, "map") and suggests that the overall findings do not depend strongly on the way how these frequencies are determined. The center frequency and FWHM for insulin monomer are in qualitative (scan along CO INM) or even quantitative (scan along [CONH] INM) agreement with experiment which, together with earlier investigations of the spectroscopy and dynamics of and around NMA,[105,116,127] provide a validation of the computational model. It is noteworthy that using one single parametrization for the -CO stretch and the multipoles on the [CONH] moiety of the peptide bond the experimentally observed FWHM for the protein is correctly described.

The fact that the stability differences between WT and mutant (here at position B24)[93] insulin dimer are also reflected in the spectroscopy and dynamics of WT and mutant insulin monomers and dimers suggests that spectroscopic investigations can be used to provide information about the association thermodynamics. This follows earlier suggestions for characterizing protein-ligand binding[96] which are supported by atomistic simulations.[97] For insulin this is particularly relevant because except for the WT dimer direct thermodynamic information about its stability appears to be missing. Replacing a thermodynamic approach by a spectroscopic characterization is an attractive alternative. The present work suggests that by combining quantitative simulations with modern experiments is a potentially useful way to obtain pharmacologically relevant information such as the strength of the modified insulin dimers.

Chapter 4

Vibrational Spectroscopy of N_3^- in the Gas- and Condensed-Phase



The results presented in this chapter have been previously published:
J. Phys. Chem. B 2019, 123, 3282-3290.
doi:10.1021/acs.jpcb.8b11430

***

Dr. Debasish Koner from the University of Basel contributed to this work as a second author.

4.1 abstract

Azido-derivatized amino acids are potentially useful, positionally resolved spectroscopic probes for studying the structural dynamics of proteins and macromolecules in solution. To this end a computational model for the vibrational modes of N_3^- based on accurate electronic structure calculations and a reproducing kernel Hilbert space representation of the potential energy surface for the internal degrees of freedom is developed. Fully dimensional quantum bound state calculations find the antisymmetric stretch vibration at 1974 cm^{-1} compared with 1986 cm^{-1} from experiment. This mode shifts by 64 cm^{-1} (from the frequency distribution) and 74 cm^{-1} (from the IR-lineshape) to the blue, respectively, compared with 61 cm^{-1} from experiment for N_3^- in water. The decay time of the frequency fluctuation correlation function is 1.1 ps, in good agreement with experiment (1.2 to 1.3 ps) and the full width at half maximum of the asymmetric stretch in solution is 18.5 cm^{-1} compared with 25.2 cm^{-1} from experiment. A computationally more efficient analysis based on instantaneous normal modes is shown to provide comparable, albeit somewhat less quantitative results compared to solving the 3-dimensional Schrödinger equation for the fundamental vibrations.

4.2 Introduction

Characterizing the structural and functional dynamics of complex systems in the condensed phase is a challenging problem, spanning several spatial and temporal scales.[137] One particularly elegant way to quantitatively assess the structure and dynamics of the solvent environment surrounding a probe molecule is to use optical spectroscopy, especially 1- and 2-dimensional infrared (1D-IR, 2D-IR) spectroscopy.[138] Two-dimensional IR spectroscopy is a powerful method for measuring the subpicosecond to picosecond dynamics in condensed-phase systems and considerably extends the toolbox of optical spectroscopy.

Using 2D-IR spectroscopy, the coupling between inter- and intramolecular degrees of freedom such as the hydrogen bonding network in solution, or structural features of biological macromolecules can be investigated by monitoring the fluctuation of fundamental vibrational frequencies of a probe molecule or ligand attached to a complex or a biological macromolecule. For quite some time, the amide-I stretching frequency has been used for this and has provided funda-

mental, molecular-level insight into the structural dynamics of small molecules and proteins alike.[128,139] For example, the CO chromophore was used as a probe to investigate the solvation dynamics of N-Methylacetamide in water.[105,139–144] Similarly, the CN^- stretching frequency has been used as a probe to investigate its own solvation dynamics[104,145–147] and structural and energetic features of a protein-ligand complex by attaching the probe to benzene[96,97].

A versatile, spatially sensitive spectroscopic probe can also report on the dynamics of a system while minimizing the perturbations induced. The fundamental modes of a triatomic ligand such as the azide ion (N_3^-) are weakly coupled to the molecular framework it is bound to and can act as a sensitive probe for the environmental dynamics. The asymmetric stretching mode of N_3^- has a large oscillator strength and a vibrational transition frequency well separated from most organic chromophores. This makes it an ideal spectroscopic probe.[148–155] Other triatomic anions such as SCN^-, NCS^- or OCN^- have also served as spectroscopic probes.[156,157]

For a quantitative molecular level description of the environmental dynamics, a high-quality potential energy surface (PES) for the internal vibrational dynamics of the probe is required. Typically, differences of a few cm^{-1} are found when immersing the same probe into two different electrostatic environments, e.g. within a wild type and a single mutant protein.[96,97] Here, we use a combination of normal mode and numerically exact bound state calculations together with MD simulations in the gas- and condensed phase to characterize the vibrational dynamics of N_3^- with the aim to provide the basis for its use as a covalently linked, positionally sensitive spectroscopic probe. The instantaneous deviation in the solute-solvent interactions is reflected by the fluctuation in the transition frequency of the fundamental modes. The time scale of the structural changes around the solute molecule can be determined from the decay of the frequency fluctuation correlation function (FFCF) from which also the lineshape of the 1D-IR spectrum can be obtained.[128]

The vibrational dynamics of N_3^- in the gas phase and solution has been investigated from experiment and computation. Experimentally, the NN asymmetric

stretching frequency of the azide ion was determined to be at 1986.47 and 2047.5 cm^{-1} in the gas phase and bulk water, respectively[151,158,159] which make it a potentially useful environmental probe for protein dynamics, see Figure 4.1. Using non-linear infrared spectroscopy vibrational lifetimes of the asymmetric stretch fundamental of azide anion in water have been measured experimentally to be $T \approx 1.2$ ps (in H_2O)[160] and $T = 1.3$ ps (in D_2O).[149] The band width of the N_3^- asymmetric stretch in bulk water is 25.2 cm^{-1}.[151,160] Due to the triple bond character of azide[161], its structure would stabilize through interaction with the solvent which leads to a blue shift (61 cm^{-1}) compared to the gas phase frequency.

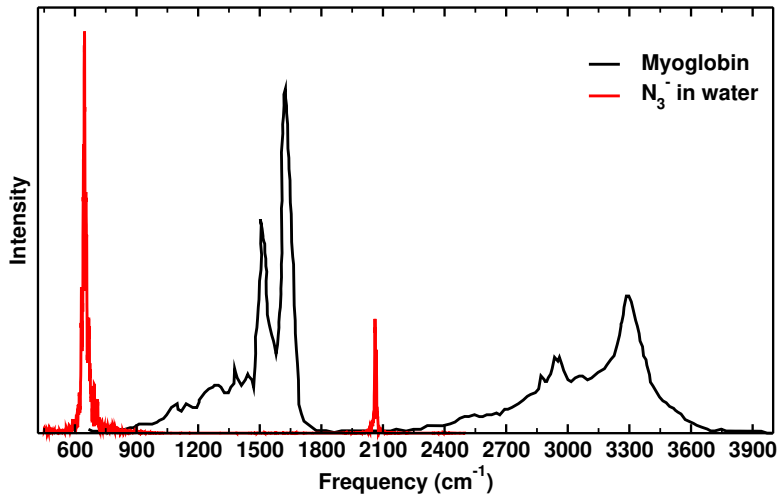


Figure 4.1: Superposition of a measured (black) infrared spectrum for Mb[24] and the computed IR sspectrum for N_3^- (red). The asymmetric stretch vibration (at $\sim$ 2055 cm^{-1} falls in an ideal range where protein absorptions are rare and make azide an ideal spectroscopic probe.

Recently, an explicitly parametrized PES was computed using an accurate composite method based on pointwise, individually scaled fc-CCSD(T*)-F12b calculations including scalar relativistic effects and the aug-cc-pV5Z basis set. Using this PES, variational calculations determined the lower bound states for N_3^- in the gas phase with spectroscopic accuracy.[162] However, such a composite method is not feasible for the ultimate purpose of the present development, which is an accurate representation of the inter- and intramolecular interactions for N_3^- as a spectroscopic probe covalently linked to another molecular building block. In order to 1) accurately capture the energetics of distorted N_3^- geometries away from the equilibrium structure and 2) to cover situations in which the probe (N_3^-) and the moiety it is linked to (e.g. an amino acid or a small organic molecule) are

coupled electronically, it was decided to compute the PES at the multi reference CI (MRCI) level of theory. This also allows to extend the grid to a range that covers geometries further away from the equilibrium to sample conformations in MD simulations in sterically demanding environments due to local interactions and constraining effects.

In an effort to quantitatively capture the environmental dynamics, the present work introduces a computational model for the azido group capable of describing the molecular vibrations of the probe in gas- and the condensed-phase in a realistic manner. To this end, an accurate intramolecular, fully dimensional PES for N_3^- is computed based on high-level electronic structure calculations and represented as a reproducing kernel Hilbert space (RKHS). The vibrational transition frequencies of the fundamental modes are calculated from instantaneous normal mode (INM) analysis and from solutions of the nuclear Schrödinger equation (SE). Results from bound state calculations including the vibrational line shapes from frequency correlation functions and the time scales for the environmental dynamics derived from them are then compared with data from experiments in the gas- and condensed phase.

First, the computational methods are presented. This is followed by quantum bound state calculations to determine the stationary states for the low lying bound states. Next, the solvation dynamics in water is considered and the 1D-IR and the frequency fluctuation correlation functions, which are directly related to 2D-IR spectroscopy, are determined and discussed.

4.3 Methods

4.3.1 The Potential Energy Surface for N_3^-

The *ab initio* energies were computed at the multi reference configuration interaction (MRCI) level of theory including the Davidson correction (MRCI+Q) with the augmented Dunning type correlation consistent polarized quadruple zeta (aug-cc-pVQZ) basis set using the Molpro software.[163,164] All calculations were carried out for the $^1A'$ electronic state in C_s symmetry. Initial orbitals for the MRCI calculations were computed using the state averaged complete active space

self-consistent field (SA-CASSCF) method for the first two $^1A'$ states with 16 electrons in 12 active orbitals and the $1s$ orbitals of the nitrogen atoms were kept frozen.

The *ab initio* energies were calculated in Jacobi coordinates (R, r, θ) (see the inset in Figure 4.2), where r is the distance between the two nitrogen atoms (N1 and N2), R is the distance between their center of mass and the third N atom (N3), and θ is the angle between $\vec{r}$ and $\vec{R}$. Evaluation of the necessary integrals in the bound state calculations (see below) is stablest and can be efficiently done if Gauss-Legendre points are used for the angular degrees of freedom.[165] Therefore, the angular grid (θ) used here contains 10 Gauss-Legendre quadrature points between 0 and 90° (6.721, 15.427, 24.184, 32.953, 41.726, 50.502, 59.278, 68.056, 76.833 and 85.611°) considering the symmetry of the system. The radial grids include 17 points along r ranging from 0.80 to 1.60 Å and 14 points along R between 1.38 and 2.14 Å.

For the bound state calculations and the molecular dynamics (MD) simulations in the gas phase and in solution a continuous and differentiable representation of the *ab initio* energies is required. Here, a reproducing kernel Hilbert space based approach[73,74] is used. A RKHS interpolation exactly reproduces a set of known function values and provides approximate values of the function at points where the values are unknown.

For the 1-dimensional kernels the linear problem $f(x_i) = \sum_j \alpha_j k(x_i, x_j)$ is solved which yields the coefficients α_j from which the value of the function $f(x) = \sum_i \alpha_i k(x_i, x)$ can be obtained for arbitrary x. The 3-dimensional kernel $K(X, X')$ is then a tensor product of three 1-dimensional kernels.[74,166]

Based on this, the 3-dimensional kernel K is

$$K(X, X') = k^{(n,m)}(R, R')k^{(n,m)}(r, r')k^{(2)}(z, z'). \tag{4.1}$$

where X stands for all dimensions involved and

$$z = \frac{1 - \cos(\theta)}{2}, \tag{4.2}$$

which maps the angle θ onto the interval $[0, 1]$. The reciprocal power decay kernel ($k^{(n,m)}$) with smoothness $n = 2$ and asymptotic decay $m = 6$ is used for the radial dimensions (i.e., r and R)

$$k^{(2,6)}(x, x') = \frac{1}{14}\frac{1}{x_>^7} - \frac{1}{18}\frac{x_<}{x_>^8}, \tag{4.3}$$

where $x_>$ and $x_<$ are the larger and smaller values of x and x', respectively. For the angular degree of freedom, a Taylor spline kernel is used

$$k^2(z, z') = 1 + z_< z_> + 2z_<^2 z_> - \frac{2}{3}z_<^3, \tag{4.4}$$

where $z_>$ and $z_<$ are similar to $x_>$ and $x_<$.

4.3.2 Quantum Bound State Calculations

For solving the 3-dimensional Schrödinger equation of N_3^-, the DVR3D[167] software was used. DVR3D employs a discrete variable representation (DVR) based on Gauss-Laguerre quadratures for the radial and Gauss-Legendre quadratures for angular coordinates and thus yields a fully point-wise representation of the wave function. Solution of the vibrational problem is based on successive diagonalisation and truncation which is possible for a number of possible coordinate systems. After solving the vibrational bound state problem using DVR3D the wave functions and expectation values are obtained. The wave functions are then inspected to assign the quantum numbers of the vibrational states. For using the RKHS PES, an interface between DVR3D and the RKHS kernel code[74] was written. This code handles the transformations between the coordinates in which the PES is expressed and the coordinates employed by DVR3D.

Here, homonuclear Jacobi coordinates are employed because the PES is expressed in this coordinate system and the necessary angular integrals can be done efficiently and accurately using Gauss-Legendre quadrature. However, the problem can also be solved using other coordinate systems, e.g. bond-length bond-angle

or Radau coordinates (ideal for heavy-light-heavy systems).[168] Ultimately, the choice of coordinate system primarily affects the rate of convergence of the eigenvalues which is relevant for highly excited states, but less so for the fundamentals considered in the present work. Here, the initial wave function is described by Morse oscillator functions for the radial coordinates. Several tests were carried out to converge the results by varying different parameters. The radial grid along r is defined by 40 points from 0.830 to 1.361 Å (1.57 to 2.57 a_0) using $r_e = 1.111$ Å (2.1 a_0), $D_e = 125.5$ kcal/mol (0.2 E_h) and $\omega_e = 3512$ cm^{-1} (0.016 E_h). For R, the grid is expanded from 1.409 to 2.106 Å (2.66 to 3.98 a_0) with 64 points and values of $r_e = 1.773$ Å (3.35 a_0), $D_e = 163.2$ kcal/mol (0.26 E_h) and $\omega_e = 2634$ cm^{-1} (0.012 E_h), respectively. For θ, 56 Gauss-Legendre quadrature points were used. Maximum dimensions of the intermediate 2D Hamiltonian and the final Hamiltonian were 800 and 1200. For rotationally excited ($J > 0$) state calculations the $z-$axis was placed along $\vec{R}$.

4.3.3 Molecular Dynamics Simulations

For the MD simulations of N_3^- in solution, CHARMM[64] was used. Again, an interface between CHARMM and the multi-dimensional RKHS code was written for this. For the simulations in water the TIP3P[117] model was used (Figure 4.2). The nonbonded interactions were treated with a 12 Å cutoff switched at 8 Å. All bonds involving hydrogen atoms are constrained using SHAKE[123] as is also typically done for simulating proteins and polypeptides in solution.

Simulations of N_3^- in water were carried out in a pre-equilibrated cubic box of size 30^3 Å^3. First, the system was minimized using 2000 steps of steepest descent (SD) and 200 steps of Newton Raphson (ABNR) followed by 20 ps of heating to 300 K and 3.5 ns of equilibration MD at 300 K. Production simulations 8 ns in length were run in the NVE ensemble using the velocity Verlet integrator.

For simulations in the condensed phase, nonbonded parameters for the N_3^- solute are required. The Lennard-Jones parameters for the nitrogen atoms were $\epsilon = 0.08485$ kcal/mol and $R_{\min}/2 = 1.66$ Å.[169] Charges are calculated following the Mulliken population analysis from the density matrix obtained at the MRCI/aug-cc-pVQZ level of theory. This yields a charge of $-0.53462e$ for the

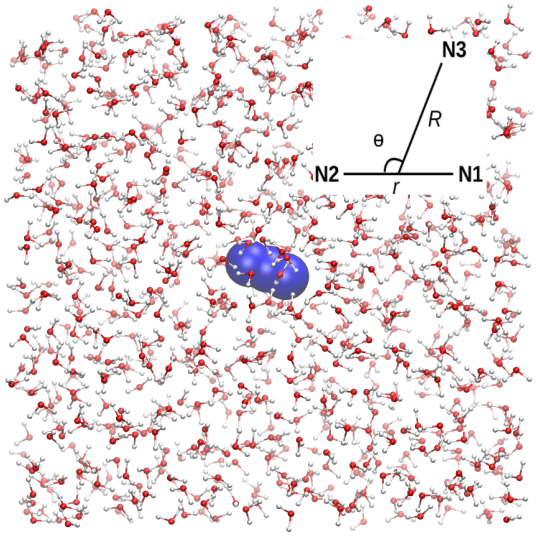


Figure 4.2: The simulation system used in the present work. N_3^- is displayed as blue van der Waals spheres and water molecules are shown as red (O atom) and white (H atom) ball and stick representation. The inset shows the Jacobi coordinates which define the *ab initio* energy grid.

terminal nitrogen atoms and $0.06924e$ for the central one. As an extension, multipolar representations of the electrostatics could also be considered although for closed-shell systems (e.g. N-methyl-acetamide) it has been found that point charge models can be quite reliable for spectroscopic[105] and thermodynamic[70,71] properties.

From the production simulation, 1.2×10^6 snapshots are taken as a time-ordered series for computing the frequency fluctuation correlation function and the 1D-IR spectrum. The FFCF was determined from either instantaneous harmonic vibrational frequencies based on a normal mode analysis or by using anharmonic frequencies calculated from solving the 3-dimensional nuclear Schrödinger equation using DVR3D. Normal modes of N_3^- were calculated for each snapshot after minimizing N_3^- and keeping the central N atom and the surrounding solvent frozen. Thus frequency trajectories were obtained for ω_1, ω_2 and ω_3 which correspond to the symmetric stretch, the bending and the asymmetric stretch vibration, respectively. The anharmonic transition frequencies ν_i were calculated using DVR3D calculations based on a 3-dimensional PES for N_3^- generated at each snapshot. To compute the 3D PES, the grid was defined in internal coordinates ($R_{\mathrm{N1-N2}}, R_{\mathrm{N2-N3}}, \angle N1N2N3$) by fixing the position of the central nitrogen,

N2, and all water molecules. The grid consists of 15 points along $R_{\mathrm{N1-N2}}$ and $R_{\mathrm{N2-N3}}$ from 0.82 to 1.35 Å and 7 points for $\angle N1N2N3$ between 155° to 180° (158.490, 161.675, 164.859, 168.043, 171.225, 174.402 and 177.561°). The total energies were then calculated using the CHARMM+RKHS-module to obtain an analytical (RKHS) 3D PES for each snapshot. Finally, the ν_i were determined by solving the 3D Schrödinger equation for bound states using DVR3D as discussed before based on the 3D RKHS PES.

From the frequency trajectory $\omega_i(t)$ (or $\nu_i(t)$; all expressions below pertain to anharmonic frequencies in a similar fashion) of a particular mode i, its frequency fluctuation correlation function, $\langle\delta\omega(0)\delta\omega(t)\rangle$ is computed. Here, $\delta\omega(t) = \omega(t) - \langle\omega(t)\rangle$ and $\langle\omega(t)\rangle$ is the ensemble average of the transition frequency. From the FFCF the line shape function is determined within the cumulant approximation

$$g(t) = \int_0^t \int_0^{\tau'} \langle\delta\omega(\tau'')\delta\omega(0)\rangle d\tau'' d\tau'. \tag{4.5}$$

To compute the line shape function $g(t)$, the FFCF is fitted to a general expression[129]

$$\langle\delta\omega(t)\delta\omega(0)\rangle = a_1 \cos(\gamma t) e^{-t/\tau_1} + \sum_{i=2}^{n} a_i e^{-t/\tau_i} + \Delta_0 \tag{4.6}$$

where a_i, τ_i, γ and Δ_0 are fitting parameters. The parametrization of this fitting function is motivated by the overall shape of the FFCF[129] and has been used in previous work.[28,104] It is an extension of the typical multiexponential decay, which is traditionally employed[149], in order to capture the anticorrelation at short times. Furthermore, this functional form also allows analytic integration[128] to obtain $g(t)$ in Eq. 4.5. The decay times τ_i of the frequency fluctuation correlation function reflect the characteristic time-scale of the solvent fluctuations to which the solute degrees of freedom are coupled.

The 1D-IR spectra is then calculated as[151]

$$I(\omega) = 2\Re \int_0^\infty e^{i(\omega - \langle\omega\rangle)t} e^{-g(t)} e^{-\frac{t}{2T_1}} e^{-2D_{\mathrm{OR}}t} dt, \tag{4.7}$$

where $\langle\omega\rangle$ is the average transition frequency obtained from the distribution, T_1 (0.8 ± 0.1 ps) is the vibrational relaxation time and $D_{\mathrm{OR}} = 1/6T_R$ with $T_R = 1.3 \pm 0.3$ ps is the rotational diffusion coefficient which account for life-

time broadening.[150]

4.4 Results

4.4.1 Analytical Potential Energy Surface

The analytical PES for N_3^- was constructed from 2231 *ab initio* energies using a RKHS. To test its quality, an additional 245 *ab initio* energies are computed at off-grid points (which were not used in the RKHS interpolation) and are compared with the energies obtained from the RKHS PES, see Figure 4.3. The correlation coefficient between the *ab initio* and analytical energies is $R^2 = 0.9999$ which confirms the high quality and interpolative power of the RKHS PES.

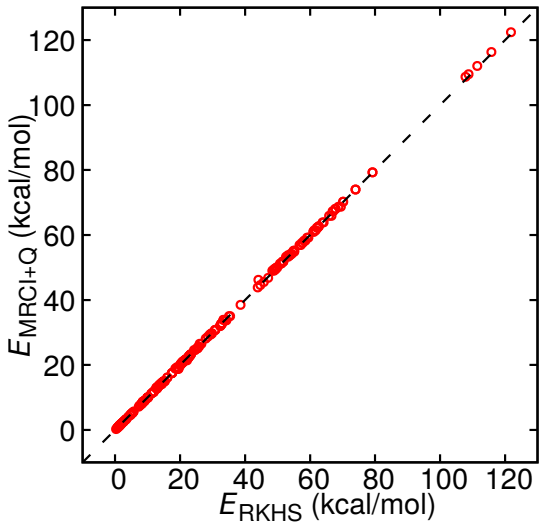


Figure 4.3: Correlation between the *ab initio* and analytical energies obtained from RKHS interpolation for a set of 245 randomly selected validation points with $R^2 = 0.9999$ and an RMSE of 30 cm^{-1} for energies up to 28.6 kcal/mol (10000 cm^{-1}). The energy of the global minimum is at $E = 0$.

The global minimum of the RKHS PES is the linear N-N-N configuration with two equal N-N bonds of length $r_e = 1.187$ Å. This is in good agreement with the *ab initio* optimized structure (1.189 Å) obtained at the MRCI+Q/aug-cc-pVQZ level of theory in the present work and with diode laser velocity modulation spectroscopy experiment[158] with a value of 1.188402 Å. In Figure 4.4 the contour plots of the RKHS PES around the global minimum are shown. The equilibrium

bond length from the composite calculations[162] is 1.18461 Å which is somewhat too short.

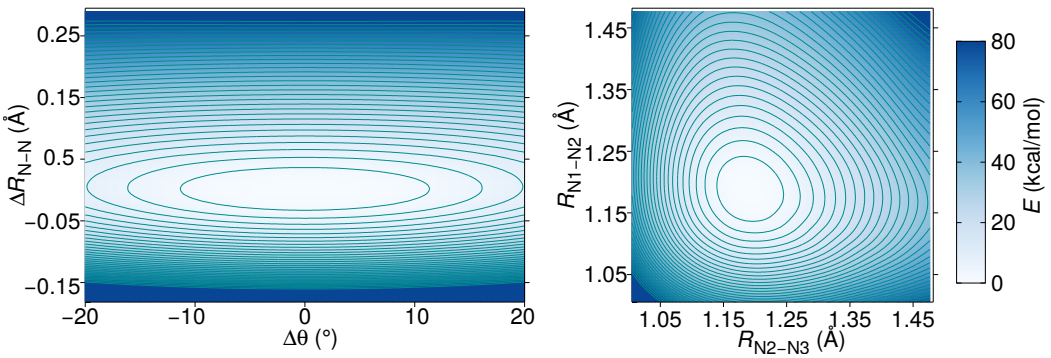


Figure 4.4: Contour diagrams of the RKHS PES. Left panel: for $R_{\mathrm{N1-N2}} = R_{\mathrm{N2-N3}}$ and $\Delta\theta = 0°$, $\Delta R_{\mathrm{N-N}} = 0$ corresponds to the equilibrium structure. Right panel: for linear N_3^-. The spacing between contours is 2.5 kcal/mol. The zero of energy is the global minimum of the system.

4.4.2 Spectroscopy and Dynamics in the Gas Phase

The quantum bound states for N_3^- were calculated using DVR3D for $J = 0$ and 1. For the fundamental stretching modes the quantum numbers were assigned upon inspection of the nodal structure of the wave functions, see Figure 4.5. The transition frequencies for the symmetric (ν_1) and asymmetric (ν_3) stretching mode from DVR3D are 1305.3 and 1973.5 cm^{-1}, respectively. Since, N_3^- is linear, the bending fundamental (ν_2) corresponds to a $J = 1$ state and is computed using DVR3D at 632.8 cm^{-1}.

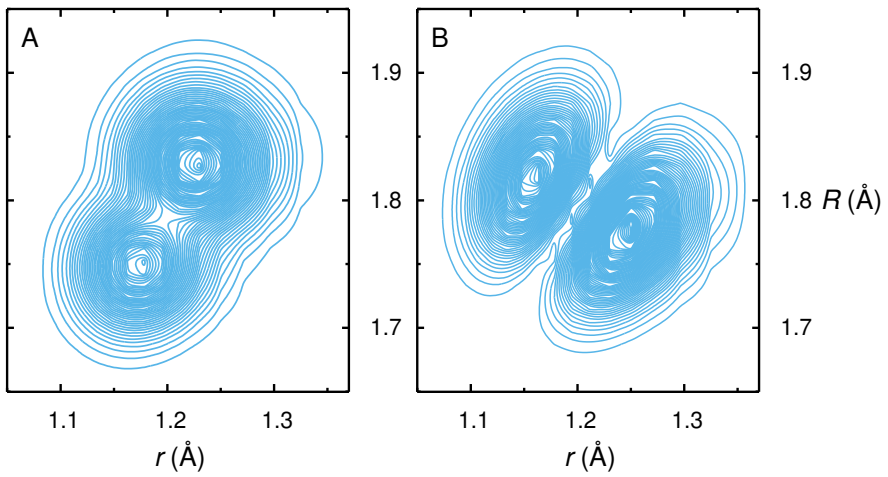


Figure 4.5: Contour plots of the squared wave function amplitude $|\Psi|^2$ associated with the ν_1 (A) and ν_3 (B) mode from DVR3D calculation on the RKHS PES for N_3^-.

The three fundamental transition frequencies obtained from different methodologies are reported in Table 4.1 along with the experimentally[158,159,170] and theoretically[162] determined frequencies from the literature. The normal modes shown in Table 4.1 were calculated using CHARMM together with the RKHS PES. It can be seen that except for the asymmetric stretch, anharmonic and harmonic frequencies agree to within < 5 cm^{-1}. Experimentally, the NN asymmetric stretch in the gas phase was found at 1986.47 cm^{-1} compared with 1973.5 cm^{-1} from the DVR3D calculation[158,159] and 1986.36 cm^{-1} from variational calculations on a parametrized fc-CCSD(T)/aug-cc-pVQZ PES (see Introduction).[162] The difference of ~ 10 cm^{-1} is most likely due to the different level of theory (MRCI vs. composite method based on fc-CCSD(T)) and the smaller basis set used here (aug-cc-pVQZ vs. aug-cc-pV5Z) because the bound state calculations are essentially converged.

Mode	Expt.	NMA	DVR3D	Power Spec. (300 K)	Ref.[162]
$[1,0,0]$	1344*	1306.8	1305.3	1307.4	1307.9
$[0,1,0]$	∼640*	636.6	632.8	637.2	629.3
$[0,0,1]$	1986.47[158,159]	2004.6	1973.5	2005.2	1986.4

* Frequencies of N_3^- in potassium halide salts[170]

Table 4.1: Fundamental transition frequencies in cm^{-1} for N_3^-. The frequencies reported in Ref.[162] are obtained from QM calculations on an analytical PES based on CCSD(T)/aug-cc-pV5Z energies.

To the best of our knowledge, there is no direct experimental data for the bending and symmetric stretching mode in the gas phase. However, infrared and Raman spectra of the azide ion in potassium azide have been determined (642.2 cm^{-1} in KN_3, 642.7 cm^{-1} in KCl cm^{-1}, 640.0 cm^{-1} in KBr, and 638.5 cm^{-1} in KI[170]). These frequencies compare with 632.8 cm^{-1} from quantum calculations. For the symmetric stretch a value of 1344 cm^{-1} was reported[170] in KN_3 somewhat higher, i.e. blue shifted, than 1305.3 cm^{-1} from DVR3D calculations in the gas phase. The variational bound state calculations on the fitted fc-CCSD(T)/aug-cc-pV5Z PES yielded $\nu_2 = 629.3$ cm^{-1}, and $\nu_1 = 1307.9$ cm^{-1}, respectively.[162] All these results agree favourably with the present calculations, see Table 4.1.

As an independent validation of the bound state calculations, NVE MD simulations with the RKHS-MRCI PES were carried out for N_3^- in the gas phase

using CHARMM. Power spectra for the distances r_i (Ni-Nj separation) were determined and compared with results from bound state calculations. Similar analyses were carried out for N_3^- in solution. Moreover, the infrared spectra for N_3^- in water was also calculated from the dipole moment autocorrelation function.

This suggests that the MD simulations provide a meaningful surrogate for the quantum calculations. Such comparisons are important when considering simulations in solution for which rigorous quantum calculations are not possible and conclusions must be drawn either from MD simulations or from other approximate treatments. For the simulations in solution the maxima of the power spectra shift to $\nu_1 = 1355.4$ cm^{-1}, and $\nu_3 = 2061.9$ cm^{-1} while for 1D-IR spectrum, the peaks are $\nu_2 = 646.8$ cm^{-1}, and $\nu_3 = 2061.1$ cm^{-1}, see Figure 4.6 and Table 4.2.

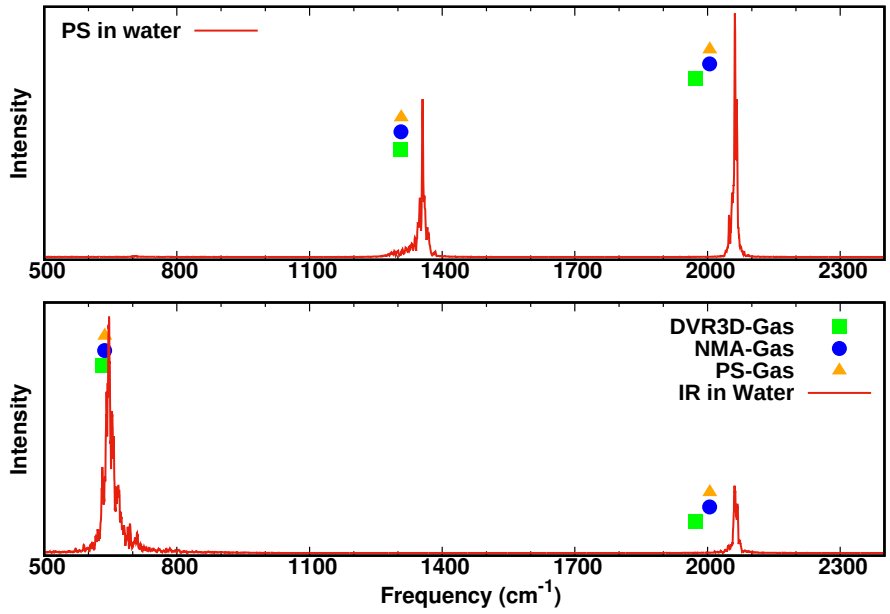


Figure 4.6: Power and IR spectra for N_3^- in solvent are shown in the upper and lower panel, respectively. The maxima of power ($\nu_1 = 1355.4$ cm^{-1}, and $\nu_3 = 2061.9$ cm^{-1}), IR spectra ($\nu_2 = 646.8$ cm^{-1}, and $\nu_3 = 2061.1$ cm^{-1}) in solvent and the results from DVR3D quantum calculations (green square), the normal mode analysis (blue circle) and the power spectra (orange triangle) all in the gas phase, are shown as well.

Mode	Expt	DVR3D	INM	Power Spectrum	IR Spectrum
$[1,0,0]$			1327.7	1355.4	
$[0,1,0]$			651.2		646.8
$[0,0,1]$	2047.5	2038.0	2047.3	2061.9	2061.1
$\Delta\nu_{[0,0,1]}$	61	64	43	57	
		74*	45*		

Table 4.2: Vibrational Frequencies cm^{-1} for N_3^- in solvent from experiment[151,160] and the present simulations. For the asymmetric stretch fundamental $[0,0,1]$ the vibrational blue shift from $P(\omega)$ and the line shape (asterisk) with respect to the gas phase values is also indicated. For INM this is a harmonic shift $\Delta\omega$ and for all other cases it is an anharmonic shift $\Delta\nu$ with respect to the fundamentals in the gas phase, see Table 4.1.

4.4.3 Dynamics and Spectroscopy in Solution

To serve as a positionally resolved spectroscopic probe, it is important to characterize the vibrational spectroscopy of N_3^- in solution. The power spectra (see Figure 4.6) indicate a blue shift of $\Delta\nu_3 = 56.7$ cm^{-1} for the asymmetric stretch. Experimentally, the anti-symmetric stretching mode of the azide ion in bulk water has been found at 2047.5 cm^{-1} which amounts to a blue shift of 61 cm^{-1} for the $[0,0,1]$ mode and serves as a comparison for the present simulations.[151] This is also consistent with the situation in CN^- for which the experimentally observed blue shift is 44 cm^{-1} compared with a value of 36 cm^{-1} from atomistic simulations.[104,171,172]

For more rigorous calculations two different analyses for the vibrational dynamics of azide in solution are considered. One of them uses normal mode calculations for a given solution configuration and the other one recomputes the *effective* 3-dimensional potential energy surface $V(R,r,\theta)$ for a given solvent configuration from which the vibrational states of interest are determined from solving the 3D Schrödinger equation, see Methods.

To determine the peak frequencies of the distributions $P(\omega)$, the raw data was fitted to a log normal probability distribution, see Figure 4.7. The peak values of $P(\omega)$, are reported in Table 4.2 along with results obtained from Power and IR spectra and from experiment. The frequency distribution peaks from instantaneous normal modes for the bending, symmetric, and asymmetric stretch are at 651.2 cm^{-1}, 1327.7 cm^{-1}, and 2047.3 cm^{-1}. This compares with the gas phase

frequencies using the same analysis at 636.6 cm^{-1}, 1306.8 cm^{-1}, and 2004.6, respectively. From comparing the frequency distributions all modes are shifted to the blue by between 15 cm^{-1} and 43 cm^{-1} (see Table 4.2). For N_3^- in solution it is expected that $P(\omega)$ deviates somewhat from a Gaussian distribution. Because all fundamentals are shifted to the blue, it is also expected that $P(\omega)$ contains a tail at higher frequencies. This is indeed the case (Figure 4.7), and the log-normal fits the raw data quite well which indicates that sampling is sufficient and the frequency trajectory is close to converged.

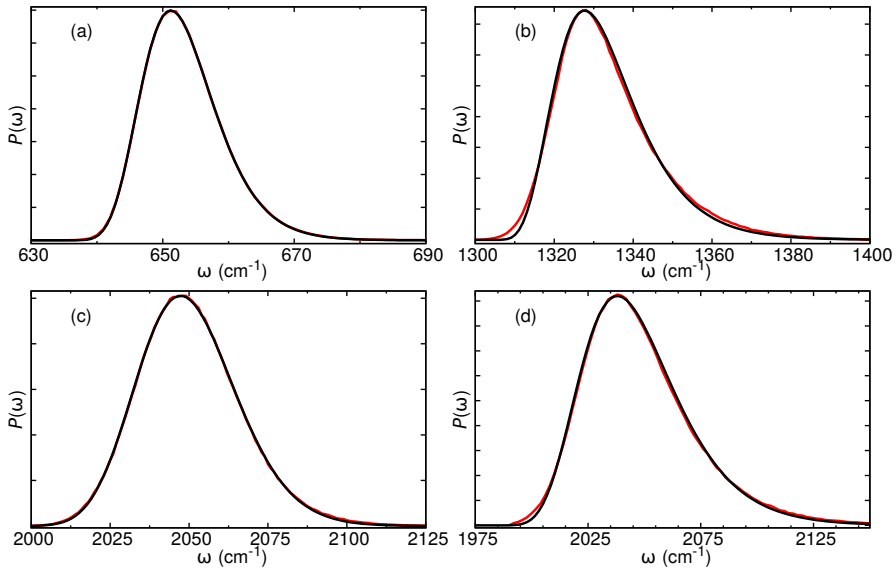


Figure 4.7: Distributions of instantaneous frequencies (red lines) calculated via normal mode analysis (INM) and quantum mechanical (QM) approach. Fits of $P(\omega)$ to 'lognormal probability distribution' functions are shown as black lines. Panel (a), (b) and (c) represents the INM results for bending, symmetric, and asymmetric stretch modes, respectively. Panel (d) shows $P(\omega)$ obtained from QM DVR3D calculations.

The blue shift from solving the 3-d SE is 64 cm^{-1} (according to $P(\omega)$) and 74 cm^{-1} when considering the IR-lineshape, see Table 4.2 and Figure 4.8. This is in reasonable to good agreement with experiment (61 cm^{-1}) and provides a meaningful basis for future applications of the present model in positionally resolved spectroscopy of peptides and proteins in solution. It is also worthwhile to note that the power spectrum, which is readily available from equilibrium MD simulations, yields a meaningful description of the blue shift ($\Delta\nu = 57$ cm^{-1}).

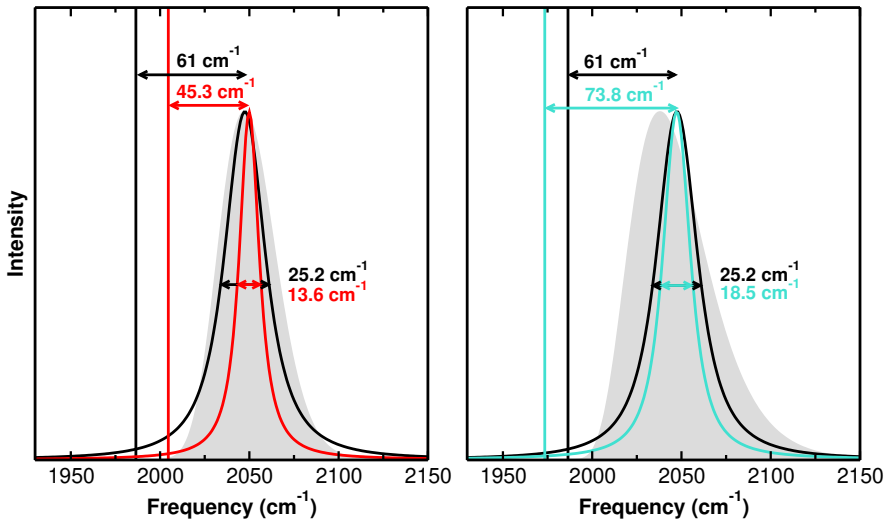


Figure 4.8: 1D-IR spectrum of N_3^- in water obtained from INM (left panel) and solution of the 3-d Schrödinger equation (3-d SE, right panel) calculations is compared with the FT-IR experiment.[151,158–160] The gas phase frequencies are shown as vertical lines and the remaining traces are experimental lineshapes and those determined from integrating the FFCFs. The black, red and turquoise lines represent experimental, INM and 3-d SE results, respectively. The grey areas indicate the distributions of transition frequencies, $P(\omega)$, for N_3^- in solution from 1.2×10^6 snapshots. The peak frequencies for the 1D-IR spectra from INM and 3-d SE calculations are at 2049.9 cm^{-1} and 2047.3 cm^{-1}, respectively.

To characterize the solvent dynamics around N_3^- in solution, the frequency correlation function for the NN asymmetric stretch from INM was determined. The data was fitted to Eq. 4.6 including three time scales, see Figure 4.9 and the corresponding fitting parameters are reported in Table 4.3. Three time scales can be distinguished in the FFCF: two sub-picosecond time scales ($\tau_1 = 0.044$ ps and $\tau_2 = 0.23$ ps) with large amplitude and one on the picosecond ($\tau_3 = 1.18$ ps) with the smallest amplitude.

Similarly, the FFCF for the asymmetric stretching mode was determined from the same 1.2×10^6 snapshots using DVR3D to solve the 3-dimensional nuclear Schrödinger equation based on scanning the 3-dimensional PES for an instantaneous solvent configuration. The FFCF determined from this time series was again fitted to Eq. 4.6 using three time scales, see Figure 4.9B. The corresponding fitting parameters are also reported in Table 4.3. The three time scales are $\tau_1 = 0.048$ ps, $\tau_2 = 0.21$ ps and $\tau_3 = 1.06$ ps. As for FFCF from INM, τ_3 agrees well with experimental values in D_2O ($\tau = 1.3$ ps)[149] and in H_2O ($\tau \approx 1.2$ ps).[160]

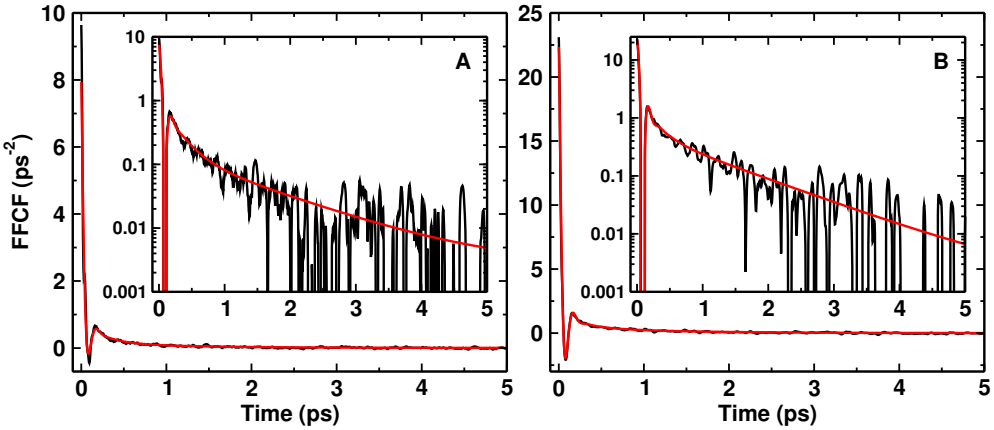


Figure 4.9: FFCF for the asymmetric stretch vibration as a function of time (black line) from A) INM and B) solution of the 3-dimensional Schrödinger equation for azide in water. Solid red line is the corresponding fit to Eq. 4.6 with three time scales. The inset shows the same data on a logarithmic scale for the $y-$axis. The rapid initial drop of the FFCF within the first 100 fs (the inertial component [173]) reflects the intermolecular hydrogen bond vibration.

Both FFCFs display a pronounced minimum at very short time ($t \sim 0.1$ ps). This feature is known from previous simulations [28] and related to the strength of the interaction between solvent and solute. [104,105,129] As this interaction is expected to decrease in going from CN^-, N_3^-, H_2O to the amide-I mode in N-methyl-acetamide, the observed behavior in the FFCF is consistent with such an interpretation. A final, less critical [129] characteristic of the FFCF is the amplitude $C(0)$ at zero time of the unnormalized FFCF. Previous simulations [28] found a value of ~ 500 cm^{-1} compared with a value of ~ 275 cm^{-1} from experiment. [149] The FFCF from the instantaneous normal modes and the quantum bound state calculations yield 272 cm^{-1} and 651 cm^{-1}, respectively. On the other hand, it has been found that the experiment is not particularly sensitive to the very short time dynamics [28] as evidenced by the absence of the short-time anticorrelation. It is noted that the general appearance of the FFCF in Figure 4.9 is comparable to that for CN^- and NMA in water [104,105] and exhibits somewhat larger fluctuations compared with previous simulations of N_3^- in water. [28] This is probably related to the fact in the present case the solute was flexible and the vibrational frequencies for the fundamentals were determined for the instantaneous solvent configurations from solutions of the 3-d Schrödinger equation whereas the stationary states for CN^- were determined from the analytical formula for bound states of a Morse oscillator, [104] and for NMA in water the chromophore was frozen during the MD

	$\langle\omega\rangle$	FWHM	a_1	a_2	a_3	τ_1	τ_2	τ_3	γ	Δ_0
INM	2049.9	13.6	7.01	0.76	0.16	0.044	0.23	1.18	30.14	0.003
			(0.09)	(0.06)	(0.06)	(0.00)	(0.03)	(0.33)	(0.35)	(0.00)
3-d SE	2047.3	18.5	20.37	1.40	0.57	0.048	0.21	1.06	33.28	0.001
			(0.09)	(0.06)	(0.07)	(0.00)	(0.02)	(0.09)	(0.11)	(0.00)
Expt.	2047.5	25.2						~ 1.2 (H_2O)		
								1.3 (D_2O)		

Table 4.3: Parameters obtained from fitting the FFCF to Eq. 4.6 for both, frequencies from INM and solutions of the 3-dimensional Schrödinger equation (SE). Average frequency $\langle\omega\rangle$ of the [0,0,1] fundamental in cm^{-1}, the full width at half maximum in cm^{-1}, the amplitudes a_1 to a_3 in ps^{-2}, the decay times τ_1 to τ_3 in ps, the parameter γ in ps^{-1}, and the offset Δ_0 in ps^{-2}. The experimental results in H_2O and in reverse micelles are from Refs.[151,160] and for the decay time in D_2O.[139] Errors, determined from R,[174] for all parameters are given in brackets.

simulations[105] which was also the approach followed for azide in water.[28]

Figure 4.8 shows the 1D-IR spectra of the NN asymmetric stretching mode using two different approaches (INM and DVR3D) to compute the frequencies. The observed and computed solvent induced shifts are 45 cm^{-1} and 74 cm^{-1}, respectively, for INM and DVR3D calculations, compared with 61 cm^{-1} from experiment[151,158,159], see Table 4.2. They both correctly capture the blue shift but the quantum calculations appear to better describe the 1D-IR spectroscopy. A similar observation is made for the full widths at half maximum (FWHM). Frequencies from instantaneous normal modes yield a FWHM of 13.6 cm^{-1} which increases to 18.5 cm^{-1} when the 3D Schrödinger equation is solved. This compares with 25.2 cm^{-1} from experiment.[151] Hence, the added computational effort in using quantum mechanical frequencies from scanning the 3-dimensional PES for a given solvent configuration indeed provides better results. On the other hand, the considerably more economical approach based on instantaneous normal modes is still a meaningful alternative.

4.5 Conclusions

This work introduces an accurate model to characterize the vibrational spectroscopy of N_3^- in solution. Based on a RKHS representation of the intramolecular interactions, the quantum bound states agree well for the fundamentals in the gas phase and in solution as determined from experiment. In particular, for the simulations in water the blue shifts, full widths at half maximum and the time scale of the solvent reorganization dynamics as probed by the asymmetric stretch

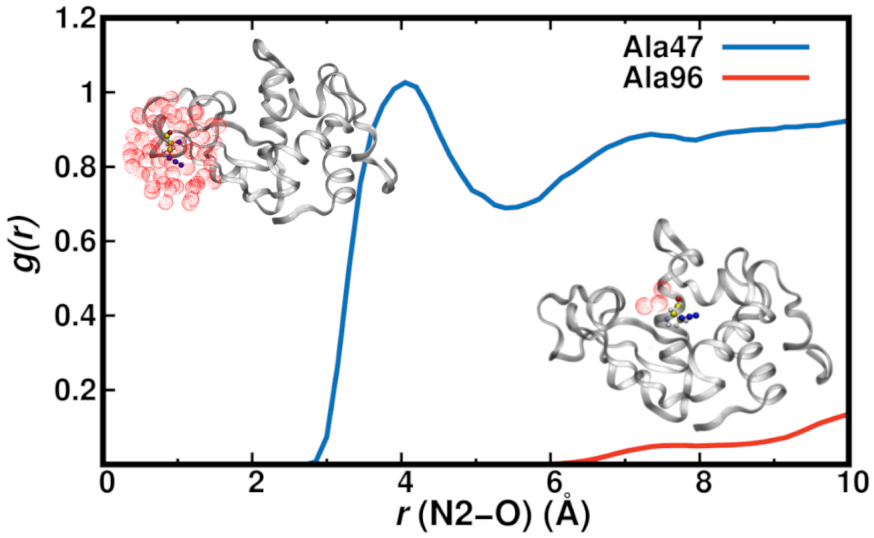
of N_3^- compare favourably with experiment.

This opens up the possibility to use this model for a spatially sensitive probe of the solvent dynamics in more complex systems, including individual molecules or proteins in solution. Independently, using a suitable flexible, spectroscopically accurate water model such as the Kumagai, Katwamura, Yokokawa (KKY) model[175–177] it will be possible to probe the coupling between the intramolecular degrees of freedom of solute and solvent as their fundamentals and overtones lead to potentially interesting dynamics. The present model can be further improved by using higher order multipoles[98,99,178,179] or polarization[180] to treat the electrostatic interactions. Also, a slight adjustment of the bonded interactions could be envisaged as using stretch and bending potentials from gas phase calculations for simulations in solution is another approximation that is used in the present work. This will be of particular importance when the present model is used for N_3^- as a covalently linked probe to a peptide or a protein residue.

Together with state-of-the art experiments the present work lays the foundation to a molecularly refined picture of the structural dynamics of complex systems in the condensed phase from a combined experimental/simulation approach.

Chapter 5

Site-Selective Dynamics of Azidolysozyme



The results presented in this chapter have been previously published:
J. Chem. Phys. 2021, 154, 165101.
doi:10.1063/5.0047330

***

5.1 abstract

The spectroscopic response of and structural dynamics around all azido-modified alanine residues ($AlaN_3$) in lysozyme is characterized. It is found that $AlaN_3$ is a positionally sensitive probe for the local dynamics, covering a frequency range of ~ 15 cm^{-1} for the center frequency of the line shape. This is consistent with findings from selective replacements of amino acids in PDZ2 which reported a frequency span of ~ 10 cm^{-1} for replacements of Val, Ala, or Glu by azidohomoalanine (AHA). For the frequency fluctuation correlation functions (FFCFs) the long-time decay constants τ_2 range from ~ 1 to ~ 10 ps which compares with experimentally measured correlation times of 3 ps. Attaching azide to alanine residues can yield dynamics that decays to zero on the few ps time scale (i.e. static component $\Delta_0 \sim 0$ ps^{-1}) or to a remaining, static contribution of ~ 0.5 ps^{-1} (corresponding to 2.5 cm^{-1}), depending on the local environment on the 10 ps time scale. The magnitude of the static component correlates qualitatively with the degree of hydration of the spectroscopic probe. Although attaching azide to alanine residues is found to be structurally minimally invasive with respect to the overall protein structure, analysis of the local hydrophobicity indicates that the hydration around the modification site differs for modified and unmodified alanine residues, respectively.

5.2 Introduction

Understanding the structural and functional dynamics of proteins in the condensed phase is a prerequisite for characterizing cellular processes at a molecular level.[181] As an example, knowledge of the mechanisms and physical principles underlying protein-ligand recognition facilitates rational drug design for treatment of diseases.[3,4] One possibility to directly and quantitatively probe the structure and dynamics of proteins and protein-ligand complexes is vibrational, in particular 2-dimensional infrared (2D-IR) spectroscopy.[128]

Given the spectroscopic response of proteins in solution that cover the range up to ~ 1700 cm^{-1} and frequencies above ~ 2800 cm^{-1}, suitable vibrational labels should absorb in the window between ~ 1700 and ~ 2800 cm^{-1}.[138,182] A range of such probes has been proposed and considered in the past, including cyanophenylalanine[183], nitrile-derivatized amino acids,[25] the sulfhydryl band of cysteines,[19]

deuterated carbons,[20] non-natural labels consisting of metal-tricarbonyl modified with a -$(CH_2)_n$- linker,[21] nitrile labels,[184] cyano[185] and SCN[17] groups, or cyanamide.[18] For the specific case of lysozyme dynamics, labeling hen egg white and human lysozyme with a ruthenium carbonyl complex successfully demonstrated to provide deeper understanding of the water dynamics from 2D-IR experiments.[186–188] Remarkably, it has been possible to demonstrate dynamic hydration and effects of collective protein hydration extending over distances 20 Å away from the protein surface, consistent with recent simulations on hydrated hemoglobin.[189,190] Another promising and sensitive label that was recently used is azidohomoalanine (AHA)[23] for which it has been demonstrated that it can be used to characterize the recognition site between the PDZ2 domain and its binding partner to provide site-specific insight into the underlying mechanisms of how signaling proteins function.[31]

The noncanonical amino acid AHA absorbs around ~ 2100 cm^{-1} with a comparatively large extinction coefficient of up to 400 $M^{-1}cm^{-1}$.[23] From a preparative perspective attachment of $–N_3$ to alanine (to give $AlaN_3$) and AHA and incorporation at almost any position of a protein through known expression techniques has been demonstrated.[191] Furthermore, attachment of an $–N_3$ probe is a spatially small modification and the chemical perturbations induced are expected to be small. This makes $AlaN_3$ and AHA worthwhile modifications to probe local protein dynamics.

Optical spectroscopy, and especially 2-dimensional infrared (2D-IR) spectroscopy, quantitatively provides information about the structure and dynamics of the solvent environment surrounding a probe molecule.[138] Such techniques can also be used to measure the subpicosecond to picosecond dynamics in condensed-phase systems. With that, the coupling between inter- and intramolecular degrees of freedom such as the hydrogen bonding network in solution, or structural features of biological macromolecules can be investigated by monitoring the fluctuation of fundamental vibrational frequencies of a probe molecule or ligand attached to a complex or a biological macromolecule. The possibility to use infrared spectroscopy for characterizing protein-ligand complexes has already been proposed for the nitrile containing inhibitor IDD743 complexed with WT and mutant human aldose reductase[96] and explicitly demonstrated for cyano-benzene in the

active site of WT and mutant lysozyme.[97]

The AHA label was previously used in 2D-IR spectroscopy studies of ligand binding to the PDZ2 domain.[23] The spectral changes observed for various modified peptidic binders were consistent with the known X-ray structure of the wild-type peptide bound to the protein. This suggests that AHA is suitable as a specific IR reporter and to highlight subtle changes of the electrostatic environment on the protein surface.[31] In the present work, attaching $-N_3$ to all alanine residues in lysozyme in succession is used to characterize the local dynamics around such modification site.

Recent investigations have demonstrated that the vibrational dynamics of N_3^- in the gas phase and in solution can be captured quantitatively.[106] Based on high-level electronic structure calculations at the multi-reference configuration interaction (MRCI) level of theory and representing the 3-dimensional potential energy surface (PES) as a reproducing kernel Hilbert space (RKHS),[73,74] the infrared spectroscopy in the gas and condensed phase was correctly described. Also, the frequency correlation function exhibited time scales consistent with experiment which suggests that the coupling between solvent and solute was correctly described.

The present work explores the local dynamics of all alanine residues in lysozyme as a typical model system by attaching $-N_3$ as a spectroscopic reporter. First, the computational methods are summarized. Then, the structural dynamics and spectroscopy for all 14 $AlaN_3$ labels is discussed and the local dynamics and hydration are explored. Finally, conclusions are drawn.

5.3 Methods

5.3.1 Molecular Dynamics Simulations

For the Molecular Dynamics (MD) simulations of WT and modified lysozyme in solution, CHARMM[64] together with the CHARMM[67] force field was used. A suitably modified version of CHARMM was employed for the simulations with the 3-dimensional RKHS PESs (see below).[106] The initial structure was the X-ray

structure of WT human lysozyme (3FE0[192]). Simulations of lysozyme in TIP3P water[117] were carried out in a cubic box of size $(62.1)^3$ Å^3. Figure 5.1 shows the structure of the system for the present work in which $-N_3$ is attached individually to each of the 14 Ala residues, replacing one hydrogen atom of the terminal CH_3 group. This yields azidoalanine-modified lysozyme.[191] Compared with protein structures in which AHA is introduced, the two modifications differ by one CH_2-group.[191].

The systems were minimized, heated for 25 ps and equilibrated for 100 ps in the NVT ensemble. Production runs, 2 ns in length, were carried out in the NVT ensemble, with coordinates saved every 5 fs for subsequent analysis. All nonbonded interactions were treated with a 14 Å cutoff switched at 10 Å,[124] and bonds involving hydrogen atoms are constrained using SHAKE[123].

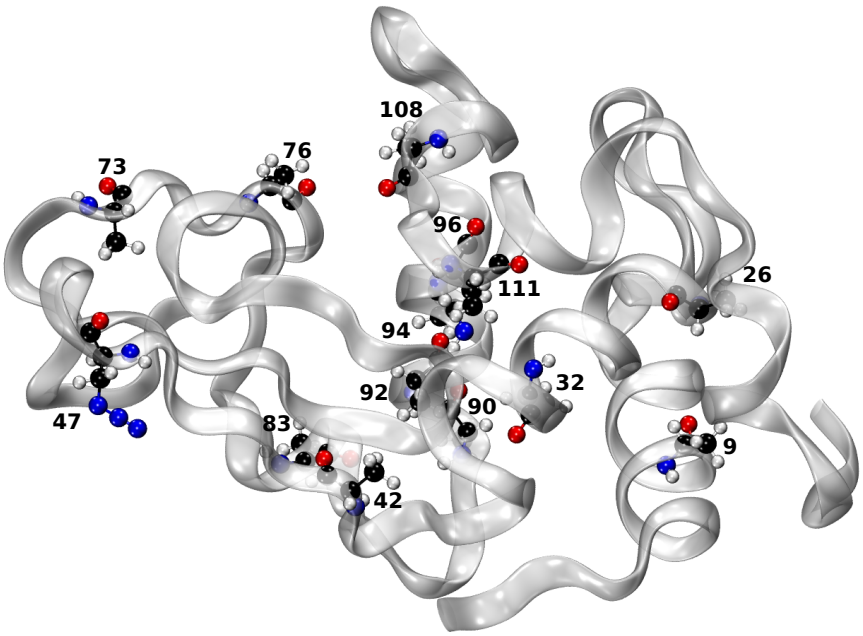


Figure 5.1: Structure of lysozyme with positions of Alanine residues indicated. The Alanine residues are at positions 9, 26, 32, 42, 47, 73, 76, 83, 90, 92, 94, 96, 108, 111. Ala residues are displayed as CPK spheres and the rest of the protein structure is shown as NewRibbons. As an example, $AlaN_3$ is shown at residue 47.

5.3.2 Energy Function for the Spectroscopic Probe

For representing the 3-dimensional energy function of the $-N_3$ label two strategies were pursued. First, the existing 3-dimensional PES for N_3^-, computed at

the MRCI+Q level of theory in the gas phase, was used to describe the stretching and bending distortions of the label attached to the CH_2 group of alanine.

Because the $-N_3$ moiety and the rest of the Ala residue are not fully electronically decoupled, a second approach was pursued. For this, the structure of AHA was optimized at the MP2/aug-cc-pVTZ level of theory. Next, the structure of AHA was frozen except for the coordinates involving the spectroscopic label. Then, a new 3-dimensional PES was computed at the pair natural orbital based coupled cluster level (PNO-LCCSD(T)-F12)[193,194] together with the aug-cc-pVTZ basis set[195] using the MOLPRO suite of codes.[164] As for the gas phase PES,[106] the *ab initio* energies were calculated in Jacobi coordinates (R, r, θ), see Figure 5.2B, where r is the distance between the nitrogen atoms N1 and N2, R is the distance between their center of mass and the atom N3, and θ is the angle between $\vec{r}$ and $\vec{R}$. The angular grid (θ) used here contains 5 Gauss-Legendre quadrature points between 156° and 180°. The radial grids include 16 points along r ranging from 0.90 to 1.51 Å and 16 points along R between 1.45 and 2.12 Å. The PNO-LCCSD(T)-F12 level of theory was chosen as it combines accuracy with feasibility for the present problem because recomputing the MRCI PES for AHA is computationally intractable.

For both PESs the parameters for the C-N3 stretch, the C-C-N3 and the C-N3-N2 bend are those from Swissparam.[196] All remaining parameters for the alanine residues were those of the CHARMM force field and were not readjusted after attaching $-N_3$ to guarantee compatibility with the CHARMM22 force field.

To carry out MD simulations for labelled lysozyme, a continuous and differentiable representation of the *ab initio* energies is required. For this, a reproducing kernel Hilbert space-based representation[73,74] is used. A RKHS representation provides approximate values for a function $f(x)$ at positions x, away from the grid points x_i. For this, the linear problem $f(x_i) = \sum_j \alpha_j k(x_i, x_j)$ for the 1-dimensional kernels is solved which yields the coefficients α_j. There are many possible choices for the kernel functions $k(\cdot, \cdot)$ but inverse powers of the distance have been found to perform well for intermolecular interactions.[73,165,197] For multi-

dimensional problems, tensor products of 1-dimensional kernels can be used.[74,166]

For the present work, the 3-dimensional kernel K is

$$K(X, X') = k^{(n,m)}(R, R')k^{(n,m)}(r, r')k^{(2)}(z, z'). \tag{5.1}$$

where X stands for all dimensions involved, r, and R are as defined above (see also Figure 5.2), and $z = \frac{1-\cos(\theta)}{2}$ maps the angle θ onto the interval $[0, 1]$. Reciprocal power decay kernels ($k^{(n,m)}$) with smoothness $n = 2$ and asymptotic decay $m = 6$

$$k^{(2,6)}(x, x') = \frac{1}{14}\frac{1}{x_>^7} - \frac{1}{18}\frac{x_<}{x_>^8}, \tag{5.2}$$

are used for r and R whereby $x_>$ and $x_<$ are the larger and smaller values of x and x', respectively. For the angular degree of freedom, a Taylor spline kernel

$$k^{(2)}(z, z') = 1 + z_< z_> + 2z_<^2 z_> - \frac{2}{3}z_<^3 \tag{5.3}$$

is used.

Charges were calculated for the optimized structure of $AlaN_3$ at the MP2/aug-cc-pVTZ level of theory from an NBO[198] analysis using Gaussian[199]and scaled to maintain overall neutrality. This yields a charge of $-0.2460e$ for the nitrogen atom N1 attached to CH_2 group, $0.1607e$ for the central N2 and $-0.0464e$ for the terminal nitrogen N3.

5.3.3 Frequency Fluctuation Correlation Function and Lineshape

From each production simulation, 4×10^5 snapshots are taken as a time-ordered series for computing the frequency fluctuation correlation function (FFCF) $\langle\delta\omega(0)\delta\omega(t)\rangle$ and line shapes. Here, $\delta\omega(t) = \omega(t) - \langle\omega(t)\rangle$ and $\langle\omega(t)\rangle$ is the ensemble average of the transition frequency. The FFCF was determined from instantaneous harmonic vibrational frequencies based on a normal mode analysis.[200] Normal modes were determined for each snapshot after minimizing the structure of the $-N_3$ label and keeping the surrounding solvent frozen. Thus, frequency trajectories $\omega_i(t)$

for label i were obtained for the asymmetric stretch vibration of $-N_3$ attached to Ala. From the FFCF the line shape function

$$g(t) = \int_0^t \int_0^{\tau'} \langle \delta\omega(\tau'')\delta\omega(0)\rangle d\tau'' d\tau'. \tag{5.4}$$

is determined within the cumulant approximation. To compute $g(t)$, the FFCF is numerically integrated using the trapezoidal rule and the 1D-IR spectra is then calculated as[151]

$$I(\omega) = 2\Re \int_0^\infty e^{i(\omega-\langle\omega\rangle)t} e^{-g(t)} e^{-\frac{t}{2T_1}} e^{-2D_{\mathrm{OR}}t} dt, \tag{5.5}$$

where $\langle\omega\rangle$ is the average transition frequency obtained from the distribution, T_1 (0.8 ± 0.1 ps) is the vibrational relaxation time and $D_{\mathrm{OR}} = 1/6T_R$ with $T_R = 1.3 \pm 0.3$ ps is the rotational diffusion coefficient which accounts for lifetime broadening.[150]

From the FFCF, the decay time can be determined by fitting the FFCF to a general expression[129]

$$\langle \delta\omega(t)\delta\omega(0)\rangle = a_1 \cos(\gamma t) e^{-t/\tau_1} + \sum_{i=2}^{n} a_i e^{-t/\tau_i} + \Delta_0 \tag{5.6}$$

where a_i, τ_i, γ and Δ_0 are fitting parameters. The decay times τ_i of the frequency fluctuation correlation function reflect the characteristic time-scales of the solvent fluctuations to which the solute degrees of freedom are coupled. In all cases the FFCFs were fitted to an expression containing two decay times (i.e. $n_{\max} = 2$) using an automated curve fitting tool from the SciPy library.[130]

5.4 Results

5.4.1 The Potential Energy Surface for the $-N_3$ Label

Two PESs for the energetics of the azide probe are considered in the present work. One is based on earlier MRCI+Q calculations with the aug-cc-pVTZ basis set for N_3^- in the gas phase[106] which was used without change for the simulation of the $AlaN_3$ unit. The second one was the LCCSD(T) PES for AHA which included coupling between the $-N_3$ probe and the amino acid framework. The RKHS rep-

resentations of the two PESs are reported in Figures 5.2A and B and the scans within CHARMM are shown in panels C and D.

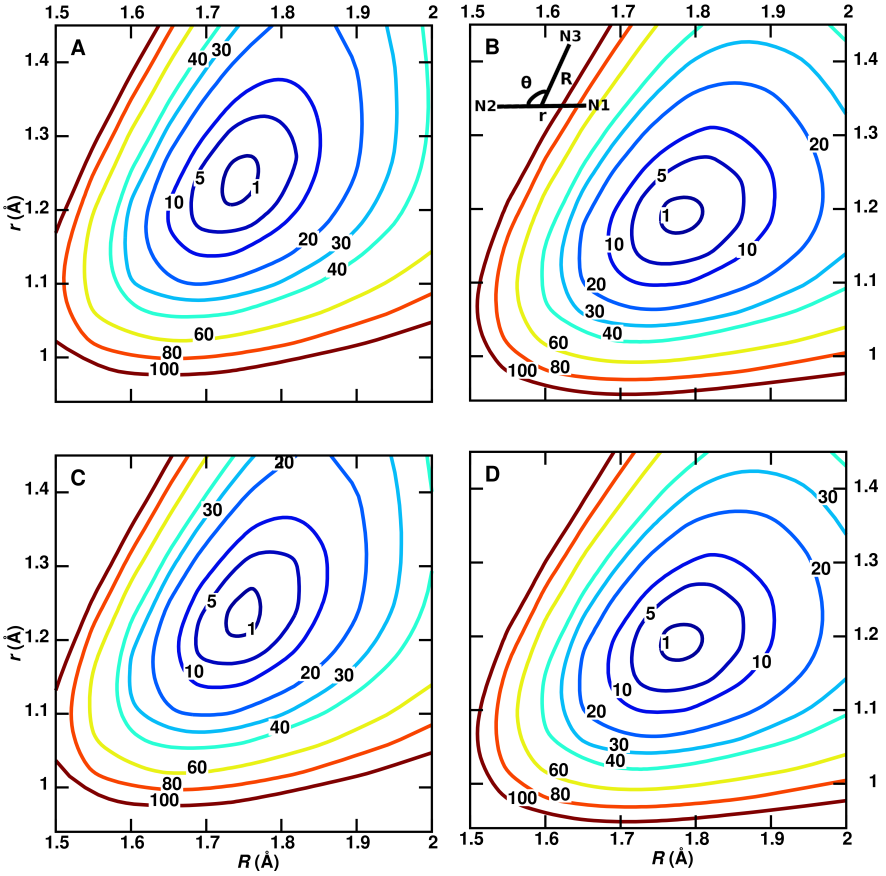


Figure 5.2: Contour diagrams of the RKHS representations for AHA (panel A, PNO-LCCSD(T)-F12) and N_3^- (panel B, MRCI+Q/aug-cc-pVQZ) PESs based on *ab initio* points calculated in Jacobi coordinates (R, r, θ) for $\theta = 176.225°$, see inset in panel B. Panels C and D report the corresponding CHARMM energies for AHA. All energies are in kcal/mol and relative to the zero of energy which is the minimum energy structure.

The RKHS representation of the PES for AHA was constructed from 1280 *ab initio* LCCSD(T)-F12 energies. An additional 230 *ab initio* energies are calculated at off-grid geometries to assess the quality of the RKHS representation. Figure 5.3 shows the correlation between the reference energies and the RKHS with a correlation coefficient of $R^2 = 0.9999$ and the root mean squared error is 0.38 kcal/mol. This confirms the high quality of the RKHS PES.

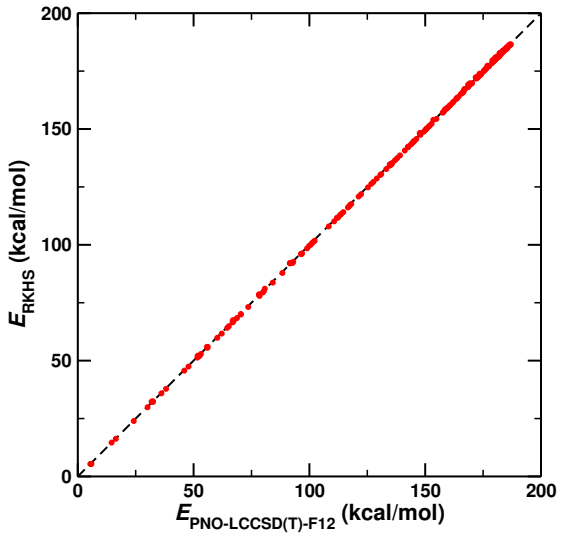


Figure 5.3: Correlation between *ab initio* and RKHS interpolation for 230 randomly selected geometries. The $R^2 = 0.9999$ and the RMSD is 0.38 kcal/mol over a range of 180 kcal/mol.

Figure 5.2A and B report the RKHS interpolation of the *ab initio* calculated energies whereas Figures 5.2C and D are from scanning the r and R coordinates for AHA in the gas phase in CHARMM. Comparing the PNO-LCCSD(T)-F12 PES (Figure 5.2A) with that at the MRCI+Q level of theory (Figure 5.2B) shows that the minima for the two are slightly displaced ($r = 1.19$ Å vs. $r = 1.24$ Å and $R = 1.77$ Å vs. $R = 1.76$ Å). Furthermore, the LCCSD(T) PES is steeper along both, the r and R coordinates, which pushes the respective vibrations up compared with the MRCI+Q PES, see Figure 5.4. Differences between the two PESs are due to both, the methods (MRCI+Q vs. PNO-LCCSD(T)-F12) and the model system (N_3^- vs. AHA) considered. Comparing the isolated, gas-phase PESs (panels A and B) with those for $AlaN_3$ (panels C and D) indicates that the PESs are close but not identical due to coupling between the spectroscopic probe and the alanine residue.

In the following, all MD simulations were carried out with the PNO-LCCSD(T)-F12 PES as it yields harmonic frequencies for $AlaN_3$ around 2110 cm^{-1} cm^{-1} (see Table 5.1) which is consistent with those experimentally observed for the replacement of AHA[31] in PDZ2 domain at 2114 cm^{-1} and for $AlaN_3$[201] in H_2O at 2116 cm^{-1}, respectively. Moreover, the influence of the covalent bonding to the Alanine residue is included in the construction of the potential energy surface. Additional refinements of the PES would, in principle, be possible through mor-

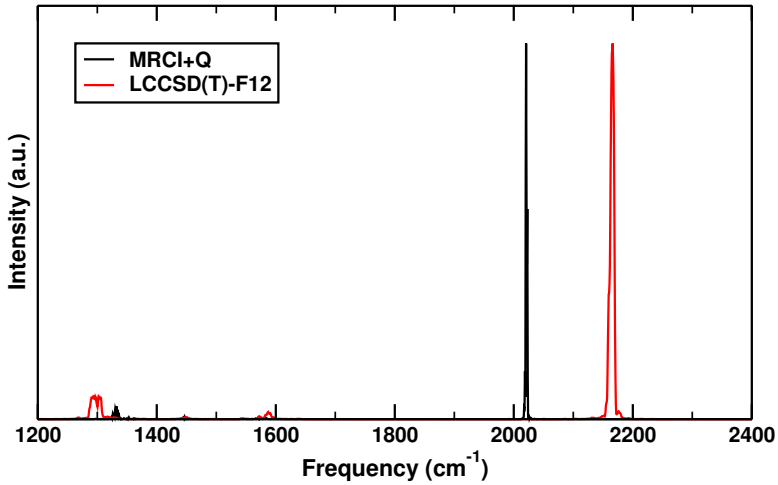


Figure 5.4: Power spectrum based on the N1-N2 separation for AHA in the gas phase and from MRCI+Q and LCCSD(T)-F12 surface.

phing[202,203] but were not deemed necessary for the present work which is mainly concerned with the differential dynamics, i.e. the relative positional sensitivity, and spectroscopy for the same label at different positions along the polypeptide chain.

5.4.2 Structural Dynamics

For the structural dynamics first the root mean squared deviation (RMSD) of unmodified and modified lysozyme in solution compared with the starting X-ray structure as the reference is analyzed. For this, the RMSD of all C_α atoms was considered. Figure 5.5 shows the RMSD for all C_α atoms (blue) and those for the 14 Alanine residues (red) specifically from the 2 ns simulation of the modified protein at position Ala47. For the WT protein similar RMSD values are reported in Figure 5.6. The RMSD values fluctuate below or around 1 Å which is indicative of a stable simulation. This suggests that attaching a $-N_3$ label to Ala has an insignificant effect on the structural dynamics of lysozyme, consistent with earlier findings for the PDZ domain for which also a minimally invasive effect was reported.[31]

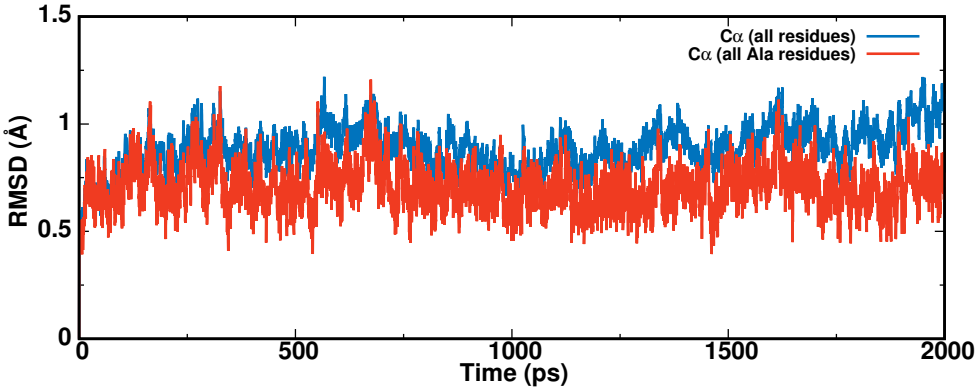


Figure 5.5: The structural RMSD for the C_α atoms from all residues (blue) and for the 14 Ala residues (red) specifically for Ala47N_3.

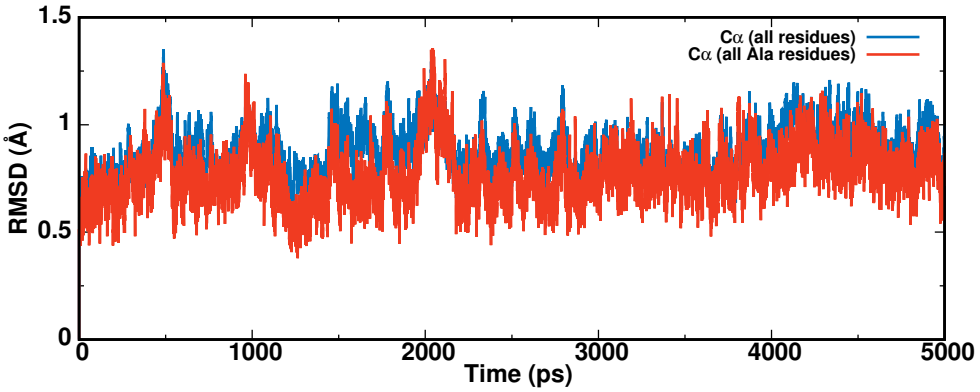


Figure 5.6: The structural RMSD for the C_α atoms from all residues (blue) and for the 14 Ala residues (red) for WT lysozyme.

5.4.3 Vibrational Spectra and Frequency Correlation Functions

First, the power spectra and frequency trajectories for the asymmetric stretch of the azide label attached to all 14 alanine residues are presented. The power spectra as determined from the Fourier transform of the N2-N3 distance correlation function are shown in Figure 5.7A for all AlaN_3 from 2 ns production runs. The peak maxima $\omega_{\max}$ cover a range of $\sim$ 20 cm^{-1} (between 2160 and 2180 cm^{-1}) and the full widths at half maximum (fwhm) of the spectra are around 20 cm^{-1}. Hence, although the same energy function was used for all modified AlaN_3 moieties, their power spectra differ depending on the position of the modified Ala residue along the polypeptide chain.

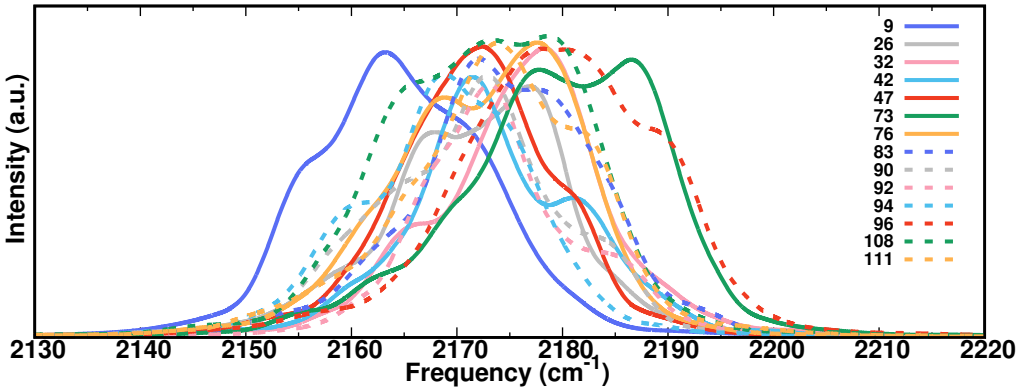


Figure 5.7: Power spectrum based on the N2–N3 separation for all modified $AlaN_3$ residues. The position of the frequency maxima differ for most of the $AlaN_3$ labels and cover a range between 2160 and 2180 cm^{-1}.

The power spectra reported in Figure 5.7 are also representative of the infrared spectrum as shown in Figure 5.8. The top panel of Figure 5.8 reports the power spectrum and peak positions of all three modes for $Ala47N_3$ with the asymmetric stretch centered around 2170 cm^{-1}, the symmetric stretch at 1333 cm^{-1} and the bending mode at 610 cm^{-1}. The bottom panel of Figure 5.8 demonstrates that the infrared spectrum (IR) determined from the dipole autocorrelation function supports the peak positions found from the power spectrum to within 2 cm^{-1}.

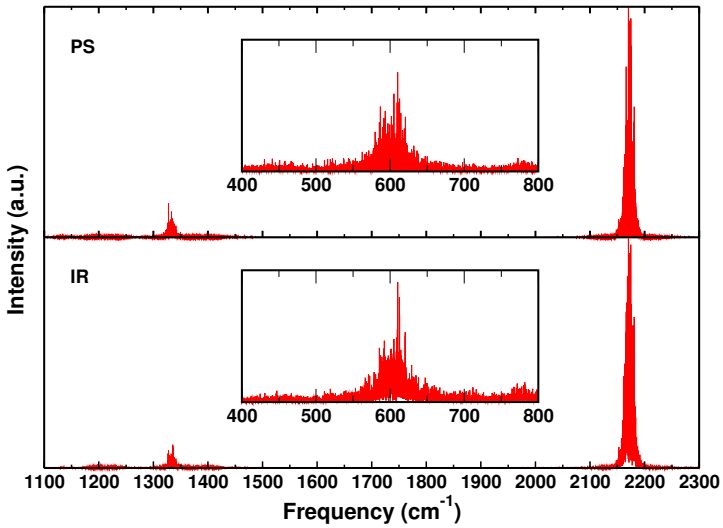


Figure 5.8: Power (PS, top panel) and IR (bottom panel) spectrum for $Ala47N_3$. The power spectrum is based on the N2-N3 bond displacement. The inset shows the bending mode of the azide group. IR spectrum is calculated the Fourier transform of the molecular dipole moment autocorrelation function.

Next, the frequency trajectories $\omega_i(t)$ for each of the spectroscopic probes i from 4×10^5 snapshots were determined from instantaneous normal mode calculations. From the frequency time series the frequency fluctuation correlation functions (FFCFs) are obtained. They contain valuable information about the environmental dynamics around each site i, i.e. the azide probes of the various Ala residues considered.

The FFCFs, shown in Figure 5.9, are fitted to Eq. 5.6 with a parametrization motivated by the overall shape of the FFCF.[129] This functional form has also been used in previous work.[28,104,129] It is an extension of the typical multiexponential decay, which is traditionally employed[204] to capture an anticorrelation at short times ($t < 1$ ps). Figure 5.9 provides a comparison between the raw data (black) and the fits (red) and Table 5.1 reports the corresponding fitting parameters.

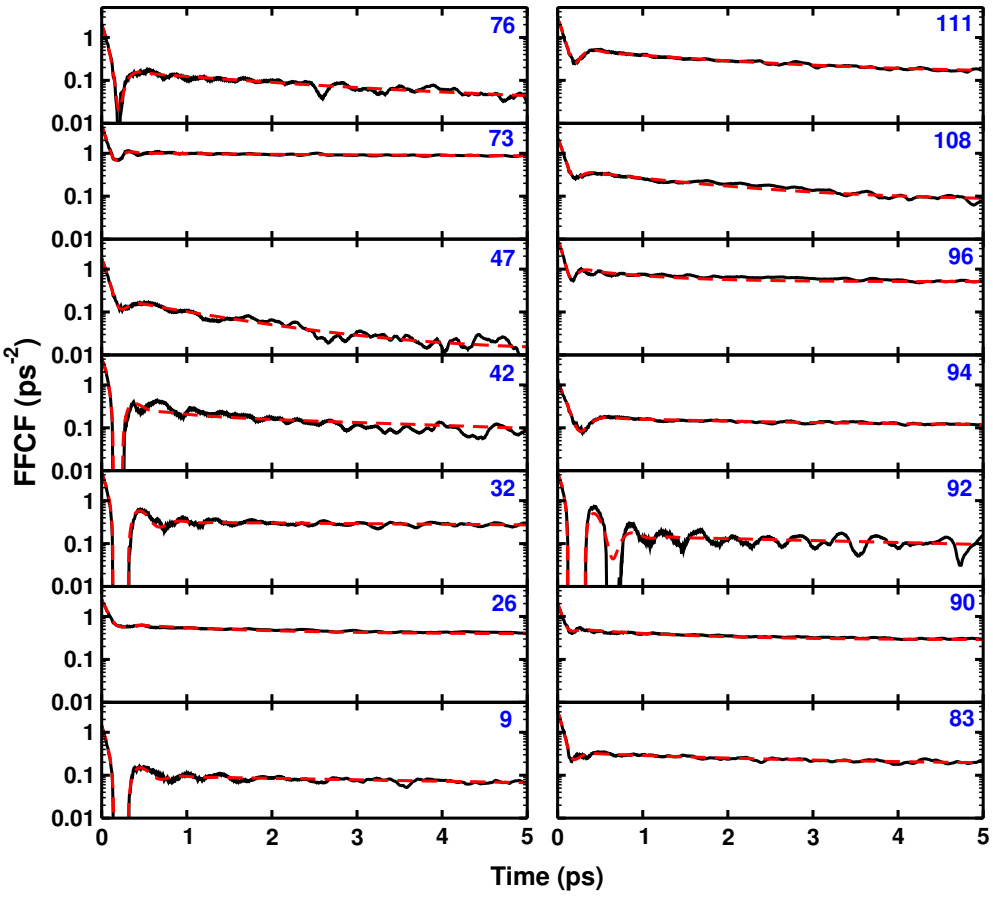


Figure 5.9: FFCFs from correlating the instantaneous harmonic frequencies for all 14 $AlaN_3$ in lysozyme. The labels in each panel refer to the alanine residue which carries the azide label. Black traces are the raw data and red dashed lines the fits to Eq. 5.6. The $y-$axis is logarithmic.

The shape of the FFCFs can differ appreciably. Some of them display a pronounced minimum at short correlation times ($t \sim 0.1$ ps) whereas others do not. This feature has also been found in previous simulations[28] and has been related to the strength of the interaction between the infrared probe and its environment.[104,105,129] Several of the FFCFs show one (Ala9, Ala32, Ala42, Ala76, Ala94, Ala96, Ala108, Ala111) or even two (Ala92) recurrences at short correlation times. For the remaining Alanine residues this feature is less pronounced (Ala47, Ala73, Ala83, Ala90) or entirely absent (Ala26). Similarly, some of the FFCFs exhibit clear static components $\Delta_0 \simeq 0.5$ ps^{-2} (Ala26, Ala73, Ala96) whereas the remaining ones decay to zero on the ~ 10 ps time scale. With respect to the correlation times, the fast correlation is generally $\tau_1 \sim 0.1$ ps whereas the long time scale ranges from $\tau_2 = 1.1$ ps to $\tau_2 < 13$ ps, see Table 5.1. Typically, the amplitude of the fast decay is one order of magnitude larger than that of the slow contribution (Table 5.1). Hence, the characteristics of the FFCFs vary considerably depending on the position at which the Alanine residue is located along the polypeptide chain. This suggests that $AlaN_3$ is a positionally sensitive probe to provide quantitative information about the local dynamics of a protein.

Res	$\langle\omega\rangle$	a_1	γ	τ_1	a_2	τ_2	Δ_0	LH (WT)	LH (–N_3)
9	2103.7	1.17	13.16	0.129	0.07	7.62	0.02	0.33	0.31
26	2107.9	1.82	8.69	0.079	0.24	1.73	0.41	1.92	1.71
32	2112.8	3.60	13.39	0.164	0.13	13.05	0.19	0.12	0.09
42	2111.4	3.57	14.08	0.104	0.32	2.15	0.04	0.96	1.36
47	2107.5	1.54	9.99	0.081	0.21	1.17	0.01	1.22	1.70
73	2114.6	3.18	15.07	0.080	0.24	3.61	0.81	1.34	0.20
76	2107.1	1.94	11.94	0.084	0.15	2.77	0.02	0.41	-0.02
83	2109.2	2.75	11.48	0.069	0.17	2.11	0.18	1.60	1.34
90	2107.5	1.57	11.71	0.056	0.22	1.42	0.30	1.14	1.51
92	2110.4	4.08	14.02	0.184	0.10	1.33	0.10	0.58	0.17
94	2104.8	1.05	9.36	0.100	0.08	5.89	0.09	1.17	1.21
96	2116.3	3.74	13.33	0.077	0.43	2.19	0.48	0.45	0.37
108	2108.4	1.83	10.77	0.084	0.35	2.43	0.04	0.55	0.41
111	2110.6	2.22	13.13	0.090	0.41	2.27	0.12	1.50	1.67

Table 5.1: Parameters obtained from fitting the FFCF to Eq. 5.6 for INM frequencies for all different $AlaN_3$ residues in lysozyme. Average frequency $\langle\omega\rangle$ of the asymmetric stretch in cm^{-1}, the amplitudes a_1 to a_3 in ps^{-2}, the decay times τ_1 to τ_3 in ps, the parameter γ in ps^{-1}, the offset Δ_0 in ps^{-2}, and the conformationally averaged local hydrophobicity (LH) for the WT and each modified protein.

Numerical integration of $g(t)$ and using Eq. 5.6 yields the 1-dimensional IR spectra for each label based on instantaneous normal modes, see Figure 5.10. Similar

to the power spectra, the center frequencies cover a range of $\sim$ 15 cm^{-1}, with center frequencies of 2104 cm^{-1} for Ala9N$_3$ and 2116 cm^{-1} for Ala96N$_3$, and the fwhm ranges from 13 to 21 cm^{-1}. Also, the $\omega_{\max}$ for Ala9N$_3$ (blue solid line in Figures 5.7 and 5.10) is lowest in frequency and those for Ala96N$_3$ (dashed red) and Ala73N$_3$ (solid green) are highest from the power spectra and the INM lineshapes, respectively. The blue shift of the power spectra compared with those from INM for the symmetric and asymmetric stretch modes was already found for N_3^- in solution.[106] The magnitude of this shift is larger in the present case probably due to coupling between the spectroscopy probe and the amino acid it is attached to.

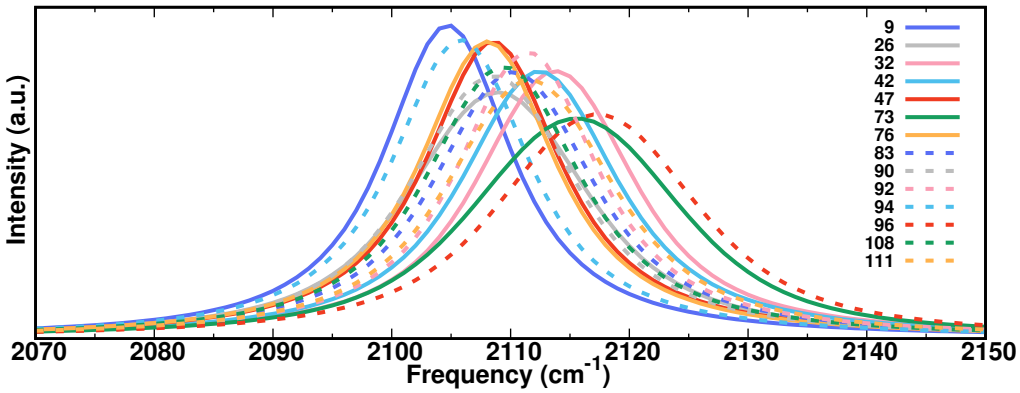


Figure 5.10: 1D-IR spectra for all 14 AlaN$_3$ residues in lysozyme. For the IR lineshape the raw FFCF from the INM analysis was numerically integrated to give $g(t)$ from which the 1D lineshape is obtained.

An alternative to instantaneous normal modes is to obtain instantaneous frequencies from solving the 1- or 3-dimensional nuclear Schrödinger equation. For this, the corresponding 1- or 3-d PES is scanned for a given snapshot with frozen environment[104,127,200] and represented as a RKHS. This is a computationally much more demanding approach, in particular in 3 spatial dimensions.[106] Here, the 1-dimensional PES along the asymmetric stretch motion was mapped out for 4×10^5 snapshots and the nuclear Schrödinger equation was solved. Then, the FFCF was again determined and fit to Eq. 5.6, see Figure 5.11. From this, the 1-dimensional IR lineshape was determined, see solid lines in Figure 5.12. This was done for Ala90N$_3$ and Ala94N$_3$. As was found for the 1-d lineshapes from INM, the frequency maximum for Ala90N$_3$ is shifted to the blue relative to Ala94N$_3$ but the shift is smaller (1 cm^{-1} vs. 3 cm^{-1}).

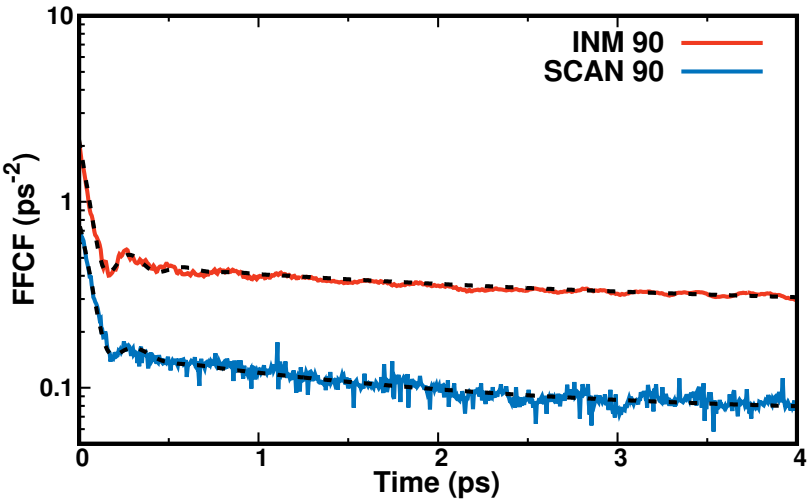


Figure 5.11: The FFCF for Ala90 based on INM (red) and scan (blue) frequencies. The dashed lines show the corresponding fit to Eq. 5.6 with 3 time scales and the fitting parameters are as follows: INM: $a_1 = 0.29$, $\gamma = 21.00$, $\tau_1 = 0.18$ ps, $a_2 = 1.38$, $\tau_2 = 0.05$ ps, $a_3 = 0.22$, $\tau_3 = 2.70$ ps, $\Delta_0 = 0.25$ and scan: $a_1 = 0.21$, $\gamma = 17.50$, $\tau_1 = 0.11$ ps, $a_2 = 0.39$, $\tau_2 = 0.08$ ps, $a_3 = 0.09$, $\tau_3 = 1.62$ ps, $\Delta_0 = 0.07$. The comparison shows that the two different ways to determine the instantaneous frequency ($\omega(t)$ and $\nu(t)$, respectively) does not affect the overall appearance of the FFCF except for the magnitude of the asymptotic value Δ_0.

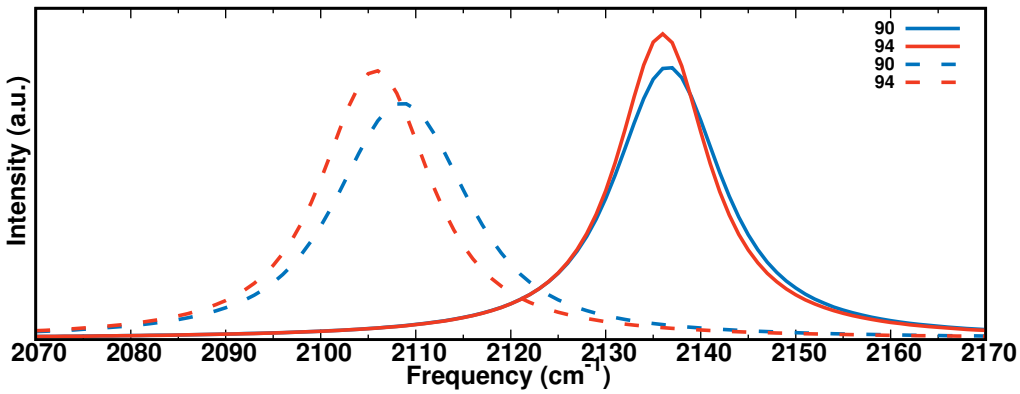


Figure 5.12: 1D-IR spectra for Ala90N_3 and Ala94N_3 of lysozyme obtained from frequency calculations using "scan" (solid lines) and INM (dashed lines). Both analyses agree in that the frequency maximum for Ala90N_3 is to the blue of that for Ala94N_3 but the magnitude of the shift differs for the two approaches.

Recently, the "INM", "scan" and "map" approaches have been compared for insulin monomer and dimer.[200] It was found that the "INM" and "scan" approaches yield comparable 1-d infrared spectra for the amide-I bands and conclusions drawn from the spectra concerning monomeric and dimeric insulin are consistent between the two. In systems with one reporter, as is also the case in the present work, and using multipolar force fields together with "scan" it was possible to

correctly assign structural substates such as the two orientations of photodissociated CO in Mb.[99,100,205–207] Similarly, from simulations of the vibrational Stark effect based on MTPs and instantaneous frequencies from "INM", the relative frequency shifts for cyano-benzene in WT and two mutants of T4-lysozyme were correctly captured.[97] Hence, "scan" and "INM" analyses together with physics-based force fields are able to describe relative frequency shifts and provide correct ordering of infrared frequencies. In systems with multiple reporters the coupling between the sites is partly included in the molecular dynamics underlying the "scan" and "INM" approaches. When evaluating the instantaneous frequencies from "scan" and "INM" it is possible to recover part of the couplings, for example from normal mode calculations of the full protein instead of freezing all but one label, or from multi-dimensional scans of the PES. For "map" the couplings are introduced as additional terms into the excitonic Hamiltonian.[27,101]

Typically, spectroscopic work has used AHA as an infrared label instead of azidoalanine as used here. To quantify the difference between $AlaN_3$ and AHA, residue Ala47 has also been replaced by AHA through inserting an additional CH_2 group before the $-N_3$ label. The parametrization of the CH_2 group is identical to that already used for alanine. Then, a 2 ns simulation for AHA in water was carried out and the IR spectrum was determined from an INM analysis, see Figure 5.13. It is found that the position of the frequency maximum for the asymmetric stretch of the azide label differs by less than 1 cm^{-1} from that with $AlaN_3$ which confirms that for IR spectroscopy, the two systems are very similar.

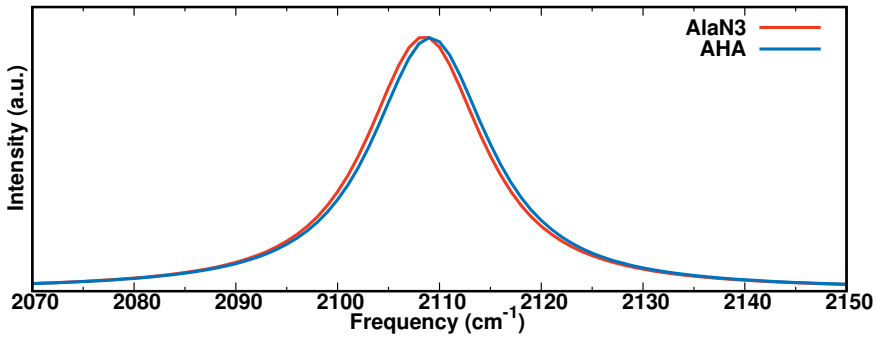


Figure 5.13: Comparison between the 1D-IR spectra of $AlaN_3$ and AHA at position Ala47 in lysozyme. For AHA, an additional CH_2 group is inserted before the $-N_3$ label. The lineshapes for $AlaN_3$ and AHA are very similar. The position of the frequency maximum for the asymmetric stretch of the azide label differs by less than 1 cm^{-1}.

Experiments for model systems to which azide was attached have reported quite complex lineshapes in the region of the azide asymmetric stretch.[29] These lineshapes have been linked to an accidental Fermi resonance (FR) between the asymmetric stretch and near resonant combination bands. By isotopic editing the nitrogen atoms in the azide probe it was demonstrated that the influence of the FR on the lineshape of the asymmetric stretch can be greatly reduced and the maximum absorption frequency shifts towards the red by different amounts, depending on the position at which the ^{15}N atom is inserted.[29] To assess the expected frequency shifts in the present case, nitrogen atoms N1 to N3 were all replaced by their heavy isotopes and independent trajectories were run for all three isotopologues. The power spectra of the asymmetric stretch vibration were determined for a qualitative characterization on the changes in the IR spectroscopy, see Figure 5.14. The maximum frequencies shift by –4 cm^{-1}, –48 cm^{-1}, and –25 cm^{-1} for the isotopologues with ^{15}N at positions N1, N2, and N3, relative to the all-^{14}N species. This compares with shifts by –2 cm^{-1}, –56 cm^{-1}, and –43 cm^{-1} from experiments on 3-azidopyridine.[29] Hence, the simulations find the correct ordering of the frequency changes upon selective ^{14}N→^{15}N replacement although the magnitudes may be somewhat system-dependent. Similar findings have been recently reported for $\beta-$azidoalanine.[208]

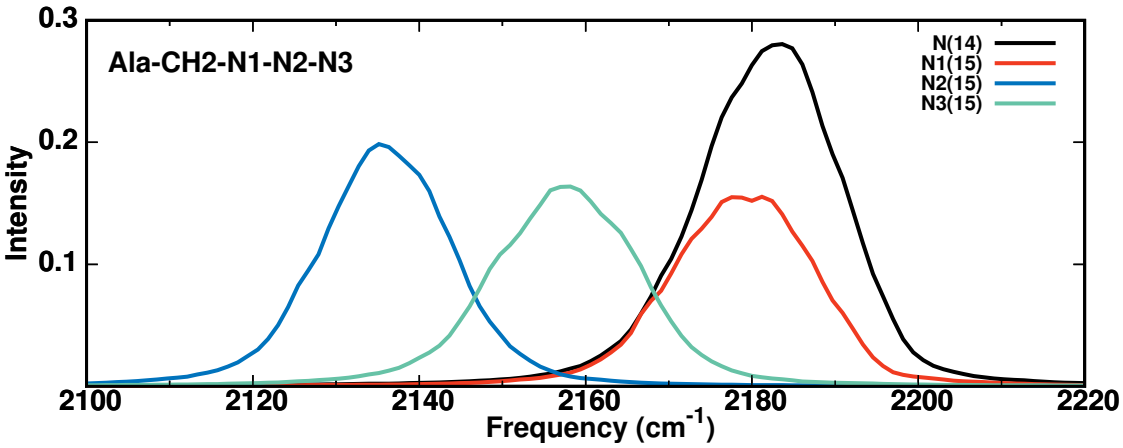


Figure 5.14: Power spectra for AHA for the all-^{14}N (black), -^{15}N1N2N3 (red), -N1^{15}N2N3 (blue), and -N1N2^{15}N3 (green) isotopologues. The positions of the frequency maxima are at 2183 cm^{-1} (all-^{14}N), 2179 cm^{-1} (-^{15}N1N2N3), 2135 cm^{-1} (-N1^{15}N2N3), and 2158 cm^{-1} (-N1N2^{15}N3), respectively.

5.5 Solvent Structure and Dynamics

Next, the solvent structuring around the modification sites is characterized. This also provides the information for an attempt to relate the spectral signatures (position of the frequency maximum, characteristics of the FFCFs) for the azide labels at different positions along the polypeptide chain with structural features and environmental properties. For this, the solvent structure around each of the 14 $AlaN_3$ probes was analyzed. First, the radial distribution functions $g(r)$ were computed along all production simulations for the 14 modified proteins, see Figure 5.15. The distance analyzed was the separation between the water-oxygen atom (O_W) and the middle nitrogen (N2) of the $-N_3$ probe in $AlaN_3$. The corresponding running coordination number $N(r)$ is

$$N(r) = 4\pi \int_0^r r^2 g(r) \rho dr$$

where ρ is the pure water density (Figure 5.15B). As is shown in Figure 5.15, the $g(r)$ and $N(r)$ differ for the 14 modification sites.

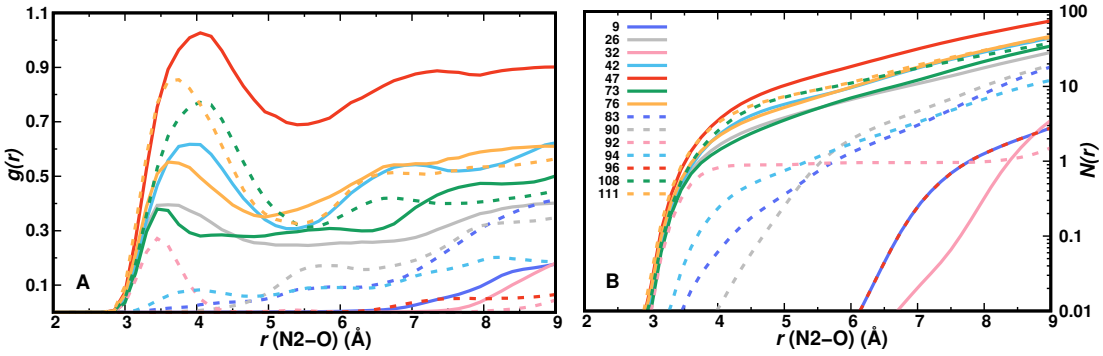


Figure 5.15: The radial distribution function $g(r)$ (panel A) and the number of water oxygen atoms $N(r)$ (panel B) between O of water and N2 of $AlaN_3$ for all alanine residues from the 2 ns production simulations. The color code for the lines is given in panel B.

For some of the residues (Ala26, Ala42, Ala47, Ala73, Ala76, Ala108, Ala111; Set1) the $g(r)$ exhibits a pronounced first maximum at $3.5 \leq r_{\mathrm{max1}} \leq 4$ Å whereas for the remaining labels (Ala9, Ala32, Ala83, Ala90, Ala92, Ala94, Ala96; Set2) such a first maximum is largely absent. This suggests that the residues in Set1 are solvent exposed whereas those in Set2 are not. The total number $N(r)$ of water molecules within a distance r supports this, see Figure 5.15B. Up to a dis-

tance of 5 Å, which is typically the extent of the first hydration shell, residues in Set1 contain 10 or more water molecules whereas those belonging to Set2 have not more than 1 water molecule in their vicinity.

A structural illustration for this observation is given in Figure 5.16 which reports all water molecules within 7 Å of Ala47N_3 (belonging to Set1) and Ala96N_3 (belonging to Set2). Consistent with Figure 5.15 only 3 water molecules are within the cutoff radius of atom N2 of Ala96N_3 whereas the hydration shell of Ala47N_3 is extensive.

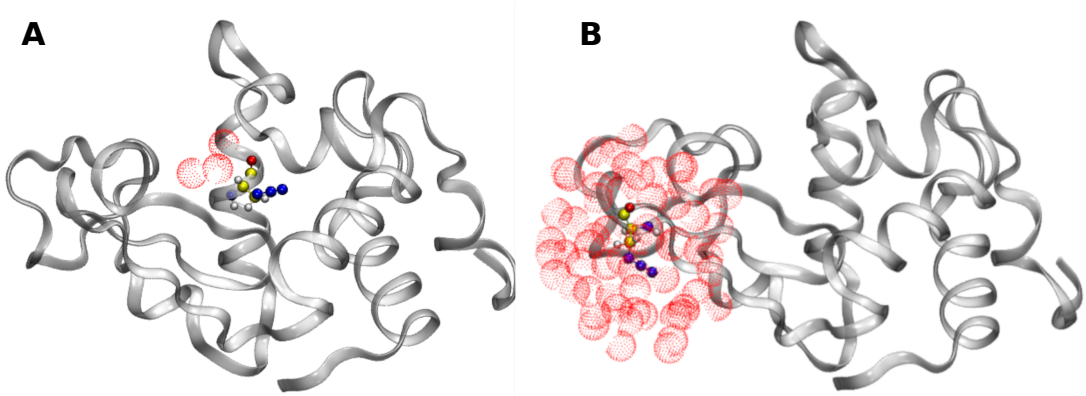


Figure 5.16: Solvent distribution based on the water-oxygen atoms within 7 Å of any atom of residues Ala96N_3 (panel A) and Ala47N_3 (panel B). The small and large hydration spheres are consistent with the $g(r)$ and $N(r)$ reported in Figure 5.15.

Another measure to quantify the solvent exposure of amino acids is to determine the time dependent quantity, $\delta\lambda^{(r)}_{\text{phob}}(t)$, which is referred to as the local hydrophobicity (LH) of residue r at time t.[209,210] This measure is based on analyzing the occupation and orientational statistics of surface water molecules at the protein/water interface, given by the 3-dimensional vector $\vec{\kappa} = (a, \cos\theta_{\text{OH1}}, \cos\theta_{\text{OH2}})$. Here, a is the distance of the water oxygen atom to the nearest atom of residue r[190], and θ_{OH1} and θ_{OH2} are the angles between the water OH1 and OH2 bonds and the interface normal. More specifically, the local hydrophobicity (LH) is $\delta\lambda^{(r)}_{\text{phob}}(t) = \lambda^{(r)}_{\text{phob}}(t) - \langle\lambda_{\text{phob}}\rangle_0$, where

$$\lambda^{(r)}_{\text{phob}}(t) = -\frac{1}{\sum_{a=1}^{N_a(r)} N_w(t;a)} \sum_{a=1}^{N_a(r)} \sum_{i=1}^{N_w(t;a)} \ln\left[\frac{P(\vec{\kappa}^{(i)}(t)|\text{phob})}{P(\vec{\kappa}^{(i)}(t)|\text{bulk})}\right] \tag{5.7}$$

and $\langle\lambda_{\text{phob}}\rangle_0$ is the ensemble average sampled from the ideal hydrophobic reference system (see below). The summation over $N_a(r)$ involves all atoms in residue r and the summation over $N_w(t;a)$ includes all water molecules within a cut-off of 6Å of atom a at time t.[190] The vector $\vec{\kappa}^{(i)}(t)$ describes the orientation (see above) of the ith water molecule in the sampled population.

The distribution $P(\vec{\kappa}^{(i)}(t)|\text{phob})$ is determined for a reference hydrophobic reference system ('phob'), whereas $P(\vec{\kappa}^{(i)}(t)|\text{bulk})$ is determined from the actual simulations ('bulk').[209] As the quantity LH includes both, the distance a of the water molecules from the interface and the orientation of a specific water molecule $(\theta_{\text{OH1}}, \theta_{\text{OH2}})$, LH can be considered as a generalization of the radial distribution function $g(r)$. The local hydrophobicity is a measure of the statistical similarity of the sampled configurations to that of an ideal hydrophobic reference system. When sampled configurations $P(\vec{\kappa}^{(i)}(t)|\text{bulk})$ are dissimilar to the hydrophobic reference system, this indicates that the site r considered is less hydrophobic, i.e. rather hydrophilic and vice versa. In other words, $\delta\lambda^{(r)}_{\text{phob}}(t) \approx 0$ for a hydrophobic environment around residue r, whereas $\delta\lambda^{(r)}_{\text{phob}}(t)$ significantly larger than zero, the environment is hydrophilic.[190,209,210] In previous work[190], sustained values of $\delta\lambda^{(r)}_{\text{phob}} > 0.5$ were considered indicative of hydrophilicity. The magnitude of such a cutoff may, however, be somewhat system-dependent.

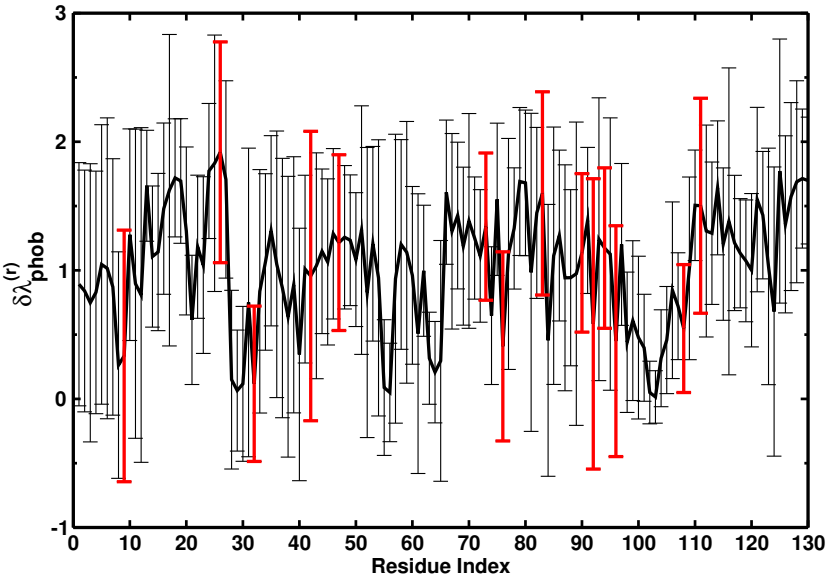


Figure 5.17: Average local hydrophobicity together with fluctuations for WT lysozyme (no $-N_3$ labels attached) for the 5 ns simulation. LH for all residues in black and for Ala in red.

Figure 5.17 gives an overview of the average LH per residue and the fluctuations around the average for WT lysozyme. The Alanine residues (in red) are found to include both, low and high values for LH, representative of more hydrophobic and hydrophilic environments, respectively. The change in LH as a function of simulation time (over 2 ns) for WT (blue) and $-N_3$ labelled (red) lysozyme for Ala76 is reported in Figure 5.18. Without spectroscopic label the Ala-residue is rather hydrophilic on average whereas with the label attached it is more hydrophobic (less hydrophilic). On the other hand, the LH can have a rather pronounced time-dependence, see Figure 5.19 (solid orange line for Ala76) from the 5 ns simulation of WT lysozyme. Thus, attaching the $-N_3$ label to Ala may modulate recruitment or displacement of solvent molecules, as was also found recently for nitrosylated Mb.[211]

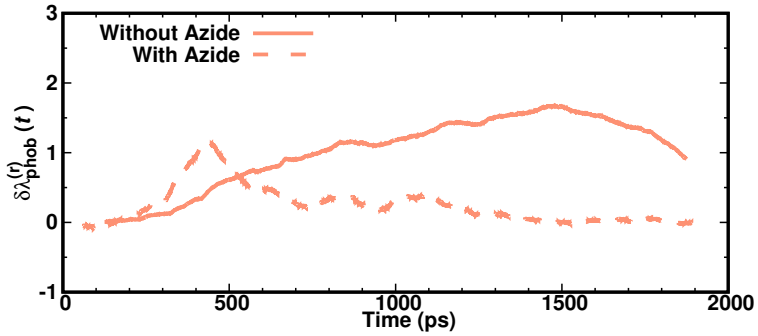


Figure 5.18: Local hydrophobicity for residue 76 with and without azide group attached to Ala for 2 ns. The effect of attaching azide on the LH, and hence the hydration itself, is clearly visible.

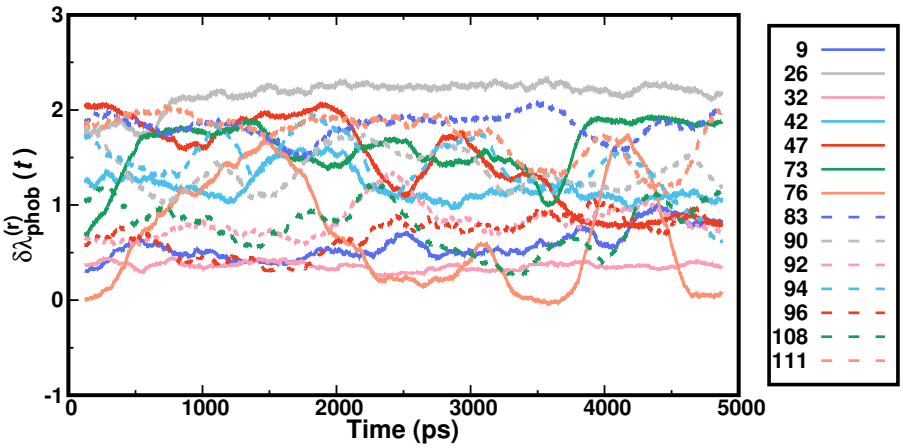


Figure 5.19: Local hydrophobicity as a function of time for all alanine residues from the simulation of WT lysozyme. The LH coefficient was determined from Eq. 5.7. Values of $\delta\lambda_{\mathrm{phob}} \approx 0$ indicate a hydrophobic environment of the site considered[190,210] whereas values around 2 point towards a hydrophilic site.

5.6 Discussion and Conclusion

The present findings confirm that azide attached to alanine residues in lysozyme is a structurally minimally invasive, specific infrared label to quantitatively probe the local dynamics around the modification site. This has already been reported for the PDZ2 domain.[31] Similar to the situation in insulin monomer and dimer, for which the amide-I vibration was found[135,200] to cover a range of ~ 20 cm^{-1}, attaching azide to give AlaN_3 spans a comparable frequency range but in a region of the infrared spectrum (around 2100 cm^{-1}) that is typically "empty". Together with their minimal impact on the overall protein structure (see Figure 5.5), and the still favourable extinction coefficient[23], such modifications bear great potential to resolve the structural dynamics of proteins and protein-ligand complexes at a molecular level. Studies that provide structural and spectroscopic information at the same time are of great interest for characterizing potential ligand-binding sites and for functional studies of protein allostery.

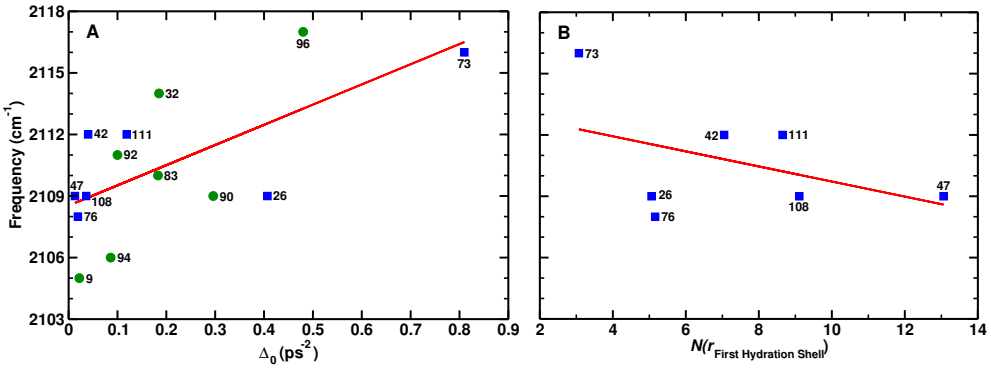


Figure 5.20: Correlation between the maximum of the 1-d lineshape from INM and he static offset Δ_0 of the FFCF (panel A) and the maximum of the 1-d lineshape from INM and the number of water molecules in the first hydration shell (panel B) for the residues that has been considered to be rather "water exposed" (Set1). Residues of Set1 are shown as blue squares and those of Set2 as green circles. The solid line is an empirical linear fit and suggests that, typically, for more blue shifted frequency maxima the static component increases while the number of water molecules in the first hydration shell decreases, i.e. with increasing hydration, $\omega_{\max}$ shifts typically to the red for alanine residues in Set1.

It is of interest to delineate whether correlations can be found between structural and spectroscopic characteristics analyzed in the present work. As the dynamics is coupled and involves a potentially complicated superposition of different structural substates, no "simple" or "obvious" correlations are expected. Rather and

at best, discovering trends can be expected from such an analysis. One example is shown in Figure 5.20B which reports the relationship between the number of water molecules in the first hydration shell (see also Figure 5.15) and the position of the frequency maximum $\omega_{\max}$ from the 1-d lineshape determined from the instantaneous normal mode analysis. Typically, with increasing hydration, the position of $\omega_{\max}$ shifts to the red. Similarly, the magnitude of the static offset Δ_0 of the FFCF is related to $\omega_{\max}$ in that larger values of Δ_0 are associated with a blue shift of the position of the frequency maximum, see Figure 5.20A.

In addition, the magnitude of the static offset may also be related to the local structural fluctuations of the protein. To assess this, the root mean squared fluctuations (RMSFs) of all C_α atoms for the WT, Ala32N_3, and Ala47N_3 variants were determined, see Figure 5.21. These two variants were selected because the FFCF for Ala47N_3 decays close to zero on the 5 ps time scale with a short $\tau_2 = 1.17$ ps whereas that for Ala32N_3 has a finite $\Delta_0 = 0.185$ ps^{-1} and the longest $\tau_2 = 13.05$ ps. Analysis of the RMSF shows that Ala32 is located in a region of the protein with particularly small fluctuations whereas the converse is true for Ala47, see Figure 5.21. Hence, it is anticipated that local flexibility/rigidity, in addition to local hydration, also contributes to differences in τ_2 and Δ_0.

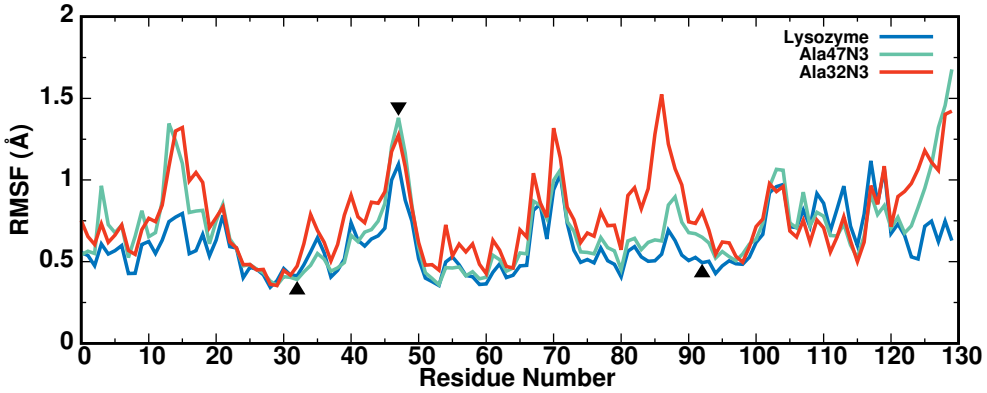


Figure 5.21: Root mean squared fluctuations (RMSF) for the C_α atoms for the WT, Ala32N_3, and Ala47N_3 proteins. The positions of residues 32, 47, and 92 are indicated as triangles. Ala32 is located in a region with low RMSF ("more rigid") whereas Ala47 is in a region with high RMSF ("more flexible"). Ala92N_3 is at the boundary between a flexible and more rigid region of the protein.

For a more comprehensive characterization of the interplay between τ_2, Δ_0, and local fluctuations, the dynamical cross correlation maps (DCCMs) and their dif-

ferences for WT and two modified proteins were determined. For this, Ala47N_3 and Ala92N_3 were considered because they show similar τ_2 (1.17 ps vs. 1.33 ps, see Table 5.1) but differ in Δ_0 and the overall appearance of the FFCF, see Figure 5.9. The DCCMs, determined using the Bio3D package,[212] report on correlated and anticorrelated motions within a protein whereas their difference maps (ΔDCCM) provide information about which of the couplings are affected upon modification of the protein. The ΔDCCM between WT and Ala92N_3 (see Figure 5.22B) reveals only small differences which suggests that the overall motions and couplings of the two proteins are similar. This contrasts with the ΔDCCM for WT and Ala47N_3 (Figure 5.22A) which indicates that couplings between Ala47 and residues 42 to 52 are reduced in Ala47N_3. As a consequence of the rigid region (residues 25 to 32, see Figure 5.21), correlated motions for residues 13 to 20 are also reduced. Interestingly, these two regions (13 to 20 and 42 to 52) are also those with particularly large RMSFs, see Figure 5.21. It is conceivable that the short helix (residues 25 to 32, see Figure 5.22C) acts as a piston that transduces motion between residues 13 to 20 and residues 42 to 52 similar to what was found in Mb where energy transfer from the solvent to a photodissociated CO ligand was also found to occur through the helices.[213] Together with the pronounced solvent exposure of Ala47N_3 (Figure 5.16) it is concluded that the FFCF for this residue is primarily determined by solvent exposure whereas that for Ala92N_3 is determined through coupling to the protein motion. This is further confirmed by the observation that azide label motion coupled to the protein dynamics can result in pronounced recurrences in the FFCF as found here for Ala92N_3 and experimentally reported for the azide anion bound to the active site of formate dehydrogenase.[214] Fourier transformation of the FFCF for Ala92N_3 (see Figure 5.9) reveals coupling between the azide bend and stretch vibrations and the low frequency motions of the protein.

Spectroscopic probes to characterize the local environment of a protein provide valuable information about local hydration. This is of particular relevance given the findings that individual water molecules can play decisive roles in protein function. For example, in HIV-I protease[215,216] a single catalytic water molecule was located in the active site of the protein or for insulin[93,135] individual water molecules were found to attack the dimerization interface to reduce the thermodynamic stability of the dimer by a factor of two. Similarly, water molecules have

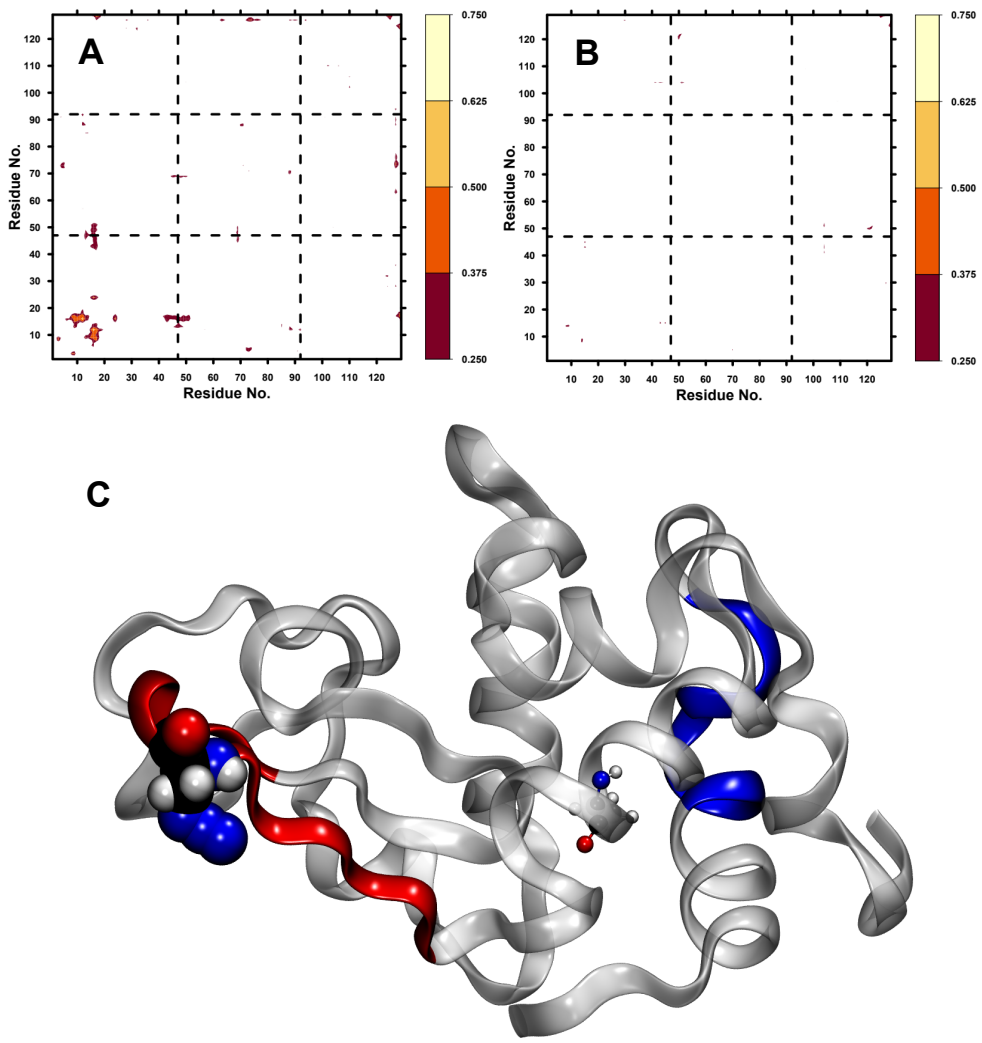


Figure 5.22: Difference dynamic cross correlation maps (ΔDCCM) between WT and Ala47N_3 (panel A) and between WT and Ala92N_3 (panel B). Panel C highlights residues 13 to 20 (blue) and 42 to 52 (red) which show pronounced differences in panel A. Residue Ala47N_3 is in van der Waals and residue Ala32 (along the helix with residues 28 to 32) is in CPK representation.

been reported to play essential roles in protein folding,[217] and for function.[218,219] Thus, probing and characterizing the local solvent environment of particular regions of a protein can provide important insights into functional aspects of proteins.

The utility of infrared spectroscopy to study the strength of protein-ligand complexes has been proposed[96] and explicitly demonstrated from molecular dynamics simulations for cyano-substituted benzene in lysozyme.[97] Using AHA as a probe,

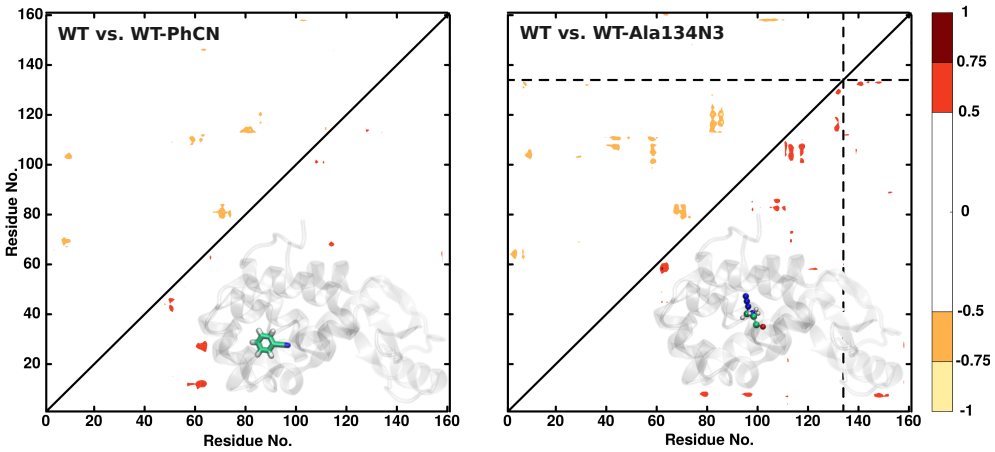
it was reported that unbound and ligand-bound PDZ2 differ in that the frequency correlation function for the two systems decay to different levels at longer correlation times. Similarly, infrared spectroscopy is also a sensitive probe - both, in terms of spectroscopy and dynamics - to characterize protein-protein interactions.[200] Together with experimental studies,[90,92,133] such efforts pave the way for functional, in vivo studies of protein-ligand and protein-protein association.[220]

In conclusion, the present work provides a comprehensive analysis of the spectroscopy and dynamics of azide-labelled alanine in lysozyme. The results demonstrate that $AlaN_3$ is a positionally sensitive probe for the local dynamics, covering a frequency range of ~ 15 cm^{-1}. This is consistent with findings from selective replacements of amino acids in PDZ2 which reported a frequency span of ~ 10 cm^{-1} for replacements of Val, Ala, or Glu by AHA.[23] Furthermore, the long-time decay constants τ_2 range from ~ 1 to ~ 10 ps which compares with experimentally measured correlation times of 3 ps.[23] Attaching azide to alanine residues can yield dynamics that decays to zero on the few ps time scale (i.e. $\Delta_0 \sim 0$ ps^{-1}) or to a remaining inhomogeneous contribution of ~ 0.5 ps^{-1} (corresponding to 2.5 cm^{-1}). One exciting prospect of this is to determine how the spectroscopy and dynamics of the modification site changes upon ligand binding to the active site for lysozyme or other proteins.

Transparent window vibrational probes allow for character- ization of specific bonds within proteins and other complex molecules or systems, which in turn provides a window into the bond's microenvironment, conformational heterogeneity, and local dynamics that is otherwise difficult or impossible to obtain.

Chapter 6

Site-Selective Dynamics of Ligand-Free and Ligand-Bound Azidolysozyme



The results presented in this chapter have been previously published:
J. Chem. Phys. 2022, 156, 105105.
doi:10.1063/5.0077361

***

6.1 abstract

Azido-modified alanine residues ($AlaN_3$) are environment-sensitive, minimally invasive infrared probes for the site-specific investigation of protein structure and dynamics. Here, the capability of the label is investigated to query whether or not a ligand is bound to the active site of Lysozyme and how the spectroscopy and dynamics change upon ligand binding. The results demonstrate specific differences for center frequencies of the asymmetric azide stretch vibration, the long time decay and the static offset of the frequency fluctuation correlation function - all of which are experimental observables - between the ligand-free and the ligand-bound, N_3-labelled protein. Changes in dynamics can also be mapped onto changes in the local and through-space coupling between residues by virtue of dynamical cross correlation maps. This makes the azide label a versatile and structurally sensitive probe to report on the dynamics of proteins in a variety of environments and for a range of different applications.

Proteins are essential for function and sustaining life of organisms. Experimentation and computational characterization has clarified that protein function involves both, structure *and* dynamics.[3,4,181] However, characterizing structural and functional dynamics of proteins at the same time under physiological conditions in the condensed phase, which is prerequisite for clarifying cellular processes at a molecular level, remains a challenging undertaking.[181] Vibrational spectroscopy, in particular 2-dimensional infrared (2D-IR) spectroscopy, has been shown to be a powerful tool for studying the structural dynamics of various biological systems[128]. One of the particular challenges is to obtain structural and environmental information in a site-specific manner. To address this, significant effort has been focused on the development and application of various infrared (IR) reporters[7,8] that absorb in the frequency range of 1700-2800 cm^{-1} to discriminate the signal from the strong protein background.[138,182] Such IR probes have provided valuable information about the structure and dynamics of complex systems. For example, nitrile probes have helped to clarify the role of electrostatic fields in enzymatic reactions[9,10] or to elucidate the mode of drug binding to proteins.[11,12] Isotope edited carbonyl spectroscopy was used to characterize the mechanism of protein folding and amyloid formation[13,14] or the structure and function of membrane proteins[15,16]. Additional molecular groups such as thiocyanate,[17] cyanamide,[18] sulfhydryl vibrations of cysteines,[19] deuterated car-

bons,[20] carbonyl vibrations of metal-carbonyls,[21] cyanophenylalanine,[22] and azidohomoalanine (AHA)[23] have also been explored.

In the present work $AlaN_3$, an analogue of azidohomoalanine (AHA) that has been shown to sensitively report on local structural changes while still being minimally invasive,[30,221] is used as the probe. This modification can be incorporated into proteins at virtually any position via known expression techniques.[191] The asymmetric stretch frequency of -N_3 is at $\sim$ 2100 cm^{-1} and has a reasonably high extinction coefficient of 300-400 $M^{-1}cm^{-1}$ which makes it an ideal spectroscopic reporter.[23] AHA has been used for biomolecular recognition after incorporation into the peptide directly[23,30] or in the vicinity of binding area of a PDZ2 domain[31], to detect the water-specific response of azide vibrations when attached to small organic molecules[32], or to probe the frequency shift and fluctuation due to its sensitivity to the local electrostatic environments and dynamics.[33,221] Such studies confirm that $AlaN_3$ and/or AHA are environment-sensitive IR probes and suitable modifications for site-specific investigations of protein structure.

With its picosecond time resolution, IR spectroscopy provides direct information about the structural dynamics around a probe molecule with high temporal resolution.[128,222] Moreover, introducing IR probes with isolated vibrational frequencies overcomes the problem of spectral congestion that complicates discrimination and analysis of desired vibrational bands. With that, the inter- and intramolecular coupling between degrees of freedom or the local structure or dynamics of biological systems can be specifically probed and characterized. Such an approach relies on the sensitivity of the probe to report on changes in the vibrational frequencies induced by alterations in the local electrostatic interactions in the vicinity of the probe.[22]

IR spectroscopy is a potentially advantageous technique to characterize ligand binding to proteins.[96,97] Its success depends in part on the notion that when a ligand binds to a protein, the frequency of an infrared active vibration shifts due to the different electric field in solution - often water - and in the protein binding site. Such an approach often requires the ligand to be modified, e.g. through addition of a suitable label such as -CN as in benzonitrile. This has been

successfully demonstrated for benzonitrile in the active site of WT and mutant lysozyme[97].

Alternatively, the protein can be selectively modified by attaching a spectroscopic label at strategic positions so that the binding process and functional dynamics can be interrogated with the functionally relevant, unmodified ligand. This has the potential advantage that interactions between the ligand and the surrounding protein are unaltered. These interactions contribute the majority of the enthalpic part to the binding free energy and therefore directly affect the affinity of the ligand and its rate of unbinding. In the present work changes in 1D- and 2D-IR signatures of the azido group attached to all alanine (Ala) residues of lysozyme upon binding of cyano-benzene (PhCN) are determined. In addition, the changes of the environmental dynamics around all $AlaN_3$ are quantified for ligand-free vs. ligand-bound lysozyme. Such differences are experimentally observable and yield valuable insight into the energetics and dynamics of ligand-protein binding.

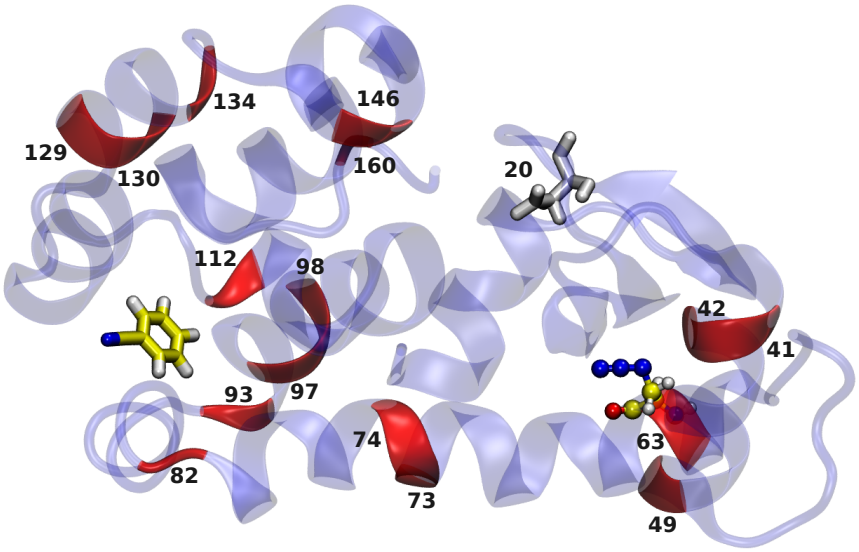


Figure 6.1: The structure of lysozyme including PhCN (Licorice) in the cavity and Ala63N3 (CPK) as an example of -N_3 label attachment. The Ala residues are at positions 41, 42, 49, 63, 73, 74, 82, 93, 97, 98 , 112, 129, 130, 134, 146, 160 (red NewCartoon). The rest of the protein is shown as blue NewCartoon except for residue Asp20 which is in white Licorice.

The dynamics of WT lysozyme without and with labeled alanine ($AlaN_3$) has been recently found to yield position-specific information about the spectroscopy

and dynamics of the modification site.[221] The structure of the protein with the labelled Ala residues is shown in Figure 6.1 together with the binding site lined by residues Leu84, Val87, Leu91, Leu99, Met102, Val111, Ala112, Phe114, Ser117, Leu118, Leu121, Leu133, Phe153. Following earlier work,[97] benzene was replaced by cyano-benzene (PhCN) maintaining Carbon atom positions. The WT structure was used here to a) compare directly with earlier results[221] and b) because PhCN has a comparatively small binding free energy towards the WT protein ($\Delta G_{\text{bind}} = -0.5$ kcal/mol) which suggests that the interaction between the ligand and the protein is weak.[97] For the L99A mutant protein $\Delta G_{\text{bind}} = -3.9$ kcal/mol for PhCN[97] which compares with an experimentally determined value of ~ -3.5 kcal/mol for iodobenzene from isothermal titration calorimetry.[223]

For the -N_3 label a full-dimensional, accurate potential energy surface (PES) calculated at the pair natural orbital based coupled cluster (PNO-LCCSD(T)-F12/aVTZ)[193,194] level and represented as a reproducing kernel Hilbert space (RKHS)[73,74] is available.[106] This energy function is suitable for spectroscopic investigations and was combined with the CHARMM force field[67] for the surrounding protein.[221] MD simulations for the WT and all modified AlaN_3 labels were carried out using an adapted version of the CHARMM program[64] with an interface to perform the simulations with the RKHS PES.[106] The protein is solvated in explicit TIP3P water[117] using a cubic box of size $(78)^3$ Å^3. First, all systems were minimized which was followed by heating and equilibration. Next, 2 ns NVT production simulations were carried out with and without the ligand present in the active site for all 16 protein variants with Ala replaced by AlaN_3. Bonds involving H-atoms were constrained using the SHAKE[123] algorithm and all nonbonded interactions were evaluated with shifted interactions using a cutoff of 14 Å and switched at 10 Å.[124] Snapshots for analysis were recorded every 5 fs.

The effect of ligand binding on the overall flexibility of the modified protein can be assessed from considering the root mean squared fluctuation (RMSF) of the C_α atoms. Depending on the position at which the -N_3 label is located, the changes in RMSF range from insignificant (Ala82N_3 or Ala160N_3) to major (Ala73N_3 or Ala112N_3), see Figure 6.2.

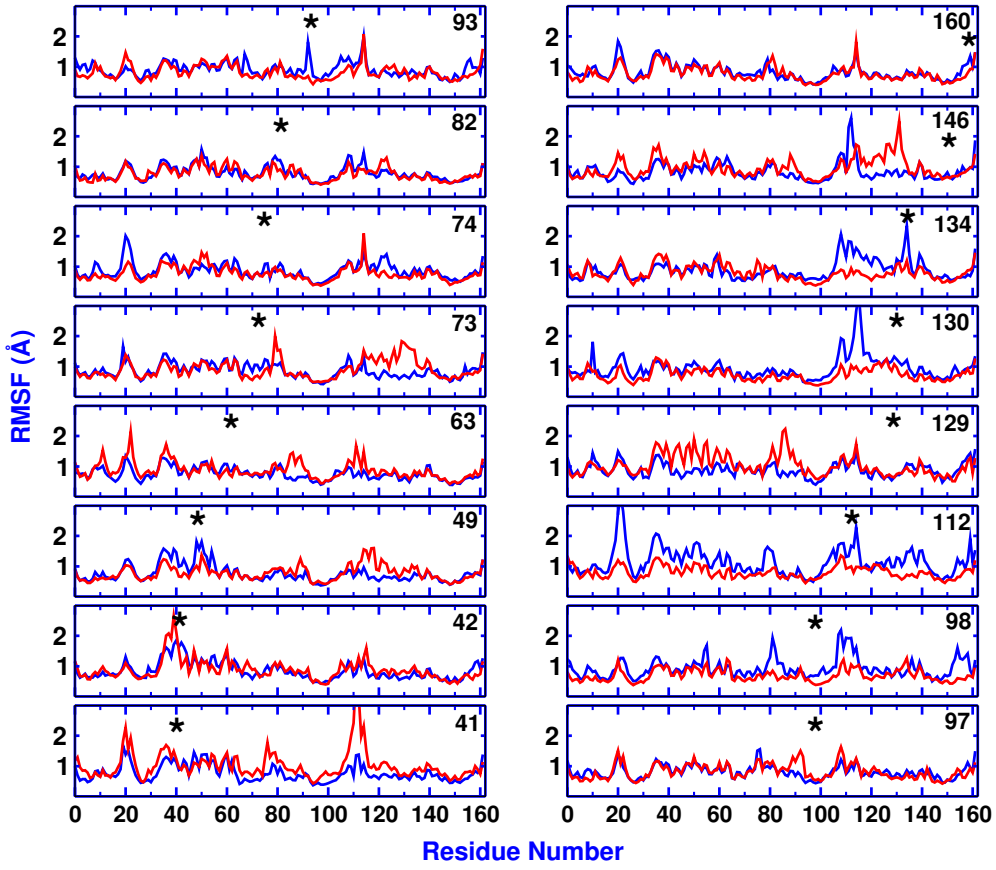


Figure 6.2: Root mean squared fluctuations (RMSFs) for the C_α atoms for ligand-free (blue) and PhCN-bound (red) lysozyme with -N_3 attached to every alanine residues. The label in each panel refers to the alanine residue number which carries the azide label and the corresponding position of residue is indicated as asterisk above the RMSF trace.

Using instantaneous normal mode (INM) analysis[106,221,224] the frequency trajectory $\omega(t)$ of the asymmetric stretch vibration of the -N_3 label was determined. Based on this, the 1D infrared spectra corresponding to the azide asymmetric stretch vibration for each of the 16 AlaN_3 residues was computed for the apo- and holo-protein, see Figures 6.3. Direct comparison of the maximum position of the infrared lineshape shows that for three N_3-modified alanine residues (Ala41, Ala98, Ala130) the difference in the absorption frequency is insignificant. Finally, for positions Ala63, Ala73 and Ala160 the differences are 8, 3, and 2 cm^{-1}, respectively, while for the other residues the change is within 1 cm^{-1}. Previous simulations of the vibrational Stark effect for the nitrile probe in PhCN with frequencies from instantaneous normal modes reported red shifts of up to 3.5 cm^{-1} for the -CN stretch in going from the WT to the L99A and L99G mutants of T4-lysozyme[97] which was also found for the nitrile probe in the active site of

human aldose reductase.[96] Similarly, the 1D and 2D infrared spectroscopy of -CO as the label for insulin monomer and dimer found that the relative shifts of the spectroscopic response was correctly described whereas the absolute frequencies may differ by some 10 cm^{-1}. In a very recent work such an approach found a splitting of 13 cm^{-1}, compared with 25 cm^{-1} from experiment, for the outer and central -CO labels in cationic trialanine in water.[225] Hence, MD simulations together with instantaneous normal modes are a successful approach to determine relative frequency shifts whereas capturing absolute frequencies in such simulations requires slight reparametrization of the underlying force field, e.g. through morphing techniques.[202,203]

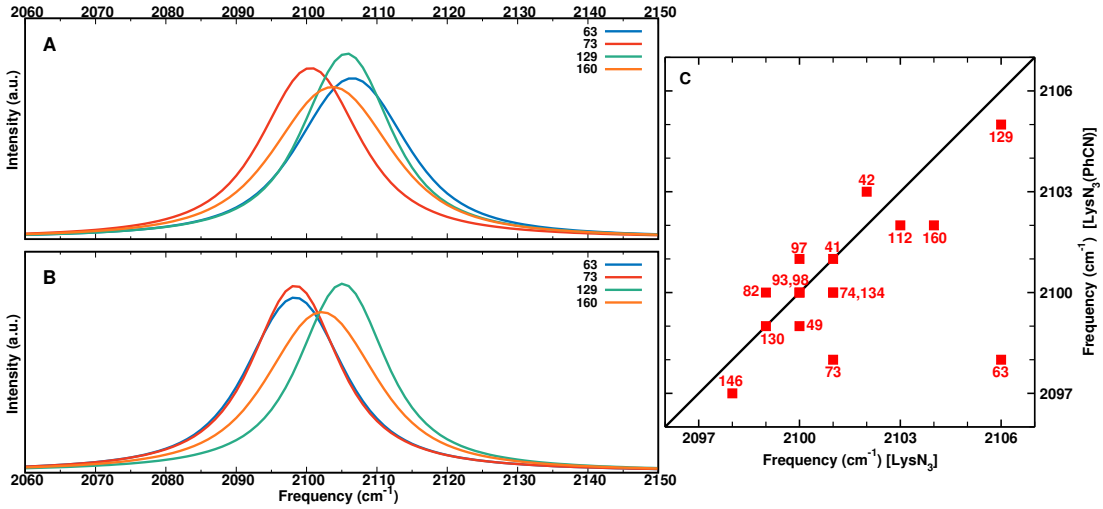


Figure 6.3: 1D IR spectra from INM for four $AlaN_3$ residues (Ala63, Ala73, Ala129, and Ala160) for ligand-free (panel A) and ligand-bound (panel B) Lysozyme. Panel C compares the maximum frequency of the 1D IR spectra for all modified Ala residues for ligand-free (along the $x-$acis) and ligand-bound (along the $y-$axis) N_3-labelled lysozyme.

The magnitude of frequency shifts is comparable to the red shift $\Delta\omega = -4$ cm^{-1} of the -CN stretch for PhCN in the active site of WT vs. L99A mutant lysozyme[97], -3 cm^{-1} from experiments of the nitrile stretch in ligand IDD743 bound to WT vs. V47N mutant hALR2[96] and 6 cm^{-1} blue shift of -CO vibrational frequency due to the binding of 19-NT to KSI compared to the WT.[226] Thus, differences of ~ 1 cm^{-1} for the frequency of the reporter in different chemical environments can be experimentally detected.[22]

From the frequency trajectories the frequency fluctuation correlation function (FFCF) can be determined which contains valuable information on time scales characteristics corresponding to the solvent variation due to the presence of solute. They are fit to an empirical functional form

$$\langle \delta\omega(t)\delta\omega(0)\rangle = a_1 \cos(\gamma t) e^{-t/\tau_1} + \sum_{i=2}^{n} a_i e^{-t/\tau_i} + \Delta_0^2 \tag{6.1}$$

which allows analytical integration to obtain the lineshape function[129] using an automated curve fitting tool from the SciPy library.[130] As for the RMSFs and 1D IR spectra, the FFCFs from the simulations with and without the ligand bound to the protein can be very similar or differ appreciably, see Figure 6.6. The slow decay time, τ_2, of the -N_3 asymmetric stretch mode of the label is typically faster for the ligand-bound protein compared to that without PhCN, see Figure 6.6B, although exceptions exist. For Ala97N_3, Ala112N_3, and Ala134N_3 the slow relaxation time τ_2 is faster by 75 % up to a factor of ~ 2.5 and for Ala146N_3 the slow time scale, τ_2, differs by a factor of ~ 3 between ligand-free ($\tau_2 = 5.13$ ps) and ligand-bound ($\tau_2 = 1.61$ ps) lysozyme. For the other Alanine residues the τ_2 times between ligand-free and ligand-bound lysozyme are similar. As an exception, for Ala129N_3 the decay is slowed down by ~ 50 % for PhCN-bound lysozyme. Figure 6.4 and 6.5 demonstrate the FFCF of all AlaN_3 for the ligand-free and ligand-bound, respectively. Furthermore, the corresponding fitting parameters are given in Table 6.1.

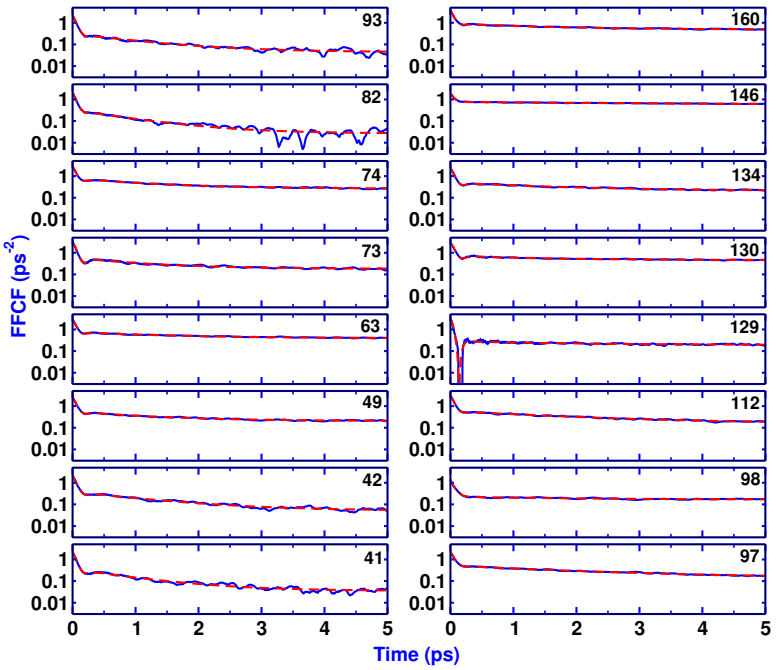


Figure 6.4: FFCFs from correlating the instantaneous harmonic frequencies for all 16 $AlaN_3$ in lysozyme. The labels in each panel refer to the alanine residue which carries the azide label. Black traces are the raw data and red dashed lines the fits to Eq. 6.1. The $y-$axis is logarithmic.

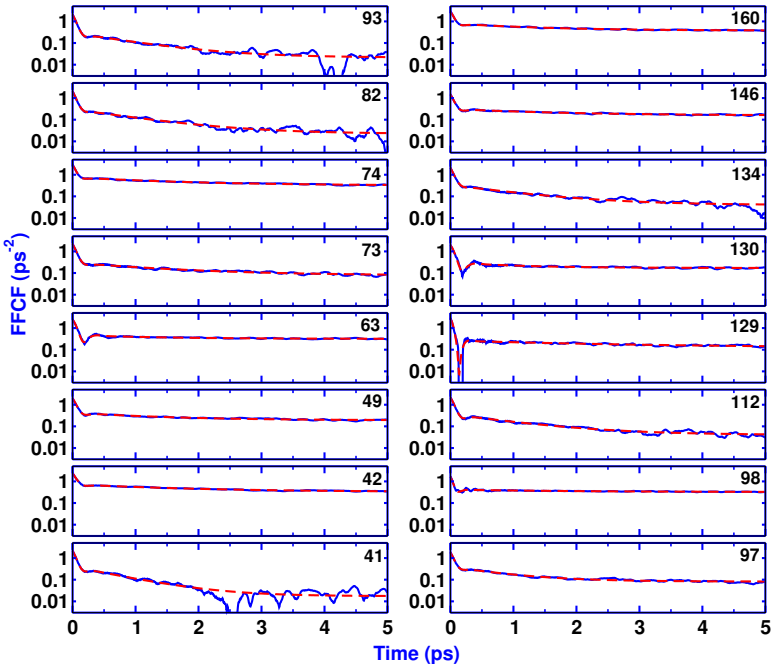


Figure 6.5: FFCFs from correlating the instantaneous harmonic frequencies for all 16 $AlaN_3$ in lysozyme with PhCN in the cavity. The labels in each panel refer to the alanine residue which carries the azide label. Black traces are the raw data and red dashed lines the fits to Eq. 6.1. The $y-$axis is logarithmic.

$LysN_3$							
Res	$\langle\omega\rangle$	a_1	γ	τ_1	a_2	τ_2	Δ_0^2
41	2099.69	1.86	11.23	0.068	0.32	0.93	0.03
42	2100.53	1.76	10.21	0.069	0.33	1.17	0.05
49	2099.02	2.00	10.67	0.069	0.36	1.18	0.20
63	2105.49	2.05	11.71	0.069	0.37	1.51	0.39
73	2099.56	2.49	13.04	0.078	0.40	1.05	0.18
74	2099.59	1.94	12.80	0.072	0.52	1.21	0.26
82	2098.36	1.76	9.15	0.062	0.30	0.90	0.02
93	2099.00	1.92	8.88	0.065	0.27	1.05	0.04
97	2098.86	1.70	7.96	0.067	0.39	2.28	0.13
98	2099.47	1.15	0.0	0.057	0.06	2.04	0.16
112	2101.62	2.40	9.41	0.072	0.44	1.99	0.15
129	2104.69	2.83	18.08	0.066	0.09	1.98	0.18
130	2097.57	2.19	12.36	0.080	0.28	1.36	0.45
134	2100.17	1.87	11.55	0.074	0.29	1.79	0.20
146	2096.84	1.16	6.15	0.057	0.24	5.13	0.52
160	2102.67	2.64	10.20	0.065	0.45	1.59	0.47
$LysN_3$–PhCN							
41	2099.90	1.77	11.78	0.068	0.38	0.70	0.01
42	2102.22	1.66	10.28	0.074	0.40	1.74	0.32
49	2098.31	1.56	10.52	0.072	0.23	1.34	0.19
63	2097.35	1.99	13.90	0.094	0.16	1.26	0.31
73	2097.38	1.94	8.92	0.067	0.22	1.40	0.07
74	2098.61	2.12	10.80	0.064	0.44	1.58	0.31
82	2098.62	1.76	9.10	0.063	0.28	0.96	0.02
93	2099.00	1.80	10.00	0.067	0.27	0.88	0.02
97	2099.65	1.46	10.88	0.064	0.31	0.81	0.08
98	2099.44	1.23	18.07	0.054	0.09	1.96	0.32
112	2100.91	1.79	11.01	0.074	0.32	1.01	0.04
129	2104.11	2.55	17.75	0.066	0.14	2.78	0.12
130	2098.13	1.64	13.09	0.090	0.13	1.00	0.17
134	2099.16	1.76	8.28	0.063	0.29	1.03	0.03
146	2096.41	1.17	10.61	0.068	0.15	1.61	0.15
160	2101.14	2.08	11.65	0.069	0.42	1.29	0.37

Table 6.1: Parameters obtained from fitting the FFCF to Eq. 6.1 for INM frequencies for all different $AlaN_3$ residues in lysozyme. Average frequency $\langle\omega\rangle$ of the asymmetric stretch in cm^{-1}, the amplitudes a_1 to a_3 in ps^{-2}, the decay times τ_1 to τ_3 in ps, the parameter γ in ps^{-1} and the static term Δ_0^2 in ps^{-2}.

As a last feature of the FFCF it is found that the static component Δ_0 can differ appreciably between ligand-free and -bound lysozyme. The static offset Δ_0 is an experimental observable and characterizes the structural heterogeneity around the modification site. There are only four alanine residues for which the static offset is similar (Ala41, Ala49, Ala82, and Ala93) for ligand-bound and ligand-free lysozyme. For all others the differences range from 15 % to a factor of ~ 3. As an example, for $Ala73N_3$ the difference for Δ_0^2 between bound and ligand-free

lysozyme is a factor of ~ 2.5 ($\Delta_0^2 = 0.18$ vs. 0.07 ps^{-2} or $\Delta_0 = 0.42$ vs. $\Delta_0 = 0.26$ ps^{-1}) and for Ala146N$_3$ they differ by a factor of ~ 3.5 ($\Delta_0^2 = 0.52$ vs. 0.15 ps^{-2}; i.e. $\Delta_0 = 0.72$ vs. $\Delta_0 = 0.39$ ps^{-1}). Thus, the environmental dynamics around the spectroscopic label can be sufficiently perturbed by binding of a ligand in the protein active site to be reported directly as an experimentally accessible quantity with typical errors[22] between 0.1 cm^{-1} and 0.3 cm^{-1} (~ 0.05 ps^{-1}). Hence, the differences found from the simulations are well outside the expected error bars from experiment.

Nonvanishing static components of the FFCF were also reported from experiments. For trialanine $(Ala)_3$ a value of $\Delta_0 = 5$ cm^{-1} was reported[227] compared with $\Delta_0 = 4.6$ cm^{-1} from MD simulations (0.94 ps^{-1} vs. 0.86 ps^{-1}) with multipolar force fields.[225] Similarly, CN^- in water features a nonvanishing tilt angle by $\tau = 10$ ps[147] with $\Delta_0 \sim 0.1$ ps^{-1} ~ 0.5 cm^{-1}.[104] Finally, 2D IR experiments for p-cyanophenylalanine bound to six distinct sites in a Src homology 3 domain reported static components ranging from $\Delta_0 = 1.0$ to 3.7 cm^{-1} (corresponding to 0.19 ps^{-1} to 0.70 ps^{-1}).[22]

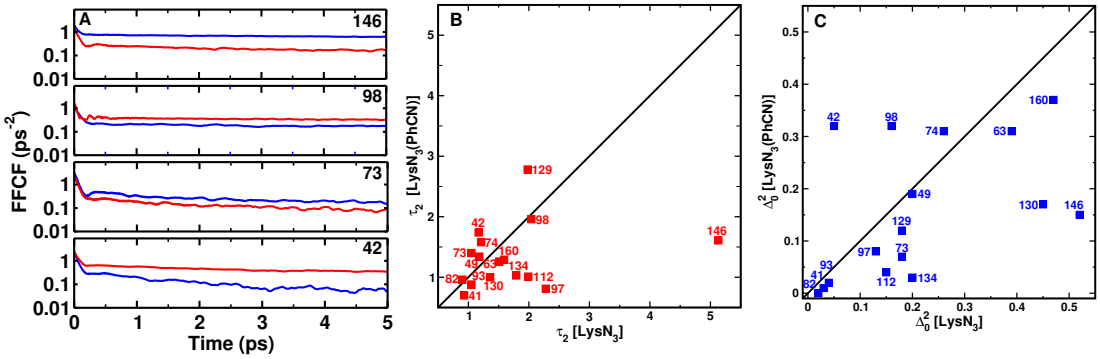


Figure 6.6: A) FFCFs with pronounced differences from correlating the INM frequencies for ligand-free and -bound Ala73, Ala146, Ala42, Ala98 in lysozyme. The labels in each panel refer to the Ala residue which carries the azide label. Blue (ligand-free) and red (ligand-bound) traces are the fits to Eq. 6.1. The $y-$axis is logarithmic. Panel B and C show the comparison of τ_2 and Δ_0^2, respectively.

To determine in which way the dynamics of residues is affected upon modification of the protein, dynamical cross-correlation maps[228,229] (DCCM) were calculated

from the trajectories using the Bio3D package.[212] Dynamic cross-correlation matrices are based on the expression

$$C_{ij} = \langle \Delta r_i . \Delta r_j \rangle / (\langle \Delta r_i^2 \langle \Delta r_j^2 \rangle)^{1/2} \tag{6.2}$$

where r_i and r_j are the spatial C_α atom positions of the respective ith and jth amino acids and Δr_i corresponds to the displacement of the ith C_α from its averaged position over the entire trajectory. DCCMs report on the correlated and anticorrelated motions within a protein and difference maps provide a global view of the positionally resolved differences in the dynamics. In the following, only absolute values for C_{ij} and differences between them that are larger than 0.5 are reported. The DCCMs are symmetrical about the diagonal and for clarity, positive correlations (for DCCM) or positive differences in C_{ij} (for ΔDCCM) are displayed in the lower right triangle and negative values or differences in C_{ij} are displayed in the upper left triangle, respectively.

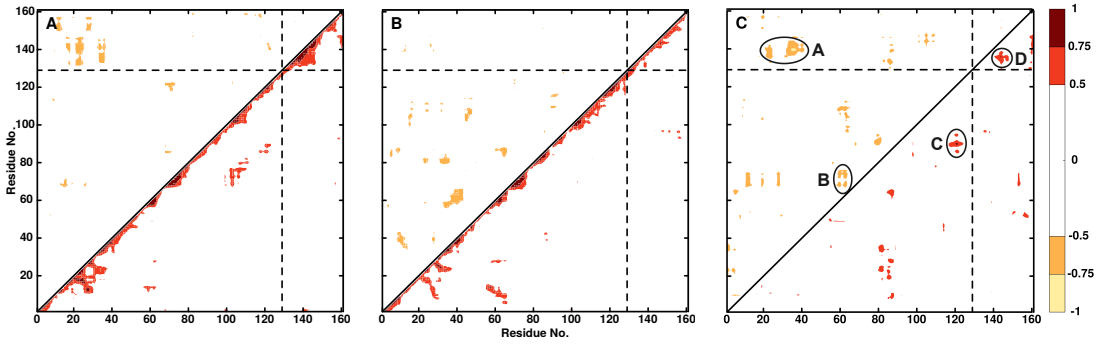


Figure 6.7: DCCM for ligand-free (panel A), ligand-bound (panel B), and ΔDCCM between ligand-free and -bound for Ala129N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

The DCCM for Lysozyme with Ala129N_3 ligand-free, ligand-bound and the difference between the two is shown in Figure 6.7. These maps reveal ligand-induced differences in the correlated and anticorrelated motions with appreciable amplitudes, see features A to D in Figure 6.7C. For ligand-free Lysozyme there are pronounced couplings between residues $[130, 147]$ and $[20, 25]/[32, 37]$ in anticorrelated motions and residues $[68, 80]$ and $[103, 112]$ for correlated motions. As demonstrated in Figure 6.7B, upon binding the ligand to Lysozyme the DCCM

shows different coupled residues compared to the ligand-free protein. As an example, residues $[35, 45]$ and $[55, 68]$ are affected more for anticorrelated motions while in the correlation ones, the coupling is between residues $[5, 15]$ and $[55, 65]$. Note that these effects may not be visible in the ΔDCCM map as the magnitude of the difference between the two systems may be below the threshold of 0.5 in the ΔC_{ij}.

In the difference map (Figure 6.7C) feature A indicates the coupling between residues $[135, 145]$ and $[20, 25]/[30, 42]$ whereas feature B refers to coupled residues $[65, 75]$ and $[58, 65]$. Furthermore, feature C demonstrates prominent variations between residues $[84, 95]$ and $[117, 125]$ while for feature D residues $[129, 140]$ and $[140, 147]$ are strongly correlated. These findings suggest that residues couple both locally (features B/D) and through space (features A/C). It should also be pointed out that residues involved in features A to C are among those with higher RMSF, see Figure 6.2. Interestingly, the region around residue Ala146 with larger differences ΔC_{ij} display correlated dynamics with spatially close residues around residue Asp20 (white licorice in Figure 6.1). On the other hand, the pronounced differences in the RMSF of Ala129N_3 (see Figure 6.2) for residues [42,57] do not show up in the ΔDCCM map because their C_{ij} coefficients are below the threshold of 0.5.

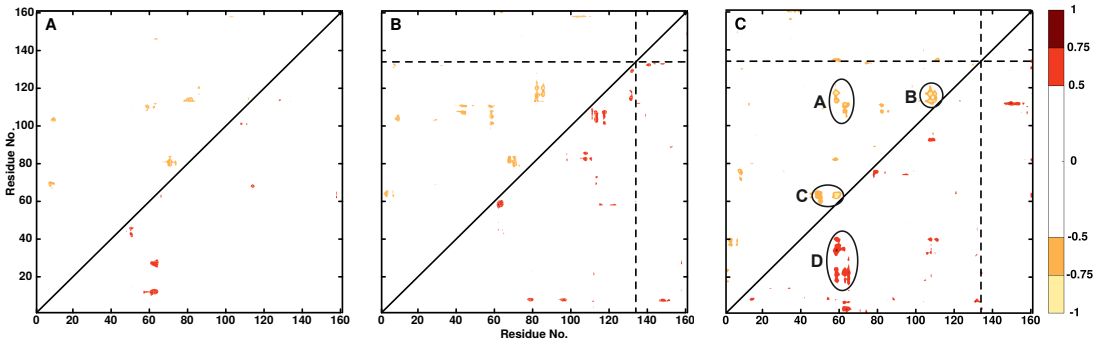


Figure 6.8: Differences in DCCM maps (ΔDCCM) between WT and WT-PhCN (panel A), WT and Ala134N_3 (panel B) and Ala134N_3 and Ala134N_3-PhCN (panel C). Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only differences in correlation coefficients with an absolute value greater than 0.5 are displayed.

Difference ΔDCCM maps between WT and ligand-bound (WT-PhCN) Lysozyme are shown in Figure 6.8A together with the difference map between WT and azido modified lysozyme at position 134 ($Ala134N_3$), see Figure 6.8B. With the PhCN ligand bound to the protein the ΔDCCM map compared with that for the ligand-free protein is sparsely populated, see Figure 6.8A. This indicates that the conformational dynamics of the two systems is similar. Contrary to that, more differences in the dynamics between WT and $Ala134N_3$ arise as Figure 6.8B shows. Finally, the ΔDCCM map between ligand-free ($Ala134N_3$) and ligand-bound ($Ala134N_3$-PhCN) labelled lysozyme at Ala134 shown in Figure 6.8C demonstrates that the "contrast" further increases. The major difference in the conformational dynamics between the ligand-free and ligand-bound protein arises for coupled residues $[57, 65]$ with $[105, 120]$ (feature A), $[103, 112]$ with $[112, 122]$ (feature B), $[60, 68]$ with $[47, 52]/[55, 62]$ (feature C), and $[57, 65]$ with $[17, 25]/[32, 42]$ (feature D). Interestingly, as mentioned before for residue 129, residues involved in features A and B are also among those with higher RMSF, see Figure 6.2. Additional ΔDCCM plots for the remaining $AlaN_3$ residues are shown in Figures 6.9 to 6.21.

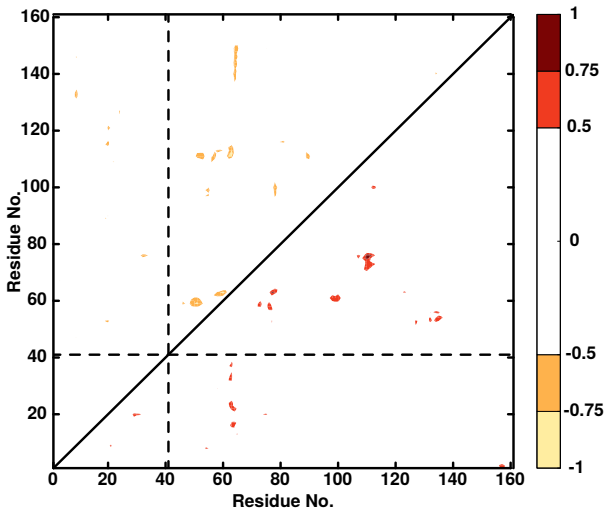


Figure 6.9: Difference dynamic cross correlation maps (ΔDCCM) between $Ala41N_3$ and $Ala41N_3$-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

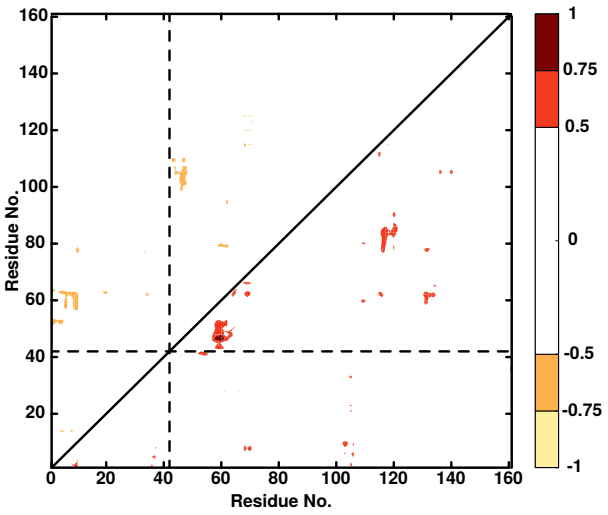


Figure 6.10: Difference dynamic cross correlation maps (ΔDCCM) between Ala42N_3 and Ala42N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

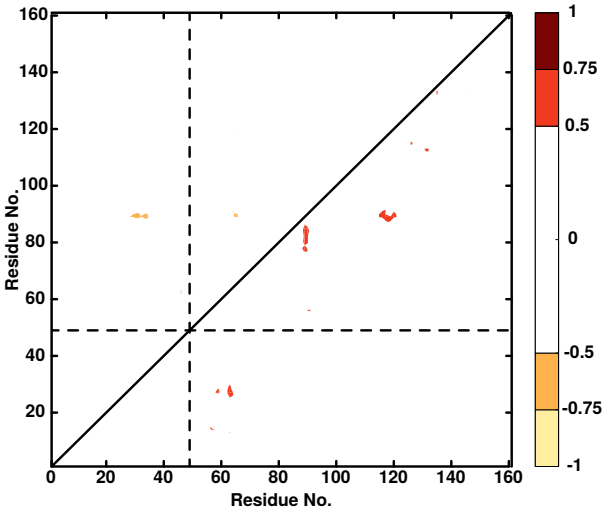


Figure 6.11: Difference dynamic cross correlation maps (ΔDCCM) between Ala49N_3 and Ala49N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

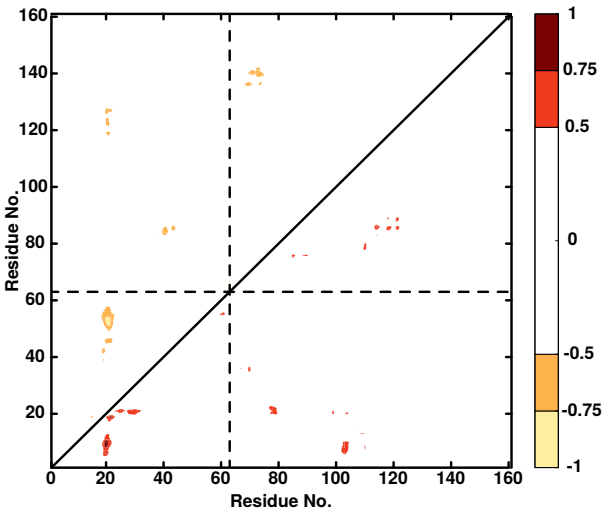


Figure 6.12: Difference dynamic cross correlation maps (ΔDCCM) between Ala63N_3 and Ala63N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

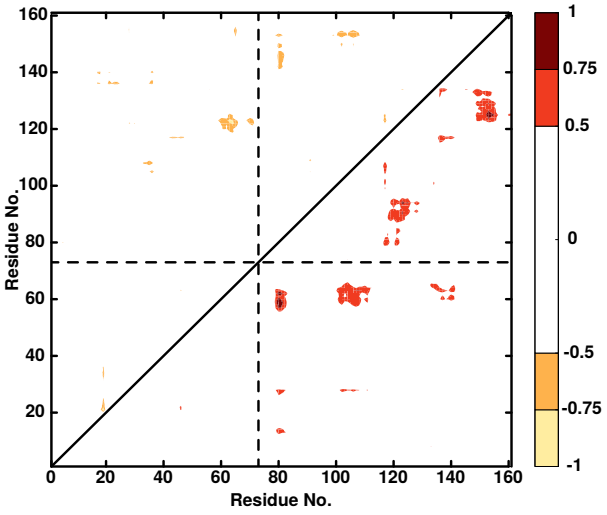


Figure 6.13: Difference dynamic cross correlation maps (ΔDCCM) between Ala73N_3 and Ala73N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

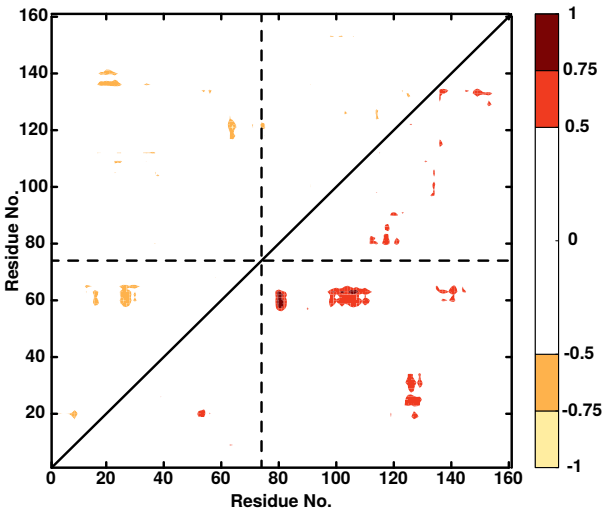


Figure 6.14: Difference dynamic cross correlation maps (ΔDCCM) between Ala74N_3 and Ala74N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

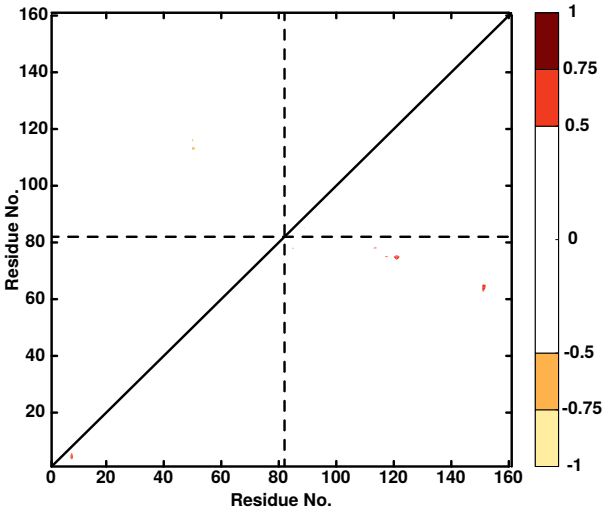


Figure 6.15: Difference dynamic cross correlation maps (ΔDCCM) between Ala82N_3 and Ala82N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

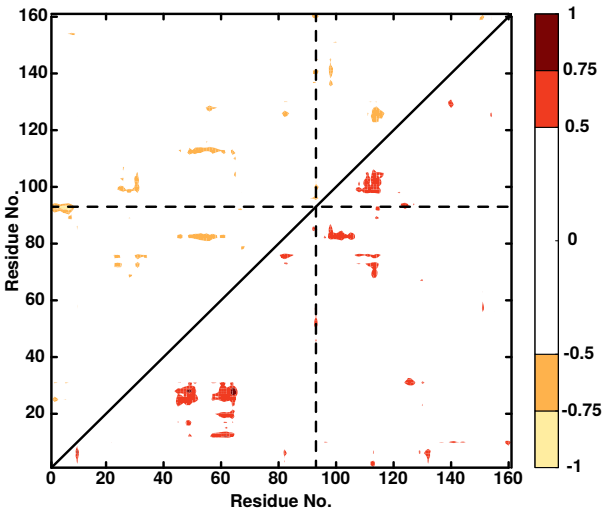


Figure 6.16: Difference dynamic cross correlation maps (ΔDCCM) between Ala93N_3 and Ala93N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

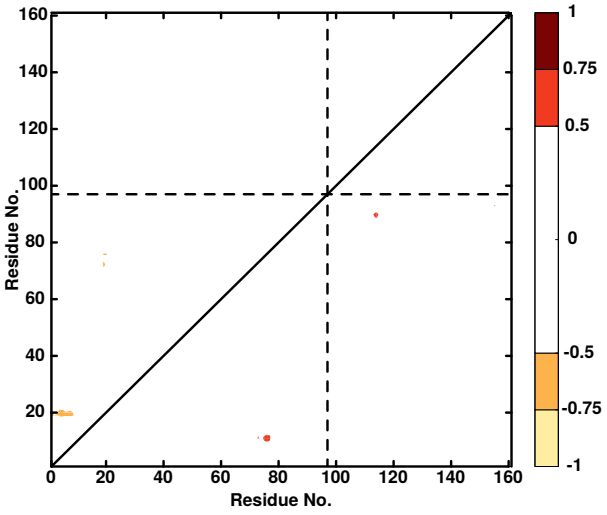


Figure 6.17: Difference dynamic cross correlation maps (ΔDCCM) between Ala97N_3 and Ala97N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

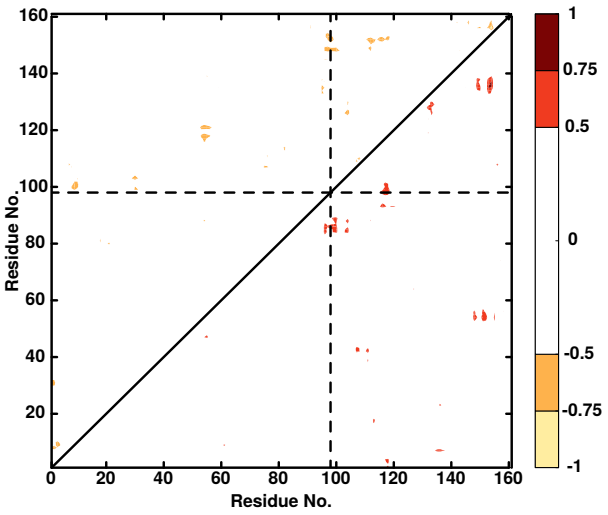


Figure 6.18: Difference dynamic cross correlation maps (ΔDCCM) between Ala98N_3 and Ala98N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

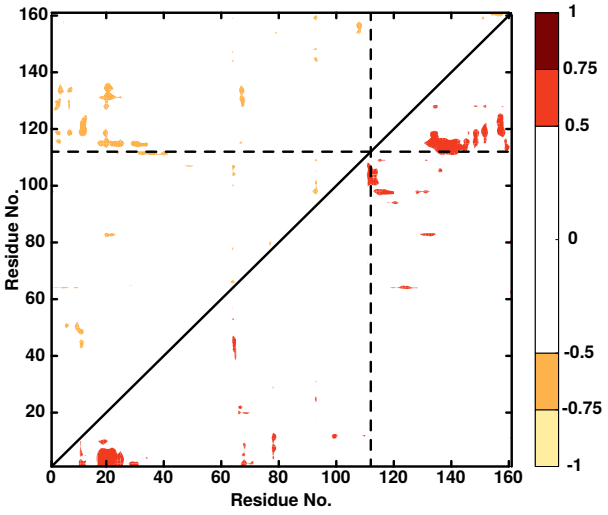


Figure 6.19: Difference dynamic cross correlation maps (ΔDCCM) between Ala112N_3 and Ala112N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

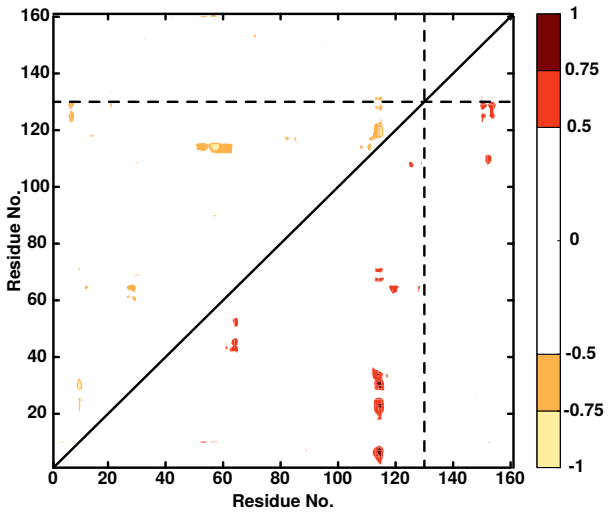


Figure 6.20: Difference dynamic cross correlation maps (ΔDCCM) between Ala130N_3 and Ala130N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

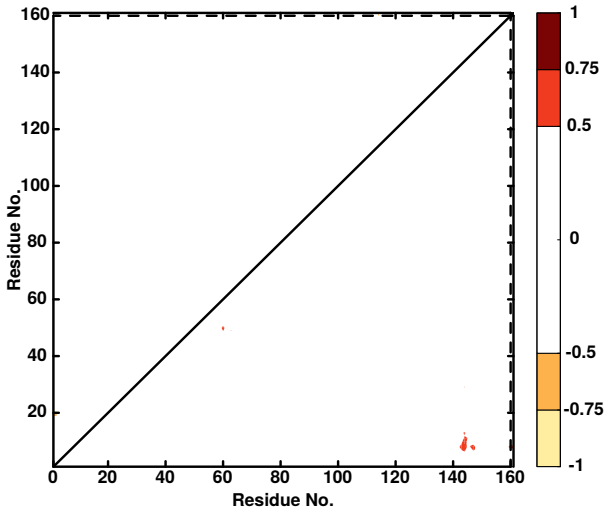


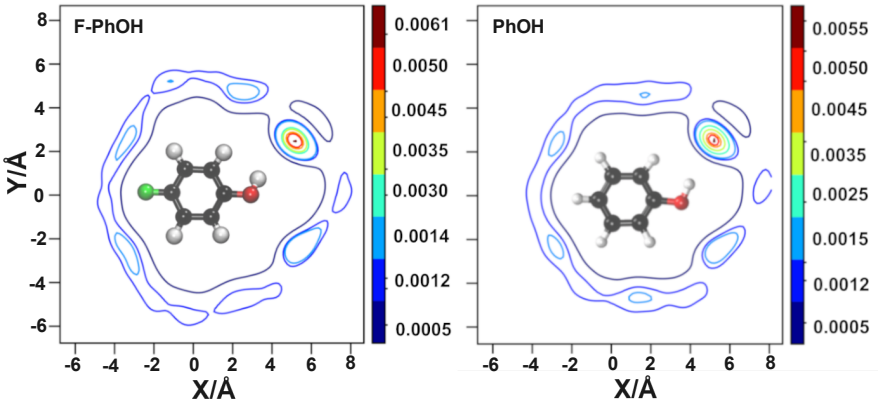
Figure 6.21: Difference dynamic cross correlation maps (ΔDCCM) between Ala160N_3 and Ala160N_3-PhCN. Positive correlations are in the lower right triangle, negative correlations in the upper left triangle. Only correlation coefficients with an absolute value greater than 0.5 are displayed.

In summary, the present work demonstrates that the 1D and 2D IR spectroscopy of azide bound to alanine residues in WT Lysozyme provides valuable site-specific and temporal information about ligand binding of PhCN to the active site of WT lysozyme. Of particular note is the increase in contrast between the ligand-free

and the ligand-bound protein when the azido-label is present, as demonstrated for Ala134N_3. Furthermore, the static component Δ_0 of the FFCF, which is an experimentally accessible observable, shows pronounced differences between the ligand-bound and ligand-free protein and can serve as a useful indicator for ligand binding. Changes in the maximum of the infrared absorbance are of the order of one to several cm^{-1} which is still detectable with state-of-the-art experiments.[22] Given that the -N_3 label can be introduced at multiple positions along the polypeptide chain with specific spectroscopic signatures for each variant of the system, it may even be possible to use the present approach to refine existing structural models based on NMR measurements[230] or from more conventional co-crystallization and X-ray structure determination efforts.

Chapter 7

Hydration Dynamics and 1D/2D Spectroscopy of 4-Fluorophenol



The results presented in this this chapter are achieved in collaboration with:

Silvan Käser and Dr. Kai Töpfer
University of Basel
(Machine-Learned PES)
***
Dr. Polydefkis Diamantis and Prof. Ursula Röthlisberger
École Polytechnique Fédérale de Lausanne (EPFL)
(QM and QM/MM Simulations)
***
Prof. Dr. Peter Hamm
University of Zurich
(Experimental Infrared Spectroscopy)

7.1 abstract

Halogenated groups are relevant in pharmaceutical applications and potentially useful spectroscopic probes for infrared spectroscopy. In this work, the structural dynamics and infrared spectroscopy of *para*-fluorophenol (F-PhOH) and Phenol (PhOH) is investigated in the gas phase and water. The gas phase and solvent dynamics around F-PhOH and PhOH is characterized by molecular dynamics (MD) simulations using CHARMM force parameterization and also with full *ab initio* (QM) and mixed quantum mechanical/molecular mechanics (QM/MM) simulations.

7.2 Introduction

Fluorination - and halogenation in general - is a common chemical modification for pharmaceuticals. Approximately 20 % of all small molecule drugs used in medicinal chemistry contain F, Cl, Br, or I or a combination thereof. Among these compounds halogenated phenyl rings constitute an important class.[39] Because of the directionality of the interaction, halogenation has emerged as one of the essential chemical modifications in medicinal materials[35–37], and supramolecular chemistry.[231,232] By changing the halogen atom, the interactions with the environment can be tuned and the hydrophobicity around the modification site can be modulated.[36,40–45] The importance of halogenation as a fundamental concept in medicinal chemistry is highlighted by the improved binding affinities of several ligands towards their receptors.[233,234] Recently, halogenation has also been employed in the context of protein modifications, such as for insulin, to fine-tune thermodynamic stability and affinity to the insulin receptor.[52]

A halogen bond "occurs when there is evidence of a net attractive interaction between an electrophilic region associated with a halogen atom in a molecular entity and a nucleophilic region in another, or the same, molecular entity."[235] So basically, halogen atom acts as an electrophile and can form an attractive interaction with a nucleophilic counterpart. Based on the analysis of the molecular surface electrostatic potential (ESP),[236] the "halogen bond" was also associated with a "σ-hole bond"[237] which is a noncovalent interaction between a covalently-bonded halogen atom X and a negative site, e.g. a lone pair of a Lewis base or an anion.[236] Such a "bond" involves a region of positive electrostatic potential,

labeled as σ-hole, on the extension of one of the covalent bonds to the atom. The σ-hole arises as a consequence of the anisotropy of the ESP around the halogen atom. Hence, halogen bonding is a subset of σ-hole interactions. Their features and properties can be fully explained in terms of electrostatics and polarization plus dispersion. The strengths of the interactions generally correlate well with the magnitudes of the positive and negative electrostatic potentials of the σ-hole and the negative site.

Despite their importance, little is known about the energetics and dynamics of protein-ligand complexes involving halogenated, pharmaceutically active compounds. The introduction of fluorine atom into organic molecules can cause major changes in the physico-chemical properties such as their solubility, the chemical reactivity and the biological activity compared to the non-fluorinated analogues.[50] In particular, fluorine often replaces hydrogen in organic molecules but the size and stereoelectronic influences of the two atoms (hydrogen vs. fluorine) are quite different and is often regarded as isosteric substitution.[37] In bio-inorganic and medicinal chemistry, the formation of intermolecular O–H/F–C and N–H/F–C hydrogen bridges was assumed to be important in binding fluorinated compounds to enzyme active sites.[49] Such interactions affect enzyme ligand binding affinity, selectively coupled with the changes in pharmaco-kinetic properties by fluorine substitution.[43] This resulted in a considerably large number of fluorine containing drugs being released for clinical use.[50] The effects of fluorine substitution on the related pharmaco-kinetic properties like lipophilicity, volatility, solubility, hydrogen bonding and steric effects affect the resulting compound binding, absorption, transport and hence the related biological activity.[238] In pharmacological application the replacement H→F is often considered to avoid metabolic transformation due to the high stability of the CF bond. Examples are drugs interacting with P450 for which fluorination has been widely used to block metabolic transformations.[38]

These observations call for a more molecularly refined picture of the energetics and dynamics involving fluorinated model compounds. The present work considers hydrated F-PhOH as a typical representative. Using linear infrared spectroscopy together with computational characterizations at different levels of theory the structural dynamics and spectroscopy of F-PhOH is characterized.

The computations use advanced empirical force fields including multipolar interactions, full *ab initio* (QM) and mixed quantum mechanical/molecular mechanics (QM/MM) simulation and machine-learned (ML) techniques. The infrared spectroscopy from the experiments can be directly compared with the computations. Also, the solvent dynamics is investigated based on frequency fluctuation correlation functions. First, the methods are described. Next, results for the spectroscopy of the CF bond are reported together with radial distribution functions. Finally, the solvent distribution is discussed. At the end, conclusions are drawn.

7.3 Methods

7.3.1 Molecular Dynamics Simulations

All MD simulations were performed with CHARMM[64]. All bonded parameters are based on CGenFF[239] except for the CF and OH bond for which a Morse potential was used to describe their anharmonicity. To that end, a scan along the CF bond is performed at the MP2/aug-cc-pVTZ level starting from an optimized structure of F-PhOH at this level of theory. The energy of 49 points is computed on a grid ranging from $r = 0.75$ Å to $r = 5.55$ Å in increments of 0.1 Å. Then, the energies are fitted to a Morse potential $V(r) = D_0[1 - \exp(-\beta(r - r_0))]^2$ which yields parameters $D_0 = 136.316$ kcal/mol, $r_0 = 1.349$ Å, and $\beta = 1.665$ Å^{-1}. Using similar approach, the calculated Morse parameters for OH bond are $D_0 = 120.234$ kcal/mol, $r_0 = 0.971$ Å, and $\beta = 2.088$ Å^{-1}. To realistically describe the electrostatic interactions, a multipolar (MTP)[118,240] model was also used with MTPs on all heavy atoms up to quadrupoles and point charges for all hydrogen atoms. These parameters were fitted to the electrostatic potential using a fitting environment[70], see Tables A.1 to A.3.

Simulations of F-PhOH and PhOH were carried out in a cubic box of (30^3 Å^3) using TIP3P[117] water molecules (Figure 7.1). Minimization, heating, and equilibration procedures for 40 ps were employed to prepare the system. Production simulations of 5 ns were run in the NVT ensemble at 300 K using a Velocity Verlet[82] integrator a and Nosé Hoover thermostat[119,120]. The timestep was $\Delta t = 1$ fs and every 5 snapshot was recorded. Lennard-Jones interactions were computed with a 12 Å cutoff switched at 10 Å.[124] The electrostatic interactions for the monopoles (point charges) are treated using Particle-Mesh Ewald[125] (PME) with

grid size spacing of 1 Å, characteristic reciprocal length $\kappa = 0.32$ Å^{-1}, and interpolation order 6. All bonds involving hydrogen atoms are constrained via SHAKE algorithm.[123] Additional MD simulations were also carried out for PhOH in water with the same setup as was used for F-PhOH in order to directly compare their spectroscopy and solvent structure around them.

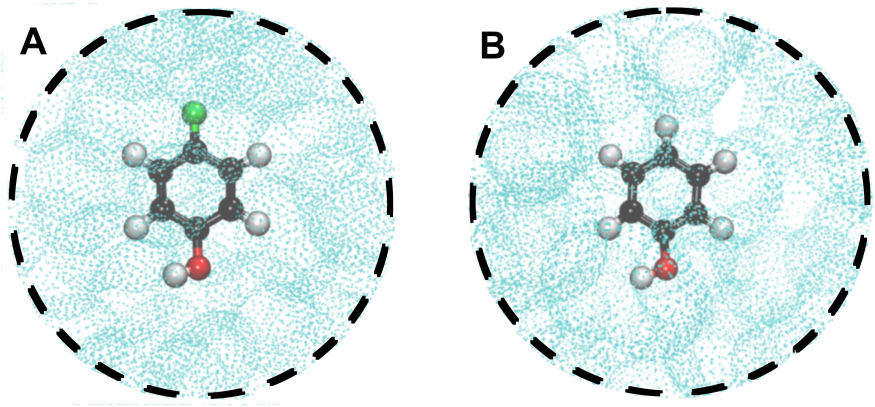


Figure 7.1: The simulation systems used in the present work. F-PhOH (panel A) and PhOH (panel B) are displayed as CPK and water molecules are shown as solvent representation.

7.3.2 Instantaneous Normal Mode

From the production simulation, 10^6 snapshots are taken as a time-ordered series for computing the frequency fluctuation correlation function (FFCF). The FFCF was determined from instantaneous harmonic vibrational frequencies based on a normal mode analysis. Such instantaneous normal modes (INM) are obtained by minimizing F-PhOH while keeping the surrounding solvent frozen. Next, normal modes were calculated in CHARMM for 5 modes in the range of 1100 to 1400 cm^{-1} with ($\nu_2 > \nu_5 > \nu_3 > \nu_4 > \nu_1$) in terms of participation ration of CF stretch in that particular mode. In a separate analysis step, the participation ratios of the CF, CO, and CH stretch and the COH bending coordinates to these 5 normal modes were determined.

7.3.3 Frequency Fluctuation Correlation Function and Lineshape

From the INMs the frequency trajectory $\omega_i(t)$ and the FFCF, $\langle\delta\omega(0)\delta\omega(t)\rangle$ is computed. Here, $\delta\omega(t) = \omega(t) - \langle\omega(t)\rangle$ and $\langle\omega(t)\rangle$ is the ensemble average of the transition frequency. From the FFCF the line shape function

$$g(t) = \int_0^t \int_0^{\tau'} \langle\delta\omega(\tau'')\delta\omega(0)\rangle d\tau'' d\tau'. \tag{7.1}$$

is determined within the cumulant approximation. To compute $g(t)$, the FFCF is numerically integrated using the trapezoidal rule and the 1D-IR spectrum is calculated according to[128]

$$I(\omega) = 2\Re \int_0^\infty e^{i(\omega-\langle\omega\rangle)t} e^{-g(t)} e^{-\frac{t\alpha}{2T_1}} dt \tag{7.2}$$

where $\langle\omega\rangle$ is the average transition frequency obtained from the distribution, $T_1 = 1.2$ ps[241] is the vibrational relaxation time and $\alpha = 0.5$ is a phenomenological factor to account for lifetime broadening.[128]

From the FFCF, the decay time can be determined by fitting the FFCF to a general expression[129]

$$\langle\delta\omega(t)\delta\omega(0)\rangle = \sum_{i=2}^{n} a_i e^{-t/\tau_i} + \Delta_0 \tag{7.3}$$

where a_i, τ_i and Δ_0 are fitting parameters. The decay times τ_i from the fits characterize the time scale of the solvent fluctuations. The absence of a minimum at short times ($\tau \sim 0.02$ ps) indicates that the interaction between F and environment is weak compare with situation in F-ACN or N_3^-.[106,241] The decay times τ_i of the FFCF reflect the characteristic time-scale of the solvent fluctuations to which the solute degrees of freedom are coupled. In all cases the FFCFs were fitted to an expression containing two decay times using an automated curve fitting tool from the SciPy library.[130]

7.3.4 Full QM and Mixed QM/MM Simulations

Full QM Simulations: The QM system was comprised of F-PhOH and 117 water molecules, in a (15.41 Å ,15.44 Å ,15.46 Å) periodic box, and initially

equilibrated classically at 300 K and 1 atm, using CHARMM. The full QM equilibration and production phases that followed with CPMD lasted for 12.5 ps and 20.4 ps respectively. For the gas phase simulation, the total equilibration and production times were 23.0 ps and 28.1 ps, respectively.

For both the gas phase and the condensed phase systems, the full QM simulation protocol consisted of (i) an equilibration of the system at 300 K first with Born-Oppenheimer (BO) MD and then with Car-Parrinello (CP) MD[84], and (ii) a production phase in the microcanonical (NVE) ensemble with CPMD.[242] The respective time steps for BO and CP MD were 10 and 2 atomic units (a.u.), respectively. In CP MD, the fictitious electron mass was equal to 400 a.u. In the production phase, frames were saved every 10 a.u., corresponding to a time interval of approximately 0.48 fs.

Mixed QM/MM Simulations: Two QM/MM MD simulations were carried out for F-PhOH and PhOH in water, respectively, using the QM/MM interface of CPMD with the Gromos code[243] and the coupling scheme developed by Rothlisberger and coworkers.[244–246] The two systems were comprised of the solute (F-PhOH and PhOH respectively), and 331 and 311 water molecules respectively. The system size was selected so that a direct comparison with the full QM simulation of F-PhOH in water can be made, and assess the impact of quantum effects such as solvent polarization in the geometric and spectral properties of F-PhOH.

The systems were first equilibrated classically, using AMBER18.[65] F-PhOH and PhOH were modelled with the GAFF2 force field[247,248], while the TIP3P model was used for the water. Following an initial minimization, the two systems were equilibrated at the isothermal-isobaric (NPT) (300 K, 1 atm) ensemble and then at the NVT ensemble, for a total of 100 ns. A time step of 2 fs was employed. In view of the small periodic box size, a cut-off of 7 Å was used for the nonbonded interactions.

In both systems, the solute was treated at the QM level and the solvent at the classical (MM) level. The QM setup, and the simulation time step were the same as described above for the full QM simulations. The QM/MM MD simulation

protocol was also similar to the one described for the full QM simulation, albeit for the use of two separate Nosé thermostats for the QM and MM parts respectively, during the equilibration with BO and CP MD. For the F-PhOH system, the equilibration and production phases lasted 10.1 ps and 35.9 ps respectively, while for the PhOH system the same phases respectively lasted 12.6 ps and 25.0 ps. During the production phase, frames were saved with the same frequency as in the full QM simulations (0.48 fs).

7.3.5 Machine-Learned PES

To validate in particular the PC- and MTP-based simulations using an empirical force field a complementary model based on a machine-learned PES was also pursued. For this PhysNet[249], a deep neural network (NN) of the message passing type[250], was used to obtain an analytical representation of the potential energy for both PhOH and F-PhOH. PhysNet uses Cartesian coordinates and nuclear charges to learn an atomic descriptor for the prediction of energies, forces, dipole moments, and partial charges to describe chemical systems and their properties, such as infrared spectra.

PhysNet was trained on *ab initio* energies, forces and dipole moments calculated at the MP2/6-31G(d,p) level of theory using Molpro[163] according to the protocol reported in Ref.[249]. The reference data, containing different geometries for both molecules, is generated from MD simulations at 50, 300 and 1000 K using CHARMM force field (5000 geometries each yielding a total of 30000 geometries) and extended with geometries obtained from normal mode sampling[251] at temperatures between 10 and 2000 K (6600 geometries for each molecule). The complete data set thus contains 43200 PhOH and F-PhOH structures.

The MD simulations for the calculation of gas phase IR spectra are run with the atomic simulation environment[252]. For each molecule 1000 trajectories, each 200 ps in length, are run to obtain an ensemble average. The NVE simulations are run at 300 K with a timestep of 0.5 fs, equilibrated for 50 ps and propagated for 200 ps. The IR spectra are then calculated from the dipole-dipole moment autocorrelation function[253–255] and averaged over all 1000 trajectories.

Results

7.3.6 Gas Phase Spectra from the Energy Functions

First, the performance of the PC- and MTP-based empirical force fields, of PhysNet, and of the QM model from the gas phase simulations is assessed. For this, simulations of F-PhOH in the gas phase were carried out and the CF-power and infrared spectra were determined and compared with experiments.

Figure 7.2 compares the infrared and CF-power spectra from simulations in the gas phase with the experimental spectrum in CCl_4[256] in the range between 1100 and 1400 cm^{-1}. For the FF simulations with PC and MTP the bonded terms were adjusted to reproduce the experimental spectrum in CCl_4 (Figure 7.2A), hence the favourable comparison with the IR and power spectra are in Figures 7.2B and C. The corresponding QM power spectrum is shown in Figure 7.2D. The two main peaks are at 1217 and 1256 cm^{-1} and appear to be shifted with respect to the experiments. Also, the peaks are broad - probably due to the comparatively short simulation time of 28.1 ps - so that additional features could be hidden below them. From MD simulations using the PhysNet representation of the MP2/6-31G(d,p) reference data the IR-spectrum also shows two peaks, centered at 1171 and 1264 cm^{-1}. The lower-frequency peak aligns nicely with the experimentally measured one at 1174 cm^{-1} whereas the high-frequency peak is shifted by 38 cm^{-1} to the blue.

FT-IR spectra for F-PhOH, PhOH and its OD-deuterated derivatives have been previously measured[256,257] in carbon tetrachloride (CCl_4) and cyclohexane solutions in the frequency range of 400-3700 cm^{-1}. Moreover, the IR spectrum for PhOH has been obtained in the gas phase,[258,259] in vapor, and in CCl_4 solvent.[258] The analysis found the OH in-plane bending vibration at 1174 cm^{-1} which was confirmed based on the fact that this band disappears in the OD deuterated species.[259] Moreover, the CF stretching mode in F-PhOH was assigned[256] to a signal at 1226 cm^{-1} although this is not a local mode (i.e. isolated CF stretch) but has contributions from other motions. Some of the measured frequencies for particular modes are reported in Table 7.1 for F-PhOH[256] and PhOH[259], respectively, together with the participation ratios to the local modes from com-

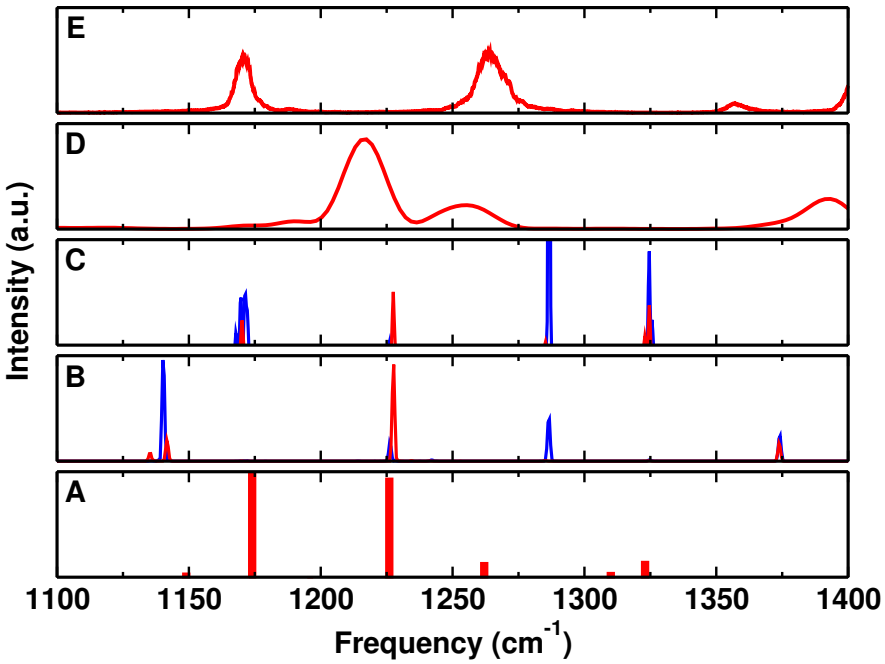


Figure 7.2: Comparison of experimental (in CCl_4) and computed (in the gas phase) spectra for F-PhOH. Panel A: experimental spectrum in CCl_4.[256] Panels B and C: IR (from Fourier transform of total dipole moment correlation function) and CF power spectra from PC (blue) and MTP (red) simulations of F-PhOH in the gas phase. Panel D: CF power spectrum from QM simulations in the gas phase. Panel E: IR spectrum obtained from simulations with PhysNet.

putations based on the calculated potential energy distributions.[256]

PhOH	1150.7	1168.9	1176.5	1261.7	1343	
	δ(CH)	δ(CH)	δ(OH)	ν(CO)	δ(CH)	
	ν(CC)	ν(CC)	δ(CH)	δ(CH)	δ(OH)	
	δ(OH)		ν(CC)			
F-PhOH	1149	1174	1226	1262	1310	1323
	δ(CH)	δ(OH)	ν(CF)	ν(CO)	δ(CH)	ν(CC)
		ν(CC)	δring	ν(CC)	ν(CC)	δ(OH)
		δ(CH)	δ(CH)	ν(CF)		δ(CH)

Table 7.1: Vibrational Frequencies in cm^{-1} for PhOH and F-PhOH in the gas phase and CCl_4 solution in the range of 1100-1400 cm^{-1}.[256,257,259] The contributions (in terms of local deformations) to each vibrational mode indicate strong mixing and are those from the literature.[256,257] The assignment of the bands has been made on the basis of the calculated potential energy distribution.[256,260] Symbols ν and δ refer to stretching and bending modes, respectively

According to the FT-IR experimental results in CCl_4[256], for F-PhOH the CF and CO stretching modes are mainly at 1226 and 1262 cm^{-1}, respectively, while they

couple to one another and potentially to other modes. On the other hand, the CO stretch for PhOH in CCl_4 and in gas phase[257,259] appear at 1257 and 1261.7 cm^{-1}, respectively. This is slightly red shifted (by 1 to 5 cm^{-1}) compared to F-PhOH. According to the reported experimental[256,259] values (see in table 7.1), it seems that the CF stretch is coupled to the in plane bending of the ring and also the C-H bend while the CO stretch is coupled to CC stretching and also CF stretching vibrations. This is also similar in PhOH where the CO stretch is coupled with C-H in plane bending. Therefore, CF stretch is highly coupled with other modes and for that reason it is not possible to assign a local CF stretching mode to a particular frequency.

The simulations so far establish that the coupling between modes in the frequency range of 1100 to 1400 cm^{-1} poses challenges, in particular for simulations with an empirical force field. These challenges are primarily related to the mechanical coupling between stretch and bend vibrations and not so much to the electrostatic interactions. For this reason, an alternative approach was explored by which a neural network-based PES was developed for PhOH and F-PhOH. Again, simulations were carried out in vacuum and the infrared spectra were determined from the dipole moment autocorrelation function, see Figures 7.2E.

The accuracy of the PhysNet model is reported in Figure 7.3 which shows the correlation between Reference MP2 and the PhysNet energies for a set of 3700 randomly selected points averaged over 980/982 trajectories with R^2= 0.9999 and RMSE = 0.0037 eV.

In summary, the gas phase spectrum from finite-T MD simulations find comparable patterns for the frequencies in the 1100 to 1400 cm^{-1} region when compared with experiment. However, achieving more quantitative agreement is challenging. For empirical force fields one of the difficulties is to correctly capture the coupling between modes if this plays a role in the spectroscopy as is the case here. On the other hand, using *ab initio*-based approaches such as QM/MD or ML-learned energy functions (here PhysNet), which do include such coupling, depend strongly on the level of theory at which the calculations are run. For example, it has been reported[261] that ML at the MP2/aug-cc-pVTZ level of theory yields anharmonic frequencies from VPT2 calculations that agree to within $\pm$30 cm^{-1}

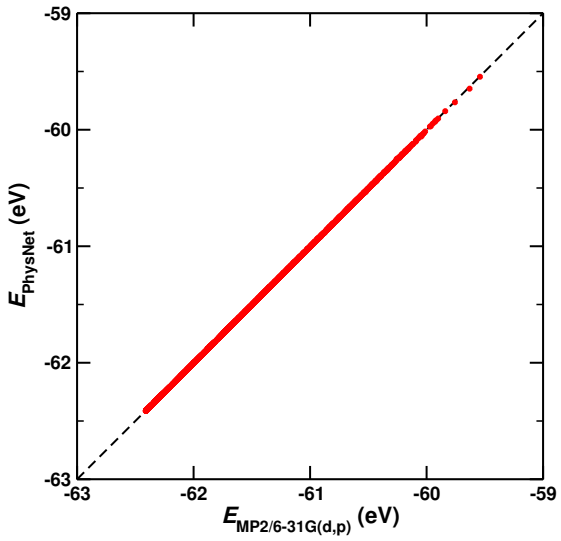


Figure 7.3: Correlation between the *ab initio* and PhysNet energies for a set of 3700 randomly selected points averaged over 980/982 trajectories with $R^2 = 0.9999$ and an root mean square error of 0.0037 eV.

with experiments which is consistent with the present findings. However, when transfer learning such a model to the CCSD(T)/aug-cc-pVTZ level of theory the agreement between experiment and VPT2 calculations improved by an order of magnitude. Hence, it is conceivable that for PhOH and F-PhOH similar improvements could be obtained with reference data at a considerably higher level of theory. However, the calculation of the necessary training set will be extremely time-consuming.

7.3.7 Spectroscopy and Dynamics in Solution

After validating the energy functions considered in the present work, the spectroscopy of F-PhOH and PhOH in water is considered. The experimental spectra for PhOH and F-PhOH in H_2O are reported in Figure 7.4. For clarity, the spectra are split into a low- (Figure 7.4A) and a high-(Figure 7.4B) frequency part. For the spectra between 1100 cm^{-1} and 1400 cm^{-1} pronounced differences between the two compounds are found. Most prominently, the single band with maximum at 1242 cm^{-1} for PhOH is shifted to the red and split into at least two (at 1201 and 1222 cm^{-1}), but possibly several more peaks, some of which overlap with the peak from PhOH. Other features, such as the broader band with peak maximum at 1381 cm^{-1} for PhOH are also shifted to the red (band maximum at 1368 cm^{-1}) for F-PhOH).

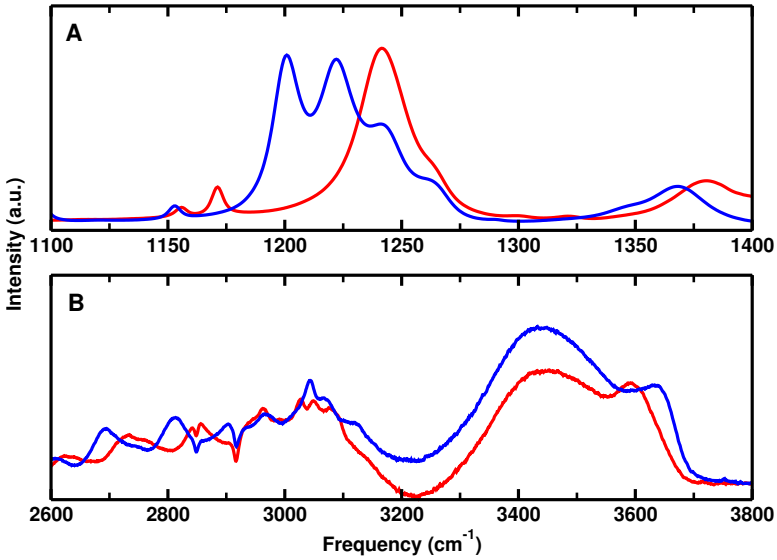


Figure 7.4: Experimental IR spectrum of PhOH (red) and F-PhOH (blue) at a concentration of 0.5M and in H_2O in a CaF_2 cell is shown in the range of (1100-1400) and (2600-3800) cm^{-1}.

For the high-frequency part (Figure 7.4B) the spectra of solvated PhOH and F-PhOH are more similar except for a pronounced absorption at 3596 cm^{-1} in PhOH which shifts to 3643 cm^{-1} upon fluorination. Towards lower frequencies a broad unstructured continuum extends from $\sim$ 3600 down to 3200 cm^{-1}, followed by spectroscopic signatures around (CH stretch) and below 3000 cm^{-1} which is tentatively assigned to the hydrogen-bonded OH stretch in PhOH and F-PhOH.

For the high frequency (OH-stretch) modes early experiments for PhOH vapor, in solution (CCl_4) and as a liquid reported band positions at $\sim$ 3650 cm^{-1}, $\sim$ 3600 cm^{-1}, and at 3500 cm^{-1}, respectively. In solution and the pure liquid an additional band was found at 3350 cm^{-1}.[258] More recently, infrared spectra were recorded for PhOH complexed with variable numbers of water molecules in the gas phase[262,263], in matrices,[264] and for PhOH at the air/water interface using vibrational sum frequency generation (SFG).[265] The cluster studies all report the phenolic-OH stretch vibration at frequencies above 3000 cm^{-1} whereas the experiment at the air/water interface assigns a very broad signature in the SFG signal extending from 2550 cm^{-1} to 3500 cm^{-1} to this mode.[265] This finding is consistent with the present experiments which find a broad absorption for both, F-PhOH and PhOH extending out to $\sim$ 2700 cm^{-1} which is also assigned to the phenol-OH stretch, see Figure 7.4B. In addition, the 1D spectroscopy finds

a sharp peak at 3596 cm^{-1}) which is assigned to the phenol-OH stretch but in a non-hydrogen bonded environment. This is supported by the observation that the peak is only marginally shifted from the PhOH OH-stretch in vapor and the fact that the spectroscopic feature is sharp and therefore can not be due to water.

Figure 7.5 shows the IR spectrum of F-PhOH and PhOH from MD and PhysNet simulations and, consistent with experiment, the lineshape for F-PhOH is more complex than that of PhOH in the 1200 to 1300 cm^{-1}. However, the pronounced double peak structure for F-PhOH between 1200 and 1250 cm^{-1} is shifted to the red of the single band at 1250 cm^{-1} for PhOH in the experiments whereas simulations with the PhysNet PES find the PhOH peak (red) at 1225 cm^{-1}, between the two peaks at 1220 cm^{-1} and 1275 cm^{-1}. It is possible that with training data from reference calculations with a larger basis set, e.g. MP2/aug-cc-pVTZ, the spectroscopy also changes. Similarly, the simulations with MTPs find a pronounced peak for PhOH at 1225 cm^{-1} whereas for F-PhOH one peak to the blue and red are found, see Figure 7.5A. On the other hand, the experiments are carried out at solute concentrations that allow F-PhOH dimers and oligomers to be formed which can affect both, the position of the absorption frequency and the line shape.

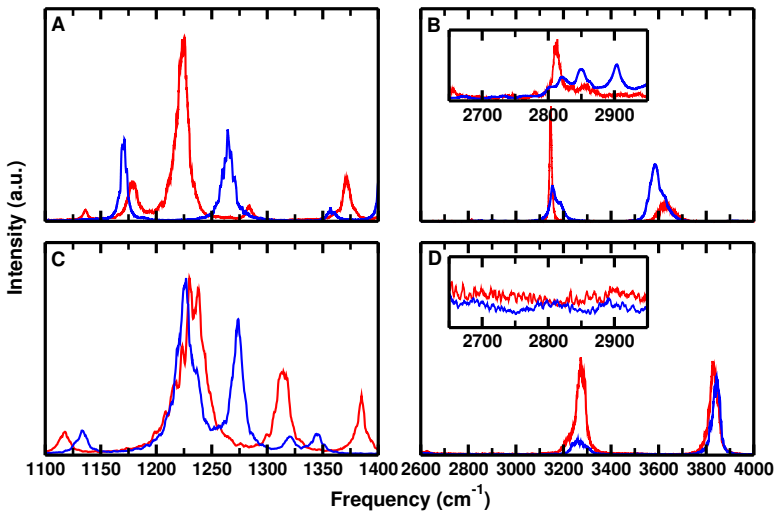


Figure 7.5: Comparison between IR spectrum of PhOH (red) and F-PhOH (blue) in H_2O obtained from MTP MD (panel A and B) and PhysNet (panel C and D) simulations in the range of 1100-1400 cm^{-1} and 2600-4000 cm^{-1}. The insets in panel B and D show a closer representation of the frequencies in the range of 2650-2950 cm^{-1}.

Figure 7.6A/B compares the experimental IR spectrum of F-PhOH in solution with MD and/or PhysNet simulations while 7.6C/D is based on QM/MM and/or QM results. From the MTP/MD simulations IR spectrum shows two prominent peaks at 1171, and 1264 cm^{-1} which are displaced from those observed experimentally. The peak at 1171 cm^{-1} is red shifted compared to double peak at 1201 and 1222 cm^{-1} of experimental spectrum while the peak at 1264 cm^{-1} is blue shifted or captured the same position compared to other two experimental peaks at 1244 and 1264 cm^{-1}. Furthermore, the smaller peak at 1357 cm^{-1} is also red shifted compared to 1368 cm^{-1} from experiment. Comparing PhysNet with experiment the strong feature is at 1226 cm^{-1} but to the blue side additional undulations can be seen which are reminiscent of the experimental finding.

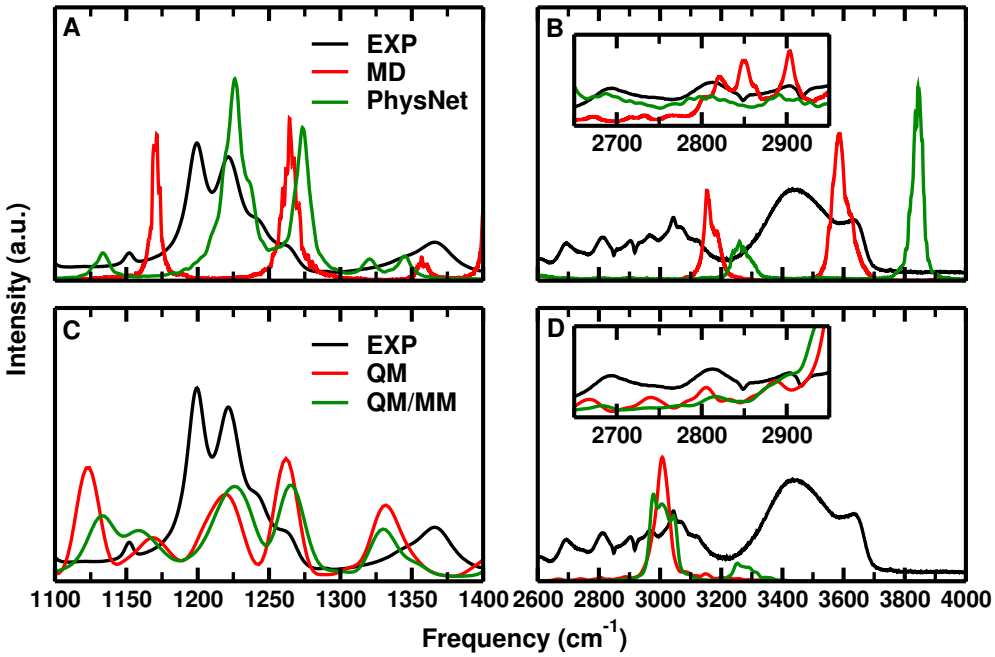


Figure 7.6: Comparison between experimental IR spectrum of F-PhOH in H_2O with MD and PhysNet simulations (panels A and B), and QM/MM and/or QM simulations (panels C and D). The ranges 1100-1400 cm^{-1} and 2600-3800 cm^{-1} are shown separately. Insets in panels B and D show a closer representation of the frequencies in the range of 2650-2950 cm^{-1}.

For the OH-stretch region above 3000 cm^{-1}, the experimental peaks are at 3432 and 3643 cm^{-1}. Furthermore, a broad absorption below 3000 cm^{-1} which also includes the CH stretch modes is found. Computationally, the high frequency peak from PhysNet is at 3843 cm^{-1}, considerably higher than that for experiment whereas that from the MTP simulations appears at approximately the

same position as in the experiment. The broad feature centered at 3432 cm^{-1} is not captured by any of the two methods. Conversely, the broad absorption below 3000 cm^{-1} appears albeit with lower intensity than in the experiments.

For the QM simulations in solution (Figure 7.6C/D and 7.7D) the two prominent bands are at 1207 and 1262 cm^{-1}. However, the number of peaks and their widths between simulations in the gas phase and in solution does not change appreciably. The two prominent bands are also found from QM/MM simulations (Figure 7.6C/D and 7.7D) with band maxima at 1225 and 1266 cm^{-1}. Thus, the splitting between the two peaks decreases for the QM/MM simulations.

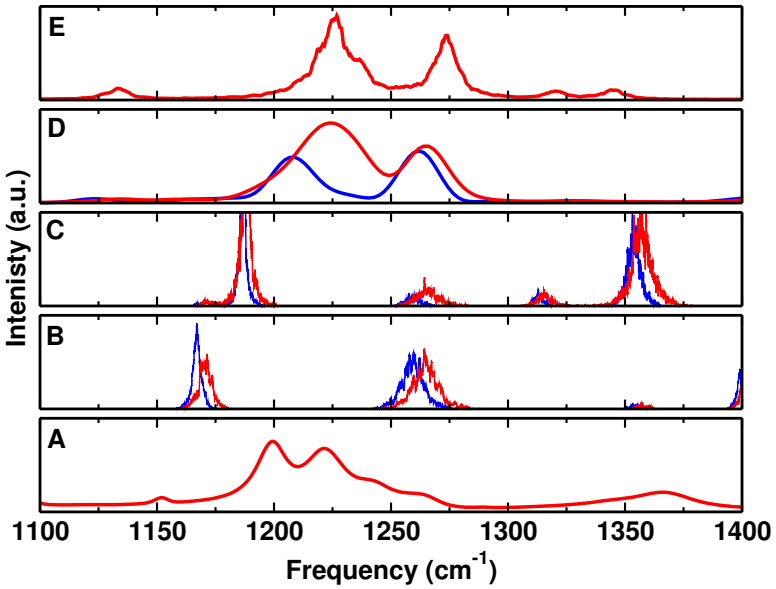


Figure 7.7: Comparison of experimental and computed spectra (in H_2O). Panel A: experimental IR spectrum in H_2O. Panels B and C: IR (from Fourier transform of total dipole moment correlation function) and CF power spectra from PC (blue) and MTP (red) simulations of F-PhOH in solution. Panel I: CF power spectra from QM (blue) and QM/MM (red) simulations in H_2O.

The frequency distribution from instantaneous normal modes (INM) in solution using the PC and MTP model of F-PhOH are compared with the normal modes in the gas phase, see Figures 7.8. The dashed lines indicate the normal mode frequencies in the gas phase and the solid lines the distributions of the frequencies from a 5 ns simulation of hydrated F-PhOH, analyzed with INM. Typically, the frequencies shift to the blue for the simulations in water. For the simulations with MTP the peak maxima for bands ν_1, ν_2, ν_4, and ν_5 are at 1140, 1170, 1286, 1324

cm^{-1} in the gas phase and shift to 1149, 1172, 1294, 1330 cm^{-1} in water, respectively, whereas the ν_3 band has a more complex distribution. For the simulation in water this band appears to consist of at least two contributions: a shoulder at 1236 cm^{-1} and a second, higher, one at 1251 cm^{-1} compared to 1225 cm^{-1} in the gas phase. These band positions compare well with the frequencies from Table 7.1.

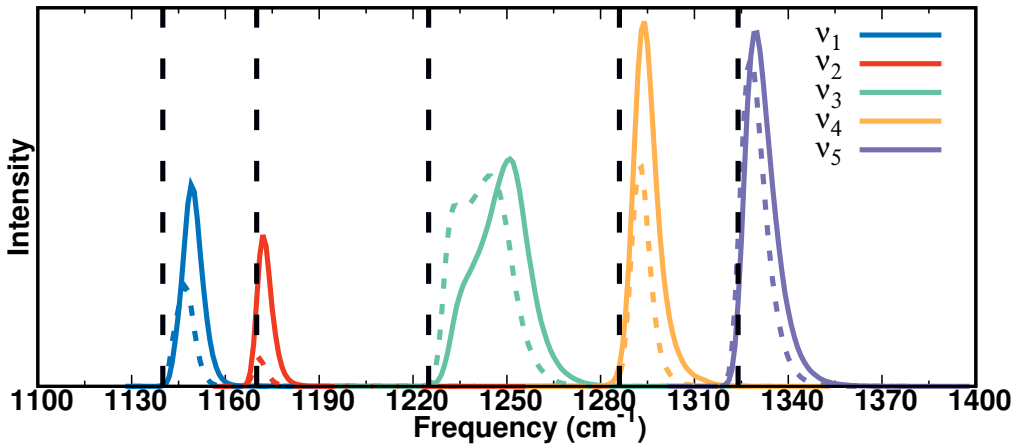


Figure 7.8: Instantaneous vibrational frequency distributions from 5 ns MTP (solid line) and PC (dashed line) simulations of F-PhOH in water for five modes between 1100 and 1400 cm^{-1}. The black dashed line correspond to the frequency of that particular mode of the optimized structure in the gas phase.

Simulations with PC and MTP electrostatics in H_2O lead to solvent induced shifts and broadening of the spectra. The CF-power spectra in solvent are shifted to higher frequencies (blue shift, panel C in Figure 7.2 and 7.7). The maximum peak frequencies are at 1140, 1171, 1226, 1286, 1324 cm^{-1} for simulations with the PC and 1141, 1170, 1227, 1285, 1324 cm^{-1} MTP models, respectively. For the infrared spectra this is less evident (panel B in Figure 7.2 and 7.7). Both, IR and power spectra from PC and MTP simulations lead to differences in the frequency maxima by ~ 5 cm^{-1}. This is also consistent with earlier work on CO in Myoglobin.[99] For the simulations in water the maximum peak frequencies from the power spectra (Figure 7.7C) are at 1163, 1184, 1265, 1310, 1343 cm^{-1} for PC and 1158, 1185, 1248, 1305, 1345 cm^{-1} for MTP model. These are differences between 1 to 17 cm^{-1} for the two models depending on the mode considered and based on the power spectra and compares with 2 to 6 cm^{-1} based on the INM frequency distributions, see Figure 7.8. These findings regarding the expected differences in the absorption frequencies from simulations with PC and MTP

models in solution are consistent with previous work.[104]

Experimentally, the absorption of F-PhOH in water is a broad distribution covering the range between 1200 and 1250 cm^{-1} whereas the frequency in CCl_4 is at 1226 cm^{-1}. This is consistent with the MTP simulations but less so with the PC-based model which find a larger blue shift. On the other hand, the experimental lineshape in the 1200 cm^{-1} region is considerably more structured than the IR spectrum computed from the dipole moment correlation function. Conversely, the frequency distribution for ν_3 in Figure 7.8 appears considerably broader than the other four frequencies but shifted somewhat too far to the blue which may be a consequence of the harmonic approximation made in INM.

It is also of interest to consider the participation ratios of the various local modes to the frequencies in the 1100 to 1400 cm^{-1} range. These were determined from the normal modes of F-PhOH over 10^5 snapshots in solution by freezing the water environment and optimizing the structure of the solute. Then the contributions of the CF, CO, and CH stretch and the COH bending modes to each of the vibrations between 1100 and 1400 cm^{-1} were determined by projection, see Figure 7.9. This analysis confirms that the modes in this frequency range are strongly mixed.

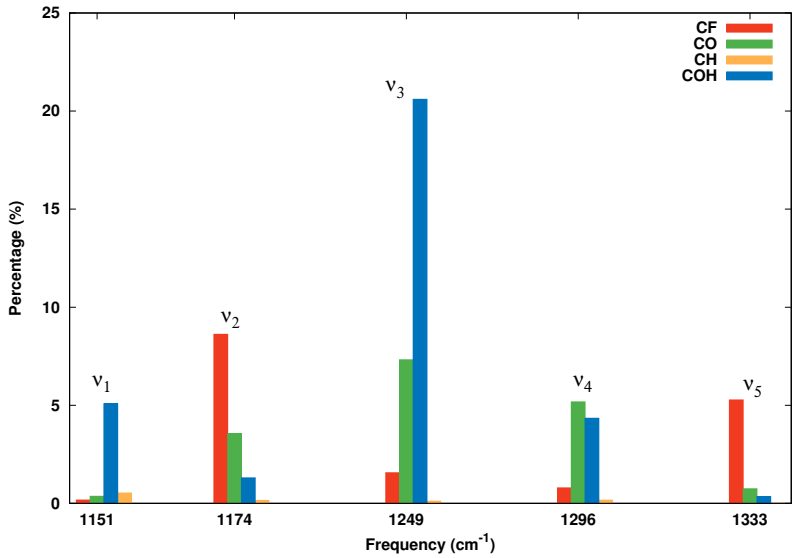


Figure 7.9: Participation ratio of the CF, CO, and CH stretching and the COH bending motions to the 5 modes between 1140 and 1350 cm^{-1} by using the "project" facility in CHARMM for 10^5 snapshots from the MTP simulation of F-PhOH in H_2O. The remaining contributions are from low frequency modes.

In summary, due to the coupling between the modes, computational spectroscopy in the frequency range 1100 to 1400 cm^{-1} for F-PhOH is challenging. None of the methods considered - empirical FF with MTP, ML-PES by PhysNet, or QM/MM and QM MD simulations - provide a consistently satisfactory description of all features compared with experiment. The experiment itself is also challenging as aggregation can occur at the concentrations necessary for recording spectra. However, at a more qualitative level, the computations correctly find that the IR spectroscopy of PhOH and F-PhOH in water differ by the number of absorption bands and that all bands experience blue and red shifts relative to the gas phase.

7.3.8 Frequency Correlation Functions and Solvent Distribution

To characterize the solvent dynamics around F-PhOH the FFCFs for each of the vibrations between 1100 and 1400 cm^{-1} was fitted to a multiexponential decay. The FFCFs (see Figure 7.10) show a biexponential decay with two sub-picosecond time constants and an insignificant static component Δ_0. The decay times of the FFCFs for the modes considered are summarized in Table 7.2. The shapes of the FFCFs are somewhat different from each other and they display several pronounced minima at different correlation times ($tau \sim 1 - 5$ ps) depending on the mode considered. Based on the correlation times, the fast correlation is generally $\tau_1 \sim 0.1$ ps whereas the longer time scale ranges from $\tau_2 = 0.28$ ps to $\tau_2 = 0.83$ ps, see Table 7.2. These correlation times are all comparatively short which points towards weak solvent/solute interactions around the CF-site. The static components are very small which implies a fast dynamics for the system. This is consistent with a less pronounced solvent structure around the CF bond as discussed further below.

It is noted that the FFCFs do not show any recurrence contrary to that of H-bonded systems, as is e.g. found in water.[129,173] Such a recurrence would be typically found around 0.2 ps, is known from previous simulations of CN^-, N_3^-, and N-methyl-acetamide in solution and has been linked to the interaction strength between solute and solvent.[28,104–106] The absence of such a feature indicates that the motions that are involved in these frequencies concern part of the solute that interacts weakly with the solvent and that the solvent organization around this site (i.e. the CF-part or the solute) is expected to be rather unstructured.

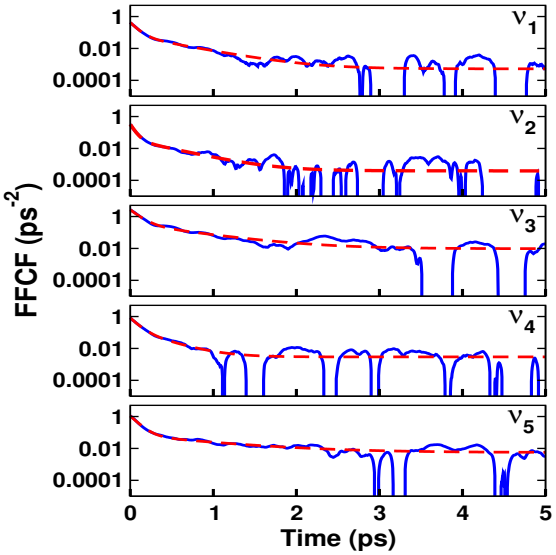


Figure 7.10: FFCFs from INM frequency calculations for F-PhOH in water from a 5 ns simulation using MTP. The FFCFs for the 5 modes ν_1 to ν_5 between 1100 and 1400 cm^{-1} are reported. The solid lines are the raw FFCF data and the dashed lines show the corresponding fit with fitting parameters reported in Table 7.2. Logarithmic scale is chosen for the y-axis.

Mode	$\langle\omega\rangle$	a_1	τ_1	a_2	τ_2	Δ_0
ν_1	1149.71	0.36	0.08	0.06	0.45	0.0005
ν_2	1173.00	0.29	0.05	0.04	0.38	0.0003
ν_3	1248.40	2.31	0.10	0.32	0.60	0.0095
ν_4	1295.18	0.68	0.08	0.14	0.28	0.0029
ν_5	1332.04	1.01	0.09	0.06	0.83	0.0055

Table 7.2: Parameters obtained from fitting the FFCF to Eq. 7.3 based on frequencies from INM using 5 ns MTP (10^6 snapshots) simulation of F-PhOH in H_2O. Average frequency $\langle\omega\rangle$ of the asymmetric stretch in cm^{-1}, the amplitudes a_1 and a_2 in ps^{-2}, the decay times τ_1 and τ_2 in ps, and the offset Δ_0 in ps^{-2}.

The 1D-IR spectra are also determined from numerically integrating the FFCFs, see Eqs. 7.1 and 7.2. The maximum peak positions of the 1D-IR spectra from this analysis and using MTPs in the simulations are at 1151, 1174, 1249, 1296, 1333 cm^{-1}, see Figure 7.11, and at 1148, 1172, 1243, 1294, 1331 cm^{-1} from using PCs in the simulations. Such shifts between simulations with PC and MTP force fields are typical. It is also noted that these frequency maxima are consistent with the analysis based on INM, see Figure 7.8. Compared with the peak maxima from the dipole moment autocorrelation function, these peak positions are displaced between 2 cm^{-1} and 6 cm^{-1} for PC and MTP and also a red shifted by $\sim$ 27 cm^{-1} (see Figures 7.2 and 7.7).

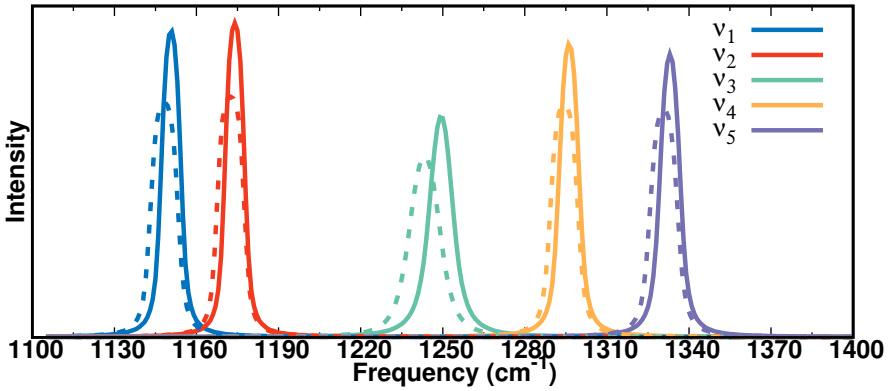


Figure 7.11: The line shapes from 5 ns simulations for F-PhOH in H_2O using PC (dashed) and MTP (solid) models for bands between 1100 and 1400 cm^{-1}. The shifts due to the two different electrostatic models are small, but noticeable and range from 2 to 6 cm^{-1}.

Radial Distribution Functions ($g(r)$): The radial distribution functions were determined using the $g(r)$ plugin from the VMD[266] software. For PC/MTP simulations, Figures 7.12A/B, 7.13A/B and for QM- and QM/MM-derived simulations, Figure 7.12C/D, 7.13C/D are respectively show g(r) function for the (fluorine, water oxygen and Hydrogen), (F-PhOH oxygen, water oxygen and hydrogen) and (F-PhOH hydrogen, water oxygen and hydrogen) pairs of the F-PhOH system. For the water-oxygen around the fluorinated end (Figure 7.12A) the radial distribution functions from PC and MTP simulations indicate that there is a first hydration shell with a maximum around 2.5 Å and 3.0 Å, respectively. The maximum is more pronounced for the PC simulation as is the first minimum, compared with the simulation using MTPs. The differences between the two models are yet more pronounced when considering the $g(r)$ between water-hydrogen atoms and the fluorine atom, see Figure 7.12B. Moreover, the radial distribution function between the water-oxygen atoms and the two carbon atoms flanking the COH group in PhOH from a 5 ns MTP simulation is shown in Figure 7.14. It is demonstrated that the solvent distribution is converged.

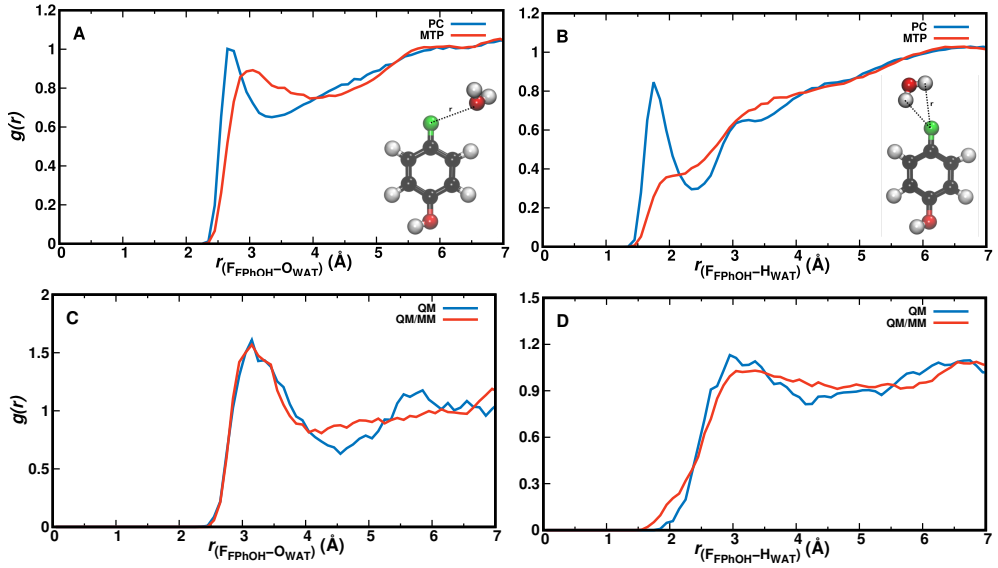


Figure 7.12: The $g(r)$ for A) F—O_W and B) F-–H_W distances as obtained from a 5 ns production run using PC (blue line) and MTP (red line) for F-PhOH in H_2O. Using PCs both $g(r)$ are more structured. Panel C and D show similar results based on QM and QM/MM-derived simulation.The scaling along all $x-$axes is identical whereas that along the $y-$axis is not.

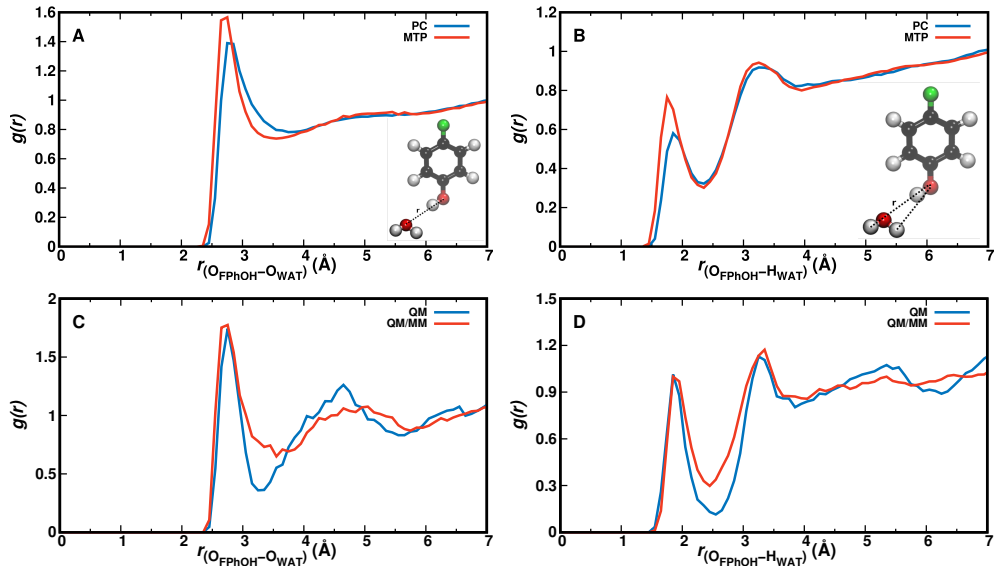


Figure 7.13: The $g(r)$ for A) O—O_W and B) O-–H_W distances as obtained from a 5 ns production run using PC (blue line) and MTP (red line) for F-PhOH in H_2O. Panel C and D show similar results based on QM and QM/MM-derived simulation.The scaling along all $x-$axes is identical whereas that along the $y-$axis is not.

The 2-dimensional solvent distribution functions were generated from the positions of the water-oxygen atoms around F-PhOH. For that, the structures of the 5000 snapshots analyzed were oriented with C1 in the origin, the C1–C4 bond

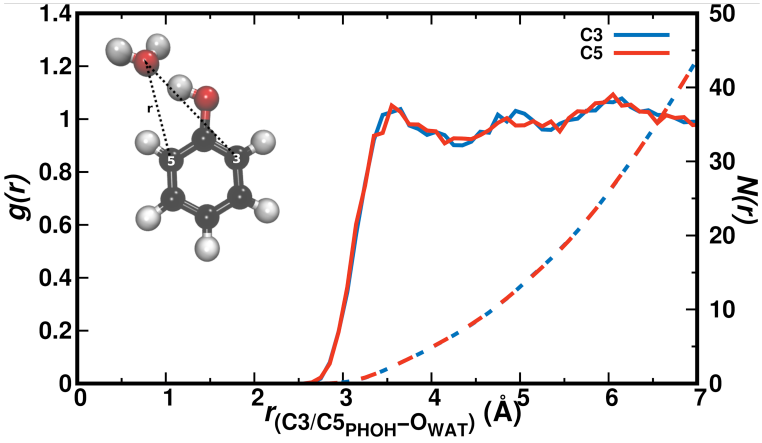


Figure 7.14: The radial distribution function between the water-oxygen atoms and the two carbon atoms flanking the COH group in PhOH from a 5 ns simulation with MTP. It is demonstrated that the solvent distribution is converged.

along the x-axis the [C1,C4,H] atoms in the $xy-$plane. A 2-dimensional histogram of the water positions was generated and then refined from kernel density estimation using Rstudio.[267] The distribution of the solvent water around the CF-part of F-PhOH is comparatively flat for all simulations with PC, MTP and PhysNet when contrasted with the COH-end of the solute, see Figure 7.15. It is also found that for PhOH the solvent distributions are similar to those for F-PhOH.

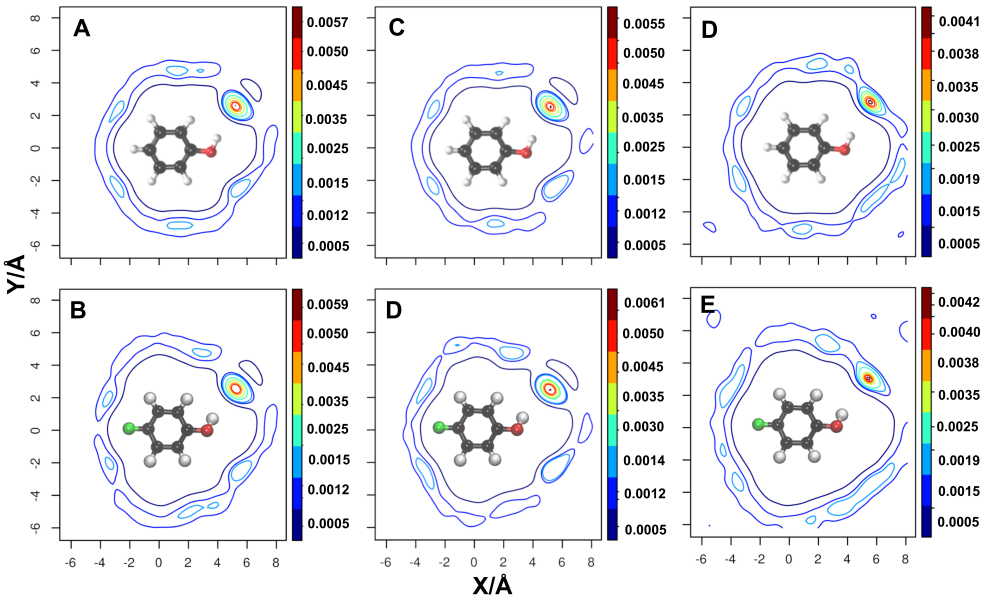


Figure 7.15: Solvent distribution based on Oxygen atoms around PhOH and F-PhOH for PC (panel A and B), MTP (panel C and D) and from PhysNet (panel E and F) model. The USO contour values are shown in each panel.

7.4 Conclusion

Computational spectroscopy in the frequency range of 1100 to 1400 cm^{-1} is challenging due to the coupling between the modes. Findings from MD simulations in the gas phase are comparable with experiment. However, since empirical force field has difficulty to obtain the correct couplings, gaining quantitative agreement is demanding. On the other hand, although *ab initio* methods such as QM/MD or ML-based energy function using PhysNet can include coupling, the the degree of precision strongly depends on the level of theory that is used for the calculation. Thus, higher accuracy is achievable at a noticeably higher level of theory which is computationally expensive. Comparing experimental results in water with computational methods - multipolar force field, ML-based PES with PhysNet, or QM/MM and QM MD - they are not consistently in good agreement with each other. Note that experiment itself is also problematic due to the possible aggregation at the same concentration of recording spectra. However, computational results qualitatively captures correct spectroscopic signatures of PhOH and F-PhOH in water in terms of the difference between absorption bands and the blue or red shifts due to the solvent effects.

Structural dynamics analysis such as FFCF show no recurrence in contrast to systems with H-bonding which implies weak interactions between solute and solvent for the motions involved in these frequencies.[129,173] Moreover, the CF-part in solute is expected to be surrounded by unstructured solvent organization. Given all of the above together with the findings from solvent distribution which show similar pattern for F-PhOH and PhOH, the present work demonstrates at a molecular level that the local hydration of CH and CF are very similar, supporting the notion that "F is a large H-atom".[38]

Chapter 8

Conclusion and Outlook

Vibrational spectroscopy is an invaluable tool for physicochemical assay, characterization of the structure and function of the biological system, and for probing the local environment of molecules. However, IR spectroscopy is challenged by the intrinsic complexity of biological systems which leads to enormous spectral congestion and difficulty in absorption bands analysis and discrimination. Thus, to alleviate this issue and provide more findings with pharmaceutical applications, this study is focused on the development of different IR probes. In chapter three, using all -CO labels in the wild type and mutant insulin (monomer and dimer), the dynamics and infrared spectroscopy are characterized. Distinct spectral responses for residues along the dimerization energy were observed. Utilizing three different approaches ("scan", "INM", "map") for calculating the frequency trajectories, it is confirmed that the overall results are similar to each other and almost independent of the computational approach. Moreover, WT and mutant monomer and dimer and also the monomers in the dimer have different spectroscopic signatures and dynamics, although the crystal structure of the dimer includes symmetric homodimer. Considering the qualitative and quantitative agreement of current findings with experimental results, this computational model together with earlier studies[105,116,127] has the potential to be used for characterizing the aggregation state and dimerization energy of modified insulin.

In chapter four, an accurate computational model is introduced to investigate the vibrational spectroscopy of N_3^- in the gas phase and solution. The model is based on accurate electronic structure calculations together with a reproducing kernel

Hilbert space representation of the potential energy surface. Fundamental vibrations are calculated using instantaneous normal mode analysis and by solving the 3D Schrödinger equation. The fundamentals, FWHM, decay time and shifts due to solvent effect are in good agreement with the experiment. To further probe the coupling between intramolecular degrees of freedom, a worthwhile prospect is to use a flexible water model such as KKY model.[175–177] Additionally, higher-order multipoles[98,99,178,179] or polarization[180] can be used to further improve the electrostatic representation.

Given the above-mentioned computational model, N_3^- can be bound to a peptide or a protein residue to probe the structural dynamics and spectroscopic responses of the system, which is shown in chapter four for all azido-modified alanine residues in lysozyme protein. The findings show that $AlaN_3$ is minimally invasive compared to a reference structure, and locally sensitive to the environment. The results are consistent with experimental findings for the selective substitutions of amino acids in PDZ2 by AHA.[23] Moreover, the offsets which correlate with degrees of hydration and long-time decay constants obtained from FFCF are comparable with experiments.[23] Thus, $AlaN_3$ is a promising probe and worthwhile modification for the site-specific investigation of protein structure and dynamics.

To further investigate the application of the $AlaN_3$ probe, it is used to query whether or not a ligand has bound to the active site of the protein, which is illustrated in chapter six. To that end, the variations in spectroscopy and dynamics of the modification site upon ligand binding are investigated. The results demonstrate pronounced differences between the ligand-free and the ligand-bound when the azide is incorporated into the protein. Thus, the covalently linked azide group to alanine is an environment-sensitive probe that can provide invaluable insights into energetics and dynamics of ligand-protein binding. Furthermore, it may be applicable to refine the existing structural determination models.[230]

In future research, -SCN reporters can also be incorporated into peptides, or proteins, as it has an exceptionally long vibrational time, owing to the insulating effect of heavy S atom which leads to control of intra-molecular relaxation over

inter-molecular vibrational relaxation in SCN.[59] The incorporation of an -SCN moiety gives rise to absorption band due to its relatively large extinction coefficient.[59] Additionally, it is sensitive to its environment and can act as a site-specific electric field probe for proteins.[58]

Chapter seven is focused on the structural dynamics and infrared spectroscopy of F-PhOH which is a pharmaceutically relevant compound due to its halogen group. Applying three different methods, empirical force field, ML-based parameterization using PhysNet, and QM/MD simulations, the results are qualitatively comparable with the experiment, albeit good quantitative agreement is challenging due to the high coupling between the modes in the frequency range of 1100 to 1400 cm^{-1}. However, improvements could be obtained using a considerably higher level of theory. Based on the results, this work supports the idea that fluorine acts as a large H atom.[38] As a future perspective, similar studies can be done to investigate the effect of substitution pattern (ortho, meta, para) on spectroscopic signatures and dynamics of halogenated compounds.

Given the above findings, computational infrared spectroscopy is a powerful tool to interrogate a wide variety of biological queries. Introducing side chain and site-specific backbone vibrational probes with spectrally isolated absorptions alleviates the problem of the congested area in the protein spectra and can be utilized to investigate the structural and environmental properties. The various moieties and vibrational modes mentioned above all have special merits, and demerits, so the selection of a specific reporter depends on several items such as the aim of the work and system of interest. The future achievements depend on the application of various vibrational probes, but can hopefully address intriguing biological questions.

Bibliography

[1] X. Pang and H.-X. Zhou, *Annu. Rev. Biophys.*, 2017, **46**, 105–130.

[2] J. McCammon, *Curr. Opin. Struct. Biol.*, 1998, **8**, 245–249.

[3] J. Guo and H.-X. Zhou, *Chem. Rev.*, 2016, **116**, 6503–6515.

[4] S. Lu, X. He, D. Ni and J. Zhang, *J. Med. Chem.*, 2019, **62**, 6405–6421.

[5] J. M. Chalmers and P. R. Griffiths, *Handbook of Vibrational Spectroscopy*, John Wiley & Sons: New York, 2002.

[6] R. Adhikary, J. Zimmermann and F. E. Romesberg, *Chem. Rev.*, 2017, **117**, 1927–1969.

[7] J. Ma, I. M. Pazos, W. Zhang, R. M. Culik and F. Gai, *Annu. Rev. Phys. Chem*, 2015, **66**, 357–377.

[8] H. Kim and M. Cho, *Chem. Rev.*, 2013, **113**, 5817–5847.

[9] N. M. Levinson and S. G. Boxer, *Nat. Chem. Biol.*, 2014, **10**, 127–132.

[10] J. P. Layfield and S. Hammes-Schiffer, *J. Am. Chem. Soc.*, 2013, **135**, 717–725.

[11] D. G. Kuroda, J. D. Bauman, J. R. Challa, D. Patel, T. Troxler, K. Das, E. Arnold and R. M. Hochstrasser, *Nat. Chem.*, 2013, **5**, 174–181.

[12] J. Liu, J. Strzalka, A. Tronin, J. S. Johansson and J. K. Blasie, *Biophys. J.*, 2009, **96**, 4176–4187.

[13] S. H. Brewer, B. Song, D. P. Raleigh and R. B. Dyer, *Biochem.*, 2007, **46**, 3279–3285.

[14] 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, *Nat. Chem.*, 2012, **4**, 355–360.

[15] A. Ghosh, J. Qiu, W. F. DeGrado and R. M. Hochstrasser, *Proc. Natl. Acad. Sci.*, 2011, **108**, 6115–6120.

[16] A. Remorino, I. V. Korendovych, Y. Wu, W. F. DeGrado and R. M. Hochstrasser, *Science*, 2011, **332**, 1206–1209.

[17] L. J. G. W. van Wilderen, D. Kern-Michler, H. M. Mueller-Werkmeister and J. Bredenbeck, *Phys. Chem. Chem. Phys.*, 2014, **16**, 19643–19653.

[18] G. Lee, D. Kossowska, J. Lim, S. Kim, H. Han, K. Kwak and M. Cho, *J. Phys. Chem. B*, 2018, **122**, 4035–4044.

[19] M. Kozinski, S. Garrett-Roe and P. Hamm, *J. Phys. Chem. B*, 2008, **112**, 7645–7650.

[20] J. Zimmermann, M. C. Thielges, W. Yu, P. E. Dawson and F. E. Romesberg, *J. Phys. Chem. Lett.*, 2011, **2**, 412–416.

[21] A. M. Woys, S. S. Mukherjee, D. R. Skoff, S. D. Moran and M. T. Zanni, *J. Phys. Chem. B*, 2013, **117**, 5009–5018.

[22] S. Ramos, R. E. Horness, J. A. Collins, D. Haak and M. C. Thielges, *Phys. Chem. Chem. Phys.*, 2019, **21**, 780–788.

[23] R. Bloem, K. Koziol, S. A. Waldauer, B. Buchli, R. Walser, B. Samatanga, I. Jelesarov and P. Hamm, *J. Phys. Chem. B*, 2012, **116**, 13705–13712.

[24] L. Miller, G. Smith and G. Carr, *J. Biol. Phys.*, 2003, **29**, 219–230.

[25] Z. Getahun, C. Huang, T. Wang, B. De Leon, W. DeGrado and F. Gai, *J. Am. Chem. Soc.*, 2003, **125**, 405–411.

[26] M. Reppert and A. Tokmakoff, in *Annu. Rev. Phys, Vol 67*, ed. Johnson, MA and Martinez, TJ, 2016, vol. 67 of Annu. Rev. Phys. Chem., pp. 359–386.

[27] L. Wang, C. T. Middleton, M. T. Zanni and J. L. Skinner, *J. Phys. Chem. B*, 2011, **115**, 3713–3724.

[28] S. Li, J. R. Schmidt, A. Piryatinski, C. P. Lawrence and J. L. Skinner, *J. Phys. Chem. B*, 2006, **110**, 18933–18938.

[29] J. S. Lipkin, R. Song, E. E. Fenlon and S. H. Brewer, *J. Phys. Chem. Lett.*, 2011, **2**, 1672–1676.

[30] P. J. M. Johnson, K. L. Koziol and P. Hamm, *J. Phys. Chem. Lett.*, 2017, **8**, 2280–2284.

[31] C. Zanobini, O. Bozovic, B. Jankovic, K. L. Koziol, P. J. M. Johnson, P. Hamm, A. Gulzar, S. Wolf and G. Stock, *J. Phys. Chem. B*, 2018, **122**, 10118–10125.

[32] M. P. Wolfshorndl, R. Baskin, I. Dhawan and C. H. Londergan, *J. Phys. Chem. B*, 2012, **116**, 1172–1179.

[33] K.-I. Oh, J.-H. Lee, C. Joo, H. Han and M. Cho, *J. Phys. Chem. B*, 2008, **112**, 10352–10357.

[34] I. Suydam and S. Boxer, *Biochemistry*, 2003, **42**, 12050–12055.

[35] M. Z. Hernandes, S. M. T. Cavalcanti, D. R. M. Moreira, J. de Azevedo, W. Filgueira and A. C. L. Leite, *Curr. Drug Targets*, 2010, **11**, 303–314.

[36] H. Matter, M. Nazaré, S. Güssregen, D. Will, H. Schreuder, A. Bauer, M. Urmann, K. Ritter, M. Wagner and V. Wehner, *Angew. Chem. Int. Ed.*, 2009, **48**, 2911–2916.

[37] K. Müller, C. Faeh and F. Diederich, *Science*, 2007, **317**, 1881–1886.

[38] R. Hevey, *Chem. Eur. J*, 2021, **27**, 2240–2253.

[39] L. N. Herrera-Rodriguez, F. Khan, K. T. Robins and H.-P. Meyer, *Chim. Oggi – Chem. Today*, 2011, **29**, 31–33.

[40] J. P. M. Lommerse, A. J. Stone, R. Taylor and F. H. Allen, *J. Am. Chem. Soc.*, 1996, **118**, 3108–3116.

[41] P. Auffinger, F. A. Hays, E. Westhof and P. S. Ho, *Proc. Natl. Acad. Sci.*, 2004, **101**, 16789–16794.

[42] K. E. Riley and P. Hobza, *Cryst. Growth Des.*, 2011, **11**, 4272–4278.

[43] L. A. Hardegger, B. Kuhn, B. Spinnler, L. Anselm, R. Ecabert, M. Stihle, B. Gsell, R. Thoma, J. Diez, J. Benz, J.-M. Plancher, G. Hartmann, D. W. Banner, W. Haap and F. Diederich, *Angew. Chem. Int. Ed.*, 2011, **50**, 314–318.

[44] K. E. Riley, J. S. Murray, J. Fanfrlik, J. Rezac, R. J. Sola, M. C. Concha, F. M. Ramos and P. Politzer, *J. Mol. Model.*, 2011, **17**, 3309–3318.

[45] K. El Hage, J.-P. Piquemal, Z. Hobaika, R. G. Maroun and N. Gresh, *J. Comp. Chem.*, 2015, **36**, 210–221.

[46] B. Park, N. Kitteringham and P. O'Neill, *Annu. Rev. Pharmacol. Toxicol.*, 2001, **41**, 443–470.

[47] D. Berkowitz and M. Bose, *J. Fluorine Chem.*, 2001, **112**, 13–33.

[48] H. Choo, Y. Chong, Y. Choi, J. Mathew, R. Schinazi and C. Chu, *J. Med. Chem.*, 2003, **46**, 389–398.

[49] T. Barbarich, C. Rithner, S. Miller, O. Anderson and S. Strauss, *J. Am. Chem. Soc.*, 1999, **121**, 4280–4281.

[50] P. Shah and A. D. Westwell, , 2007, **22**, 527–540.

[51] J.-H. Choi and M. Cho, *J. Chem. Phys.*, 2011, **134**, 154513.

[52] K. El Hage, V. Pandyarajan, N. B. Phillips, B. J. Smith, J. G. Menting, J. Whittaker, M. C. Lawrence, M. Meuwly and M. A. Weiss, *J. Biol. Chem.*, 2016, **291**, 27023–27041.

[53] M. M. Waegele, M. J. Tucker and F. Gai, *Chem. Phys. Lett.*, 2009, **478**, 249–253.

[54] C. Huang, T. Wang and F. Gai, *Chem. Phys. Lett.*, 2003, **371**, 731–738.

[55] W. Hu and L. J. Webb, *J. Phys. Chem. Lett.*, 2011, **2**, 1925–1930.

[56] H. A. McMahon, K. N. Alfieri, C. A. A. Clark and C. H. Londergan, *J. Phys. Chem. Lett.*, 2010, **1**, 850–855.

[57] C. G. Bischak, S. Longhi, D. M. Snead, S. Costanzo, E. Terrer and C. H. Londergan, *Biophys. J.*, 2010, **99**, 1676–1683.

[58] A. T. Fafarman, L. J. Webb, J. I. Chuang and S. G. Boxer, *J. Am. Chem. Soc.*, 2006, **128**, 13356–13357.

[59] J. M. Schmidt-Engler, L. Blankenburg, B. Blasiak, L. J. G. W. van Wilderen, M. Cho and J. Bredenbeck, *Anal. Chem.*, 2020, **92**, 1024–1032.

[60] M. Born and R. Oppenheimer, *Annalen der Physics*, 1927, **84**, 0457–0484.

[61] O. T. Unke, D. Koner, S. Patra, S. Kaser and M. Meuwly, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 013001.

[62] F. Jensen, *Introduction to Computational Chemistry*, Wiley, Chichester, 1999.

[63] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*, Dover: Mineola, 1996.

[64] B. R. Brooks, C. L. Brooks III, A. D. MacKerell Jr., 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, E. W. Pastor, J. Z. Post, C. B. andPu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, *J. Comp. Chem.*, 2009, **30**, 1545–1614.

[65] S. J. Weiner, P. A. Kollman, D. A. Case, U. Singh, C. Ghio, G. Alagona, S. Profeta Jr and P. Weiner, *J. Am. Chem. Soc.*, 1984, **106**, 765–784.

[66] W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1988, **110**, 1657–1666.

[67] A. MacKerell, D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586–3616.

[68] J. Alexander D. Mackerell, M. Feig and I. Charles L. Brooks, *J. Comp. Chem.*, 2004, **25**, 1400–1415.

[69] A. Stone, *Chem. Phys. Lett.*, 1981, **83**, 233–239.

[70] F. Hedin, K. El Hage and M. Meuwly, *J. Chem. Theo. Comp.*, 2016, **56**, 1479–1489.

[71] F. Hedin, K. El Hage and M. Meuwly, *J. Chem. Theo. Comp.*, 2017, **57**, 102–103.

[72] G. H. Golub and C. F. Van Loan, *Matrix Computations*, JHU Press: Baltimore, 2012.

[73] T.-S. Ho and R. Rabitz, *J. Chem. Phys.*, 1996, **104**, 2584–2597.

[74] O. T. Unke and M. Meuwly, *J. Chem. Inf. Model.*, 2017, **57**, 1923–1931.

[75] K. Hornik, *Neural Networks*, 1991, **4**, 251–257.

[76] B. Huang and O. A. von Lilienfeld, *arXiv preprint arXiv:1707.04146*, 2017.

[77] O. T. Unke, S. Brickel and M. Meuwly, *J. Chem. Phys.*, 2019, **150**, 074107.

[78] S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

[79] S. A. Hollingsworth and R. O. Dror, *Neuron*, 2018, **99**, 1129–1143.

[80] M. Karplus and J. McCammon, *Nat. Struct. Biol.*, 2002, **9**, 646–652.

[81] A. Leach, *Molecular Modelling: Principles and Applications*, Pearson Education; Prentice Hall, 2001.

[82] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *J. Chem. Phys.*, 1982, **76**, 637–649.

[83] E. Hairer, C. Lubich and G. Wanner, *Acta Numerica*, 2003, **12**, 399–450.

[84] R. Car and M. Parrinello, *Phys. Rev. Lett.*, 1985, **55**, 2471–2474.

[85] A. Kukol, *Molecular modeling of Proteins: 2nd Edition*, Humana Press: Totowa, 2015.

[86] S. Strazza, R. Hunter, E. Walker and D. W. Darnall, *Archives of Biochemistry and Biophysics*, 1985, **238**, 30–42.

[87] B. Tidor and M. Karplus, *J. Mol. Biol.*, 1994, **238**, 405–414.

[88] V. Zoete, M. Meuwly and M. Karplus, *Proteins: Structure, Function, and Bioinformatics*, 2005, **61**, 79–93.

[89] E. Baker, T. Blundell, J. Cutfield, S. Cutfield, E. Dodson, G. Dodson, D. Hodgkin, R. Hubbard, N. Isaacs and C. Reynolds, *Phil. Trans. R. Soc. Lond. B Biol.*, 1988, **319**, 369–456.

[90] X.-X. Zhang and A. Tokmakoff, *J. Phys. Chem. Lett.*, 2020, **11**, 4353–4358.

[91] P. Banerjee, S. Mondal and B. Bagchi, *J. Chem. Phys.*, 2018, **149**, 114902.

[92] A. Antoszewski, C.-J. Feng, B. P. Vani, E. H. Thiede, L. Hong, J. Weare, A. Tokmakoff and A. R. Dinner, *J. Phys. Chem. B*, 2020, **124**, 5571–5587.

[93] S. Raghunathan, K. El Hage, J. L. Desmond, L. Zhang and M. Meuwly, *J. Phys. Chem. B*, 2018, **122**, 7038–7048.

[94] P. Wang, X. Wang, L. Liu, H. Zhao, W. Qi and M. He, *Biophys. J.*, 2019, **117**, 533–541.

[95] P. Banerjee and B. Bagchi, *Proc. Natl. Acad. Sci.*, 2020, **117**, 2302–2308.

[96] I. T. Suydam, C. D. Snow, V. S. Pande and S. G. Boxer, *Science*, 2006, **313**, 200–204.

[97] P. Mondal and M. Meuwly, *Phys. Chem. Chem. Phys.*, 2017, **19**, 16131–16143.

[98] T. Bereau, C. Kramer and M. Meuwly, *J. Chem. Theo. Comp.*, 2013, **9**, 5450–5459.

[99] N. Plattner and M. Meuwly, *Biophys. J.*, 2008, **94**, 2505–2515.

[100] D. Nutt and M. Meuwly, *Biophys. J.*, 2003, **85**, 3612–3623.

[101] T. l. C. Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst and J. Knoester, *J. Chem. Phys.*, 2006, **125**, 044312.

[102] T. l. C. Jansen, A. G. Dijkstra, T. M. Watson, J. D. Hirst and J. Knoester, *J. Chem. Phys.*, 2012, **136**, 209901.

[103] M. Reppert and A. Tokmakoff, *J. Chem. Phys.*, 2013, **138**, 134116.

[104] M. W. Lee, J. K. Carr, M. Göllner, P. Hamm and M. Meuwly, *J. Chem. Phys.*, 2013, **139**, 054506.

[105] P.-A. Cazade, T. Bereau and M. Meuwly, *J. Phys. Chem. B*, 2014, **118**, 8135–8147.

[106] S. M. Salehi, D. Koner and M. Meuwly, *J. Phys. Chem. B*, 2019, **123**, 3282–3290.

[107] Q. X. Hua, S. E. Shoelson, M. Kochoyan and M. A. Weiss, *Nature*, 1991, **354**, 238–241.

[108] S. Shoelson, M. Fickova, M. Haneda, A. Nahum, G. Musso, E. Kaiser, A. Rubenstein and H. Tager, *Proc. Natl. Acad. Sci.*, 1983, **80**, 7390–7394.

[109] M. Haneda, M. Kobayashi, H. Maegawa, N. Watanabe, Y. Takata, O. Ishibashi, Y. Shigeta and K. Inouye, *Diabetes*, 1985, **34**, 568–573.

[110] H. Tager, N. Thomas, R. Assoian, A. Rubenstein, M. Saekow, J. Olefsky and E. Kaiser, *Proc. Natl. Acad. Sci.*, 1980, **77**, 3181–3185.

[111] L. Žáková, E. Kletvíková, V. Veverka, M. Lepšík, C. J. Watson, J. P. Turkenburg, J. Jiráček and A. M. Brzozowski, *J. Biol. Chem.*, 2013, **288**, 10230–10240.

[112] H. Chen, M. Shi, Z.-Y. Guo, Y.-H. Tang, Z.-S. Qiao, Z.-H. Liang and Y.-M. Feng, *Protein Engineering*, 2000, **13**, 779–782.

[113] M. R. DeFelippis, R. E. Chance and B. H. Frank, *Crit Rev Ther Drug Carrier Syst.*, 2001, **18**, 201–264.

[114] A. Mackerell, M. Feig and C. Brooks, *J. Comp. Chem.*, 2004, **25**, 1400–1415.

[115] A. MacKerell, M. Feig and C. Brooks, *J. Am. Chem. Soc.*, 2004, **126**, 698–699.

[116] P.-A. Cazade, F. Hedin, Z.-H. Xu and M. Meuwly, *J. Phys. Chem. B*, 2015, **119**, 3112–3122.

[117] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.

[118] C. Kramer, P. Gedeck and M. Meuwly, *J. Comp. Chem.*, 2012, **33**, 1673–1688.

[119] S. Nosé, *J. Chem. Phys.*, 1984, **81**, 511–519.

[120] W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695–1697.

[121] H. C. Andersen, *J. Chem. Phys.*, 1980, **72**, 2384–2393.

[122] S. Nosé and M. L. Klein, *Mol. Phys.*, 1983, **50**, 1055–1076.

[123] W. V. Gunsteren and H. Berendsen, *Mol. Phys.*, 1997, **34**, 1311–1327.

[124] P. J. Steinbach and B. R. Brooks, *J. Comput. Chem.*, 1994, **15**, 667–683.

[125] T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

[126] D. T. Colbert and W. H. Miller, *J. Chem. Phys.*, 1992, **96**, 1982–1991.

[127] D. Koner, S. M. Salehi, P. Mondal and M. Meuwly, *J. Chem. Phys.*, 2020, **153**, 010901.

[128] P. Hamm and M. Zanni, *Concepts and Methods of 2D Infrared Spectroscopy*, Cambridge University Press: New York, 2011.

[129] K. Moller, R. Rey and J. Hynes, *J. Phys. Chem. A*, 2004, **108**, 1275–1289.

[130] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and S. Contributors, *Nature Methods*, 2020, **17**, 261–272.

[131] V. Zoete, M. Meuwly and M. Karplus, *J. Mol. Biol.*, 2004, **342**, 913–929.

[132] B. Dhayalan, A. Fitzpatrick, K. Mandal, J. Whittaker, M. A. Weiss, A. Tokmakoff and S. B. H. Kent, 2016, **17**, 415–420.

[133] X.-X. Zhang, K. C. Jones, A. Fitzpatrick, C. S. Peng, C.-J. Feng, C. R. Baiz and A. Tokmakoff, *J. Phys. Chem. B*, 2016, **120**, 5134–5145.

[134] M. Falconi, M. Cambria, A. Cambria and A. Desideri, *J. Biomol. Struc. Dyn.*, 2001, **18**, 761–772.

[135] J. L. Desmond, D. Koner and M. Meuwly, *J. Phys. Chem. B*, 2019, **123**, 6588–6598.

[136] M. Reppert, A. R. Roy and A. Tokmakoff, *J. Chem. Phys.*, 2015, **142**, 125104.

[137] K. El Hage, S. Brickel, S. Hermelin, G. Gaulier, C. Schmidt, L. Bonacina, S. C. van Keulen, S. Bhattacharyya, M. Chergui, P. Hamm, U. Rothlisberger, J.-P. Wolf and M. Meuwly, *Struct. Dyn.*, 2017, **4**, 061507.

[138] K. L. Koziol, P. J. M. Johnson, B. Stucki-Buchli, S. A. Waldauer and P. Hamm, *Curr. Op. Struct. Biol.*, 2015, **34**, 1–6.

[139] P. Hamm, M. Lim and R. M. Hochstrasser, *J. Phys. Chem. B*, 1998, **5647**, 6123–6138.

[140] M. T. Zanni, M. C. Asplund and R. Hochstrasser, *J. Chem. Phys.*, 2001, **114**, 4579–4590.

[141] P. Hamm and R. M. Hochstrasser, in *Ultrafast Infrared and Raman Spectroscopy*, ed. M. D. Fayer, Marcel Dekker, New York, 2001, pp. 273–347.

[142] M. F. DeCamp, L. DeFlores, J. M. McCracken, A. Tokmakoff, K. Kwac and M. Cho, *J. Phys. Chem. B*, 2005, **109**, 11016–11026.

[143] A. Bastida, M. A. Soler, J. Zuniga, A. Requena, A. Kalstein and S. Fernandez-Alberti, *J. Chem. Phys.*, 2010, **132**, 224501.

[144] A. Bastida, M. A. Soler, J. Zuniga, A. Requena, A. Kalstein and S. Fernandez-Alberti, *J. Phys. Chem. B*, 2012, **116**, 2969–2980.

[145] R. Rey and J. T. Hynes, *J. Chem. Phys.*, 1998, 142–153.

[146] M. W. Lee and M. Meuwly, *J. Phys. Chem. A*, 2011, **115**, 5053–5061.

[147] M. Koziński, S. Garrett-Roe and P. Hamm, *Chem. Phys.*, 2007, **341**, 5–10.

[148] M. Li, J. Owrutsky, M. Sarisky, J. P. Culver, A. Yodh and R. M. Hochstrasser, *J. Chem. Phys.*, 1993, **98**, 5499–5507.

[149] P. Hamm, M. Lim and R. Hochstrasser, *Phys. Rev. Lett.*, 1998, **81**, 5326–5329.

[150] Q. Zhong, A. Baronavski and J. Owrutsky, *J. Chem. Phys.*, 2003, **118**, 7074–7080.

[151] H. Maekawa, K. Ohta and K. Tominaga, *Phys. Chem. Chem. Phys.*, 2004, **6**, 4074–4077.

[152] S. Li, J. R. Schmidt and J. L. Skinner, *J. Chem. Phys.*, 2006, **125**, 244507.

[153] S. Li, C. Lawrence and J. Skinner, *Abstr. Pap. Am. Chem. Soc.*, 2005, **229**, U775.

[154] J.-H. Choi, D. Raleigh and M. Cho, *J. Phys. Chem. Lett.*, 2011, **2**, 2158–2162.

[155] J. Borek, F. Perakis, F. Klaesi, S. Garrett-Roe and P. Hamm, *J. Chem. Phys.*, 2012, **136**, 224503.

[156] K. Ohta, J. Tayama and K. Tominaga, *Phys. Chem. Chem. Phys.*, 2012, **14**, 10455–10465.

[157] Q. Zhong, A. Baronavski and J. Owrutsky, *J. Chem. Phys.*, 2003, **119**, 9171–9177.

[158] M. Polak, M. Gruebele, Saykally and R. Kaldor, *J. Am. Chem. Soc.*, 1987, **109**, 2884–2887.

[159] M. Polak, M. Gruebele, G. Peng and R. Saykally, *J. Chem. Phys.*, 1988, **89**, 110–114.

[160] H. Maekawa, K. Ohta and K. Tominaga, *Res. Chem. Intermed.*, 2005, **31**, 703–716.

[161] M. Garcia-Viloca, K. Nam, C. Alhambra and J. Gao, *J. Phys. Chem. B*, 2004, **108**, 13501–13512.

[162] P. Sebald, C. Stein, R. Oswald and P. Botschwina, *J. Phys. Chem. A*, 2013, **117**, 13806–13814.

[163] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby and M. Schütz, *WIREs Comput. Mol. Sci.*, 2012, **2**, 242–253.

[164] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, S. J. Bennie, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, S. J. R. Lee, Y. Liu, A. W. Lloyd, Q. Ma, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, T. F. Miller III, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson and M. Wang, *MOLPRO, Version 2012.1, A Package of ab Initio Programs*, 2012.

[165] M. Meuwly and J. Hutson, *J. Chem. Phys.*, 1999, **110**, 3418–3427.

[166] T. Hollebeek, T.-S. Ho and H. Rabitz, *Ann. Rev. Phys. Chem.*, 1999, **50**, 537–570.

[167] J. Tennyson, M. A. Kostin, P. Barletta, G. J. Harris, O. L. Polyansky, J. Ramanlal and N. F. Zobov, *Comp. Phys. Comm.*, 2004, **163**, 85–116.

[168] B. Sutcliffe and J. Tennyson, *Int. J. Quant. Chem.*, 1991, **39**, 183–196.

[169] A. Morita and S. Kato, *J. Chem. Phys.*, 1998, **109**, 5511–5523.

[170] R. Lamoureux and D. Dows, *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.*, 1975, **31**, 1945–1949.

[171] S. E. Bradforth, E. H. Kim, D. W. Arnold and D. M. Neumark, *J. Chem. Phys.*, 1993, **98**, 800–810.

[172] P. Hamm, M. Lim and R. M. Hochstrasser, *J. Chem. Phys.*, 1997, **107**, 10523–10531.

[173] D. Laage, G. Stirnemann, F. Sterpone, R. Rey and J. T. Hynes, *Annu. Rev. Phys. Chem.*, 2011, **62**, 395–416.

[174] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[175] N. Kumagai, K. Kawamura and T. Yokokawa, *Mol. Sim.*, 1994, **12**, 177–186.

[176] N. Plattner and M. Meuwly, *ChemPhysChem*, 2008, **9**, 1271–1277.

[177] M. Pezzella, O. T. Unke and M. Meuwly, *J. Phys. Chem. Lett.*, 2018, **9**, 1822–1826.

[178] M. Devereux, N. Plattner and M. Meuwly, *J. Phys. Chem. A*, 2009, **113**, 13199–13209.

[179] K. El Hage, P. K. Gupta, R. Bemish and M. Meuwly, *J. Phys. Chem. Lett.*, 2017, **8**, 4600–4607.

[180] V. Anisimov, G. Lamoureux, I. Vorobyov, N. Huang, B. Roux and A. MacKerell, *J. Chem. Theo. Comp.*, 2005, **1**, 153–168.

[181] J. M. Plitzko, B. Schuler and P. Selenko, *Curr. Op. Struct. Biol.*, 2017, **46**, 110–121.

[182] M. M. Waegele, R. M. Culik and F. Gai, *J. Phys. Chem. Lett.*, 2011, **2**, 2598–2609.

[183] R. E. Horness, E. J. Basom and M. C. Thielges, *Anal. Chem.*, 2015, **7**, 7234–7241.

[184] S. Bagchi, S. G. Boxer and M. D. Fayer, *J. Phys. Chem. B*, 2012, **116**, 4034–4042.

[185] J. Zimmermann, M. C. Thielges, Y. J. Seo, P. E. Dawson and F. E. Romesberg, *Angew. Chem. Int. Ed.*, 2011, **50**, 8333–8337.

[186] J. T. King and K. J. Kubarych, *J. Am. Chem. Soc.*, 2012, **134**, 18705–18712.

[187] J. T. King, E. J. Arthur, C. L. Brooks III and K. J. Kubarych, *J. Phys. Chem. B*, 2012, **116**, 5604–5611.

[188] J. T. King, E. J. Arthur, C. L. Brooks, III and K. J. Kubarych, *J. Am. Chem. Soc.*, 2014, **136**, 188–194.

[189] K. El Hage, F. Hedin, P. K. Gupta, M. Meuwly and M. Karplus, *Elife*, 2018, **7**, e35560.

[190] M. Pezzella, K. El Hage, M. J. Niesen, S. Shin, A. P. Willard, M. Meuwly and M. Karplus, *J. Phys. Chem. B*, 2020, **124**, 6540–6554.

[191] K. Kiick, E. Saxon, D. Tirrell and C. Bertozzi, *Proc. Natl. Acad. Sci.*, 2002, **99**, 19–24.

[192] K. Chiba-Kamoshida, T. Matsui, A. Ostermann, T. Chatake, T. Ohhara, I. Tanaka, K. Yutani and N. Niimura, *Acta Crystallogr., Sect. A: Found. Adv.*, 2002, **58**, C305.

[193] M. Schwilk, Q. Ma, C. Koeppl and H.-J. Werner, *J. Chem. Theo. Comp.*, 2017, **13**, 3650–3675.

[194] Q. Ma, M. Schwilk, C. Koeppl and H.-J. Werner, *J. Chem. Theo. Comp.*, 2018, **14**, 6750.

[195] T. H. Dunning, Jr., *J. Chem. Phys.*, 1989, **90**, 1007–1023.

[196] V. Zoete, M. Cuendet, A. Grosdidier and O. Michielin, *J. Comp. Chem.*, 2011, **32**, 2359–2368.

[197] P. Soldán and J. M. Hutson, *J. Chem. Phys.*, 2000, **112**, 4415–4416.

[198] J. P. Foster and F. Weinhold, *J. Am. Chem. Soc.*, 1980, **102**, 7211–18.

[199] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, *Gaussian 09, Revision A.02*, Gaussian, Inc., Wallingford, CT, 2009.

[200] S. M. Salehi, D. Koner and M. Meuwly, *J. Phys. Chem. B*, 2020, **124**, 11882–11894.

[201] M. Okuda, K. Ohta and K. Tominaga, *J. Chem. Phys.*, 2015, **142**, 212418.

[202] J. M. Bowman and B. Gazdy, *J. Chem. Phys.*, 1991, **94**, 816–817.

[203] M. Meuwly and J. M. Hutson, *J. Chem. Phys.*, 1999, **110**, 8338–8347.

[204] P. Hamm, M. Lim and R. M. Hochstrasser, *J. Phys. Chem. B*, 1998, **5647**, 6123–6138.

[205] D. Nutt and M. Meuwly, *Proc. Natl. Acad. Sci.*, 2004, **101**, 5998–6002.

[206] K. Nienhaus, J. S. Olson, S. Franzen and G. U. Nienhaus, *J. Am. Chem. Soc.*, 2005, **127**, 40–41.

[207] M. Meuwly, *Chem. Phys. Chem.*, 2006, **7**, 2061–2063.

[208] J. Y. Park, H.-J. Kwon, S. Mondal, H. Han, K. Kwak and M. Cho, *Phys. Chem. Chem. Phys.*, 2020, **22**, 19223–19229.

[209] S. Shin and A. P. Willard, *J. Chem. Theo. Comp.*, 2018, **14**, 461–465.

[210] S. Shin and A. P. Willard, *J. Phys. Chem. B*, 2018, **122**, 6781–6789.

[211] H. T. Turan and M. Meuwly, 2021, **125**, 4262–4273.

[212] B. J. Grant, A. P. C. Rodrigues, K. M. ElSawy, J. A. McCammon and L. S. D. Caves, *Bioinformatics*, 2006, **22**, 2695–2696.

[213] J. Helbing, M. Devereux, K. Nienhaus, G. U. Nienhaus, P. Hamm and M. Meuwly, *J. Phys. Chem. A*, 2012, **116**, 2620–2628.

[214] P. Pagano, Q. Guo, A. Kohen and C. M. Cheatum, *J. Phys. Chem. Lett.*, 2016, **7**, 2507–2511.

[215] E. T. Baldwin, T. N. Bhat, S. Gulnik, B. Liu, I. A. Topol, Y. Kiso, T. Mimoto, H. Mitsuya and J. W. Erickson, *Structure*, 1995, **3**, 581–590.

[216] V. Prashar, S. Bihani, A. Das, J.-L. Ferrer and M. Hosur, *PloS one*, 2009, **4**, e7860.

[217] G. Schiro, Y. Fichou, F.-X. Gallat, K. Wood, F. Gabel, M. Moulin, M. Haertlein, M. Heyden, J.-P. Colletier, A. Orecchini, A. Paciaroni, J. Wuttke, D. J. Tobias and M. Weik, *Nuovo Cim.*, 2015, **6**, 1–8.

[218] Y. Pocker, *Cell. Mol. Life Sci.*, 2000, **57**, 1008–1017.

[219] S. Pal and A. Zewail, *Chem. Rev.*, 2004, **104**, 2099–2123.

[220] L. M. Miller, M. W. Bourassa and R. J. Smith, *Bioch. Bioph. Acta*, 2013, **1828**, 2339–2346.

[221] S. M. Salehi and M. Meuwly, *J. Chem. Phys.*, 2021, **154**, 165101.

[222] A. L. Le Sueur, R. N. Schaugaard, M.-H. Baik and M. C. Thielges, *J. Am. Chem. Soc.*, 2016, **138**, 7187–7193.

[223] L. Liu, W. A. Baase and B. W. Matthews, *J. Mol. Biol.*, 2009, **385**, 595–605.

[224] M. Cho, G. Fleming, S. Saito, I. Ohmine and R. Stratt, *J. Chem. Phys.*, 1994, **100**, 6672–6683.

[225] P. Mondal, P.-A. Cazade, A. K. Das, T. Bereau and M. Meuwly, *arXiv preprint arXiv:2106.10142, in print in J. Phys. Chem. B*, 2021.

[226] S. D. Fried, S. Bagchi and S. G. Boxer, *Science*, 2014, **346**, 1510–1514.

[227] S. Woutersen, R. Pfister, P. Hamm, Y. Mu, D. S. Kosov and G. Stock, *J. Chem. Phys.*, 2002, **117**, 6833–6840.

[228] T. Ichiye and M. Karplus, *Protein Struct. Funct. Genet.*, 1991, **11**, 205–217.

[229] G. Arnold and R. Ornstein, *Biophys. J.*, 1997, **73**, 1147–1159.

[230] W. Becker, K. C. Bhattiprolu, N. Gubensäk and K. Zangger, *ChemPhysChem*, 2018, **19**, 895.

[231] P. Metrangolo, H. Neukirch, T. Pilati and G. Resnati, *Acc. Chem. Res.*, 2005, **38**, 386–395.

[232] P. Metrangolo, F. Meyer, T. Pilati, G. Resnati and G. Terraneo, *Angew. Chem. Int. Ed.*, 2008, **47**, 6114–6127.

[233] Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang and W. Zhu, *J. Med. Chem.*, 2009, **52**, 2854–2862.

[234] R. Wilcken, M. O. Zimmermann, A. Lange, A. C. Joerger and F. M. Boeckler, *J. Med. Chem.*, 2013, **56**, 1363–1388.

[235] G. R. Desiraju, P. S. Ho, L. Kloo, A. C. Legon, R. Marquardt, P. Metrangolo, P. Politzer, G. Resnati and K. Rissanen, *Pure Appl. Chem.*, 2013, **85**, 1711–1713.

[236] T. Clark, M. Hennemann, J. S. Murray and P. Politzer, *J. Mol. Model.*, 2007, **13**, 291–296.

[237] H. Wang, W. Wang and W. J. Jin, *Chem. Rev.*, 2016, **116**, 5072–5104.

[238] D. Chopra and T. N. G. Row, *CrystEngComm*, 2011, **13**, 2175–2186.

[239] k. Vanommeslaeghe, E. hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. Mackerell, *J. Comp. Chem.*, 2010, **31**, 671–690.

[240] T. Bereau, C. Kramer and M. Meuwly, *J. Chem. Theo. Comp.*, 2013, **9**, 5450–5459.

[241] P.-A. Cazade, H. Tran, T. Bereau, A. K. Das, F. Klaesi, P. Hamm and M. Meuwly, *J. Chem. Phys.*, 2015, **142**, 212415.

[242] *CPMD, Copyright IBM Corp 1990-2019, Copyright MPI für Festkörperforschung Stuttgart 1997-2001*, http://www.cpmd.org/, `http://www.cpmd.org/`.

[243] J. Hermans, H. J. C. Berendsen, W. F. van Gunsteren and J. P. M. Postma, *Biopol.*, 1984, **23**, 1513–1518.

[244] A. Laio, J. VandeVondele and U. Rothlisberger, *J. Chem. Phys.*, 2002, **116**, 6941–6947.

[245] M. C. Colombo, L. Guidoni, A. Laio, A. Magistrato, P. Maurer, S. Piana, U. Röhrig, K. Spiegel, M. Sulpizi, J. VandeVondele, M. Zumstein and U. Röthlisberger, *CHIMIA International Journal for Chemistry*, 2002, **56**, 13–19.

[246] E. Brunk and U. Rothlisberger, *Chem. Rev.*, 2015, **115**, 6217–6263.

[247] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, **25**, 1157–1174.

[248] *GAFF2 is a public domain forcefield, an upgrade of the previously released general AMBER forcefield (GAFF). It is available with the distribution of AmberTools17 and can be downloaded from http://ambermd.org. A publication for it is currently under preparation.*, http://ambermd.org/, `http://ambermd.org/`.

[249] O. T. Unke and M. Meuwly, *J. Chem. Theo. Comp.*, 2019, **15**, 3678–3693.

[250] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, Proc. of the 34th Int. Conf. on Machine Learning-Volume 70, 2017, pp. 1263–1272.

[251] J. S. Smith, O. Isayev and A. E. Roitberg, *Sci. Data*, 2017, **4**, 170193.

[252] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dulak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiotz, O. Schutt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *J. Phys. Condens. Matter*, 2017, **29**, 273002.

[253] M. Thomas, M. Brehm, R. Fligg, P. Vöhringer and B. Kirchner, *Phys. Chem. Chem. Phys.*, 2013, **15**, 6608–6622.

[254] M. Schmitz and P. Tavan, *J. Chem. Phys.*, 2004, **121**, 12233–12246.

[255] M. Schmitz and P. Tavan, *J. Chem. Phys.*, 2004, **121**, 12247–12258.

[256] W. Zierkiewicz and D. Michalska, *J. Phys. Chem.*, 2003, **107**, 4547–4554.

[257] D. Michalska, W. Zierkiewicz, D. C. Bienko, W. Wojciechowski and T. Zeegers-Huyskens, *J. Phys. Chem.*, 2001, **105**, 8734–8739.

[258] J. Evans, *Scientific American*, 1960, **16**, 1382–1392.

[259] H. Bist, J. C. Brand and D. Williams, *J. Mol. Spectrosc.*, 1967, **24**, 402–412.

[260] Y. Morino and K. Kuchitsu, *J. Chem. Phys.*, 1952, **20**, 1809–1810.

[261] S. Käser, E. D. Boittier, M. Upadhyay and M. Meuwly, *J. Chem. Theo. Comp.*, 2021, **17**, 3687–3699.

[262] T. Hamashima, K. Mizuse and A. Fujii, *J. Phys. Chem. A*, 2011, **115**, 620–625.

[263] T. Shimamori and A. Fujii, *J. Phys. Chem. A*, 2015, **119**, 1315–1322.

[264] P. Banerjee, I. Bhattacharya and T. Chakraborty, *Scientific American*, 2017, **181**, 116–121.

[265] R. Kusaka, T. Ishiyama, S. Nihonyanagi, A. Morita and T. Tahara, *Phys. Chem. Chem. Phys.*, 2018, **20**, 3002–3009.

[266] W. Humphrey, A. Dalke and K. Schulten, *J. Mol. Graph.*, 1996, **14**, 33–38.

[267] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020.

Appendix A

Mutipolar Parametrization

Atom (PhOH)	Charge (e)	Atom (F-PhOH)	charge (e)
C1	-0.095	C1	0.161
C2	-0.074	C2	-0.060
C3	-0.079	C3	-0.061
C4	0.075	C4	0.064
C5	-0.079	C5	-0.061
C6	-0.074	C6	-0.060
H7	0.086	H7	0.113
H8	0.102	H8	0.105
H9	0.102	H9	0.105
H10	0.086	H10	0.113
O11	-0.392	O11	-0.389
H12	0.259	H12	0.260
H13	0.082	F13	-0.291

Table A.1: Molecular monopoles calculated using a fitting environment with GDMA algorithm for PhOH and F-PhOH.[70]

Atom	Q10 (e)	Q11c (e)	Q11s (e)
PhOH			
C1	-0.015	0.0	0.033
C2	-0.016	-	0.030
C3	-0.0002	-	0.054
C4	0.028	0.0	0.089
C5	-0.0002	-	0.054
C6	-0.016	-	0.030
O11	0.0	0.099	-0.070
H13	0.0	0.0	0.0
F-PhOH			
C1	-0.011	0.0	0.180
C2	-0.022	-	0.073
C3	-0.025	-	0.017
C4	0.015	0.0	0.122
C5	-0.025	-	0.017
C6	-0.022	-	0.073
O11	0.0	0.121	-0.140
F13	0.147	0.0	0.008

Table A.2: Atomic dipoles for PhOH and F-PhOH from fitting to the molecular electrostatic potential.[70] Qxx are the spherical MTP coefficients expressed in the local axis system.

Atom	Q20 (e)	Q21c (e)	Q21s (e)	Q22c(e)	Q22s(e)
PhOH					
C1	-0.029	0.0	0.001	-0.005	0.0
C2	-0.026	-	-0.0003	-0.0013	-
C3	-0.012	-	0.002	0.0039	-
C4	8.04×10^{-5}	0.0	0.013	0.0015	0.0
C5	-0.012	-	0.0029	0.0039	-
C6	-0.0260	-	-0.0003	-0.0013	-
O11	-0.006	0.0	0.0	0.015	-0.0278
H13	0.0	0.0	0.0	0.0	0.0
F-PhOH					
C1	-0.024	0.0	0.006	-0.007	0.0
C2	-0.042	-	0.003	0.0005	-
C3	-0.029	-	0.002	0.009	-
C4	0.005	0.0	0.022	-0.009	0.0
C5	-0.029	-	0.002	0.009	-
C6	-0.042	-	0.003	0.0005	-
O11	-0.034	0.0	0.0	0.029	-0.0798
F13	-0.034	0.0	0.0016	0.009	0.0

Table A.3: Atomic quadrupoles for PhOH and F-PhOH from fitting to the molecular electrostatic potential.[70] Qxx are the spherical MTP coefficients expressed in the local axis system.

Appendix B

List of publications

- ☐ Salehi, S. M; Meuwly, M. **"Site-Selective Dynamics of Ligand-Free and Ligand-Bound Azidolysozyme"** *J. Chem. Phys.* **156**, 105105 (2022)
  doi:10.1063/5.0077361
- ☐ Salehi, S. M; Meuwly, M. **"Cross-Correlated Motions in Azidolysozyme"**, *Molecules* **27**, 839 (2022)
  doi:10.3390/molecules27030839
- ☐ Salehi, S. M; Meuwly, M. **"Site-Selective Dynamics of Azidolysozyme"** *J. Chem. Phys.* **154**, 165101 (2021)
  doi:10.1063/5.0047330
- ☐ Salehi, S. M; Koner, D; Meuwly, M. **"The Dynamics and Infrared Spectroscopy of Monomeric and Dimeric Wild Type and Mutant Insulin"** *J. Phys. Chem. B* **124**, 52, 11882–11894 (2020)
  doi:10.1021/acs.jpcb.0c08048
- ☐ Koner, D; Salehi, S. M; Mondal, P; Meuwly, M. **"Non-conventional Force Fields for Applications in Spectroscopy and Chemical Reaction Dynamics."** *J. Chem. Phys.* **153**, 010901 (2020)
  doi:10.1063/5.0009628
- ☐ Salehi, S. M; Koner, D; Meuwly, M. **"Vibrational Spectroscopy of N_3^- in the Gas and Condensed Phase"** *J. Phys. Chem. B* **123**, 123, 15, 3282–3290 (2019)
  doi:10.1021/acs.jpcb.8b11430