SIMULATED RAMAN SPECTRAL ANALYSIS OF ORGANIC MOLECULES

By

LU LU

A THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Applied Optics Graduate Program of Delaware State University

DOVER, DELAWARE
August 2017

This thesis is approved by the following members of the Final Oral Review Committee:

Dr. Jun Ren, Committee Chairperson, Department of Physics and Engineering, Delaware State University
Dr. Essaid Zerrad, Committee Member, Department of Physics and Engineering, Delaware State University
Dr. Gour Pati, Committee Member, Department of Physics and Engineering, Delaware State University
Dr. Jinjie Liu, External Committee Member, Department of Mathematical Sciences, Delaware State University

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Jun Ren for the support of my master study and research, for her patience, motivation, enthusiasm, encouragement and immense knowledge. Her guidance helped me in all the time of research and preparing of the thesis. I have been extremely lucky to have an advisor who cared so much about my work, and who responded to my questions and queries so promptly. I would also like to express my gratitude to Dr. Melikechi, without his kind help, I will not be able to settle down at DSU and start my research.

Besides my advisor and Dr.Melikechi, I would like to thanks the rest of my thesis committee: Dr. Essaid Zerrad, Dr. Gour Pati, and Dr. Jinjie Liu, for their encouragement and insightful comments.

I want to thank my labmates in Delaware State University Scientific Computing Group: Deshaun Crawford, Aleem Sayles, Monica Elavarthi, Anthony Falice, and Hongbin Lu for the stimulating discussions and all the fun we have had in the last two years.

I also want to thank the faculty and staff members of the Optical Science Center for Applied Research (OSCAR), Jacquelyn Jones and Debbie roussell, and the Department of Physics and Engineering, Dr. Mukti Rana, Amal Juracka and Rose Shields for their many professional and caring help over the years.

Last but not least, I would like to thank the Funding support from Delaware State DeDo grant and NSF support (Award Number: 1649173).

SIMULATED RAMAN SPECTRAL ANALYSIS OF ORGANIC MOLECULES

Lu Lu

Faculty Advisor: Dr. Jun Ren

ABSTRACT

The advent of the laser technology in the 1960s solved the main difficulty of Raman spectroscopy, resulted in simplified Raman spectroscopy instruments and also boosted the sensitivity of the technique. Up till now, Raman spectroscopy is commonly used in chemistry and biology. As vibrational information is specific to the chemical bonds, Raman spectroscopy provides fingerprints to identify the type of molecules in the sample. In this thesis, we simulate the Raman Spectrum of organic and inorganic materials by General Atomic and Molecular Electronic Structure System (GAMESS) and Gaussian, two computational codes that performs several general chemistry calculations. We run these codes on our CPU-based high performance cluster (HPC). Through the message passing interface (MPI), a standardized and portable message-passing system which can make the codes run in parallel, we are able to decrease the amount of time for computation and increase the sizes and capacities of systems simulated by the codes. From our simulations, we will set up a database that allows search algorithm to quickly identify N-H and O-H bonds in different materials. Our ultimate goal is to analyze and identify the spectra of organic matter compositions from meteorites, and compared these spectra with terrestrial biologically-produced amino acids and residues.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACCD	Aug-cc-pvdz Basis Set
DFT	Density Functional Theory
GAMESS	General Atomic and Molecular Electronic Structure System
GROMACS	GROningen MAchine for Chemical Simulations
GUI	Graphical User Interface
GTO	Gaussian-type Orbitals
HF Method	Hartree-Fock Method
HPC	High-Performance Computing
IR	Infrared
MD	Molecular Dynamics
MPI	Message Passing Interface
RMSD	Root-Mean-Square Deviation

CHAPTER I: INTRODUCTION

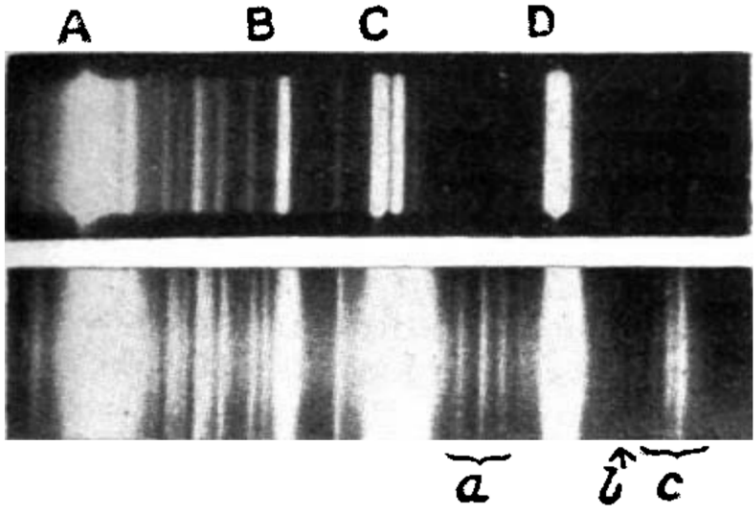
1.1 A Brief History of Raman Spectroscopy

Twenty percent of the human body is made up of protein. Protein plays a crucial role in almost all biological processes, including forming orgasmic tissues and building up enzymes to keep our body functioning normally. Amino acids, which are the building blocks of proteins, also play a crucial role in our lives. A large amount of our cells, tissue, and muscles build by amino acids, meaning they perform many significant important bodily functions. For example, in the human brain, glutamate and gamma-aminobutyric acids are, respectively, the main excitatory and inhibitory neurotransmitters [1]. Therefore, the study of amino acids' biological and chemical properties can help provide a deeper understanding of the origin of life and the production of drugs, biodegradable plastics, and chiral catalysts.

Due to the fact that an amino acid is a small molecule, a proper tool is needed to study its properties. Raman Spectroscopy is exactly the tool to carry out this job.

In 1923, a paper titled "The quantum theory of dispersion" was published in the science journal *Naturwissenschaften* by the Austrian quantum physicist A. Smekal [2]. This paper theoretically predicted the scattering of monochromatic radiation with a change in frequency of light-material interaction, later called Raman scattering. Scattering light passing through various mediums was studied after the first prediction, but no change in wavelength was observed until 1928, when Indian scientist C.V. Raman and his coworker K.S. Krishnan first discovered this kind of inelastic scattering using a quartz mercury vapor lamp [3]. By comparing the spectra in Figure 1 from the incident light and the scattering light, one sees several other lines in addition to the lines in the incident spectrum. Raman and Krishnan published their finding in *Nature* with the title "The optical analogue of Compton effect," and 2 years later they received the Nobel

Figure 1: Spectrum of incident light (above) and scattered light (below)



Prize in physics for their observation of light inelastic scattering, which was named Raman scattering in honor of his contribution.

Developments in Raman spectroscopy based on the theory of Raman scattering occurred slowly during the period from 1930 to 1950, since there was no proper monochromatic radiation source. In the early experiments during the 1930s, the mercury lamp, filtered to offer the monochromatic light, was the most common radiation source. The mercury Toronto arc lamp was introduced as the ultimate source later in 1952 [4]. However, the intensity of the mercury arc lamp was so weak that they needed significant exposure time for the photographic receiver to create a readable Raman spectrum. The invention of laser technology in 1960 solved this problem and provided a monochromatic source, which improved the Raman scattering intensity and shortened the time for exposure. In 1962, Porto and Wood [5] reported the first use of a pulsed ruby laser for exciting Raman spectra. After that, Raman spectroscopy developed quickly and became a practical tool for studying vibrational information on the molecular atomic scale in many fields.

1.2 Thesis Objective

The ultimate objective of this thesis is to utilize an advanced computational technique to simulate the Raman spectrum of inorganic and organic material as compelling data for analyzing and identifying the organic components in an unknown sample. Since Raman spectroscopy can provide fingerprints to identify the type of molecules in the sample, it is important to understand the pattern of Raman Spectrum for different organic materials. We start with the simplest amino acid, water, and glycine, and simulate their Raman spectra using the General Atomic and Molecular Electronic Structure System (GAMESS) a quantum chemistry computational code, on our High Performance Computing (HPC) cluster. We then determine some characteristic spectrum peaks that are the fingerprint of glycine and develop a search algorithm for identifying these peaks. And furthermore to identify different kinds of amino acid from their Raman spectra. Once we verify the accuracy of our search algorithm, we will set up a database that allows search algorithm to quickly identify N-H and O-H bonds in different materials. Since Raman spectroscopy has some unique properties, such as its ability to be used with solids and liquids without the additional work of sample preparation, it is used widely in mineral identification and characterizations of bio-molecules. With our Raman spectrum analyzing and search scheme, the Raman spectroscopy technique can be applied on a deep space explorer for analyzing and detecting exterritorial organic material, which would avoid the present delays in long-distance data transition.

CHAPTER II: BASIS THEORY

In this chapter, we will introduce some basic concepts related to our topic. Since the Raman Spectroscopy is an important tool for our study, it will be the first section in this chapter. In the next section, vibrational modes of a molecule will be explained. Section 2.3, describes, the basic chemical structure of our research materials, amino acids. In the next two sections 2.4-2.5, the Hartree-Fock method for calculating the simulated Raman Spectrum is discussed. In the last two sections (2.6-2.7), both the simulation software we used and the platform for the software are introduced.

2.1 Raman Scattering

When light encounters matter, either absorption or scattering occurs. Infrared spectroscopy is based on the absorption process, and the Raman Spectroscopy is based on the scattering process.

The process of absorption requires the energy of the incident radiation to be exactly equal to the energy difference between the ground state and the excited state of a molecule. After the molecule absorbs the energy from the incident light, the electron in the ground state transits to the excited state. Therefore, the spectrum of transmitted light will have some frequency bands are missing. The frequency of the missing band equals to the vibrational frequency of the molecule. By plotting out the intensity of the transmitted light versus frequency, we can get the IR spectrum and identify specific chemical groups within the molecule.

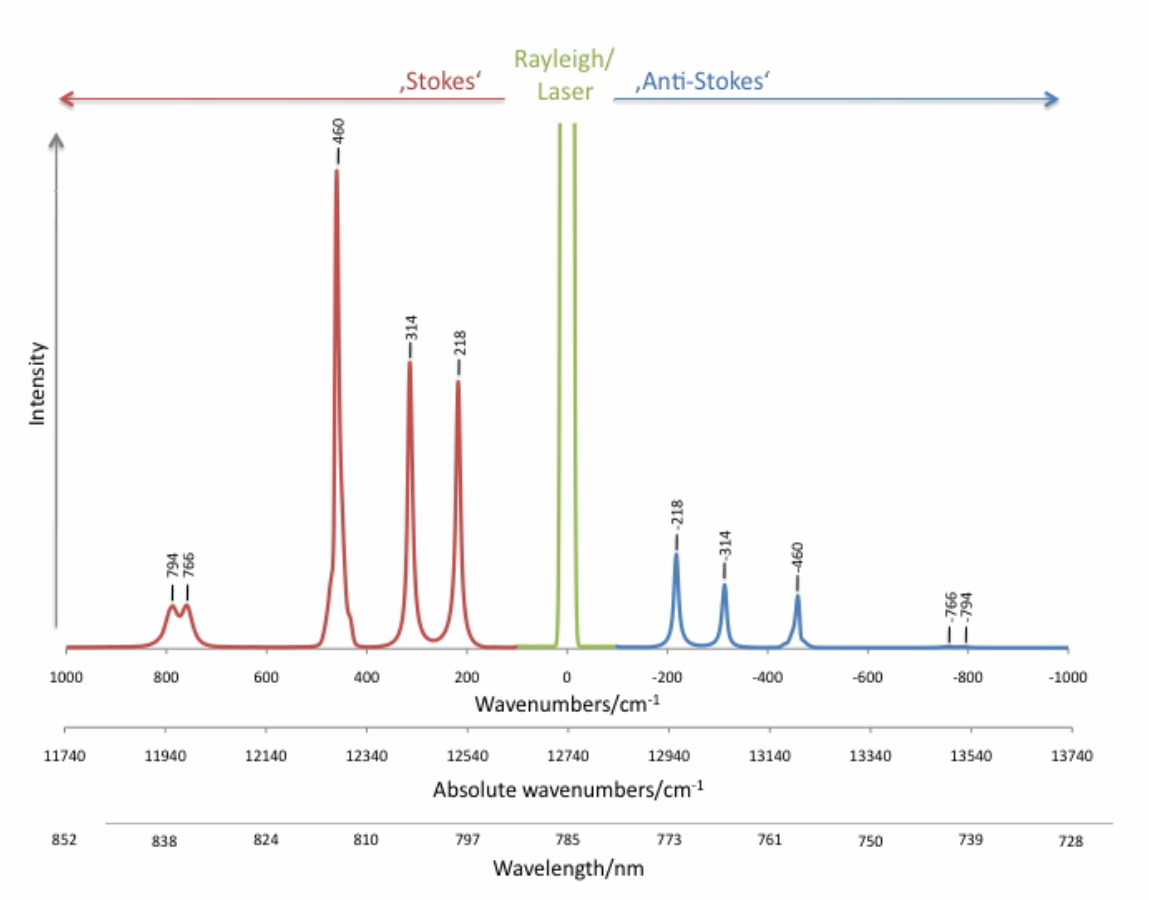
In contrast, scattering does not require the incident radiation to match the energy difference between the ground and excited states. A light wave, considered as an oscillating dipole, when passing through a molecule, interacts and distorts the clouds of electrons orbiting the nuclei. Energy is released in the form of scattered radiation. As the wavelength of visible

light is much greater than the size of a common molecule, the light-molecule interaction polarizes the electrons and excites the molecule to a higher energy state called a “visual state,” which takes place in a short time. Unlike electrons, which have a smaller mass, the nuclei do not have time to respond. This process results in the molecule reaching a high-energy state by changing the electron geometry without moving the nuclei. Since these high-energy-state electrons are unstable and cannot last for a long time, they return back to their ground state and release photons in random directions. This release light is called scattering radiation.

There are two types of scattering: Rayleigh and Raman scattering.

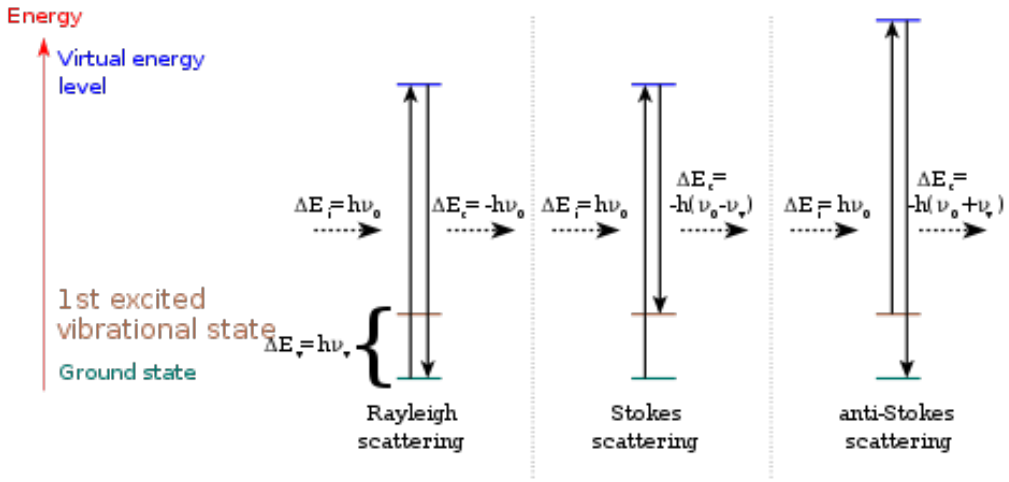
When, the electrons fall back to the ground state and radiate light with the same frequency as the

Figure 2: Stokes/anti-Stokes bands shifted from Rayleigh lines.



incident light in random directions, this elastic process is called Rayleigh scattering. Rayleigh scattering is the most intense scattering observed comparing to other types of scattering. In Figure 2, the green line in the middle is the Rayleigh line.

Figure 3: The different energy states of light scattering: Rayleigh scattering, Stokes effect and anti-Stokes effect.



On the other hand, if the frequency of the scattering light is different from that of the incident light, this inelastic process is called Raman scattering. If the scattering frequency is lower than the incident frequency, it is the Stoke effect. If the scattering frequency is higher than the incident frequency, it is the anti-Stoke effect.

Due to thermal energy, some molecules at room temperature are initially in a higher state (such as in the first excited vibrational state, colored brown in Figure 4. After being excited by the incident light to the “visual state”, the electrons fall back to the ground state shortly and emit light with a frequency higher than that of the original light. This means that during this process, the energy in the molecule transfers to the scattering light.

The opposite with anti-Stoke effect is the Stoke effect. A molecule starts from the ground state and ends in a higher state. During this interaction, energy flows from the light to the

molecule, which causes the frequency of the scattering light to be lower than that of its incident light.

The energy difference between the incident light and the scattering light is equal to the vibrational energy of the molecule. By measuring this frequency shift, we can determine the vibrational frequency of a molecule. This is the basic theory behind Raman spectroscopy. Since different chemical bonds in a molecule correspond to a variety of frequency shifts, by assigning and identifying different peaks from the Raman spectrum of an unknown molecule, we can deduce the structure and the chemical group of this molecule and identify it.

From the thermal statistics perspective, the number of particles at different energy states of a molecule at room temperature should obey the Maxwell-Boltzmann distribution. The rate of the excited state and the ground state is shown below:

$$\frac{N_n}{N_m} = \frac{g_n}{g_m} exp[\frac{-(E_n - E_m)}{kT}] \qquad (1.1)$$

In equation 1.1, N_n is the number of molecules on the excited vibrational energy level, N_m is the number of molecules on the ground vibrational energy level, g is the degeneracy of levels n and m, $E_n - E_m$ represents the energy difference between n and m, and k is Boltzmann's constant. At room temperature, the number of molecules in the ground state is much more than that in the excited state, causing the magnitude of Stoke scattering in the spectrum much higher than that of anti-Stoke scattering. Therefore, spectroscopists usually choose the frequency shift to the Stoke line to create the Raman spectrum. In Figure 2, the x-axis represents the frequency shift from incident frequency, with wavenumber cm^{-1} as a unit. The y-axis shows the intensity of the scattering rate. The frequency region from 500 to 1640 cm^{-1} is usually referred to as the

fingerprint region; because most bonds for organic material have their characteristic frequency lines in this region.

2.2 Vibrational modes of a molecule

A molecule is made up of numbers of atoms forming chemical bonds. These chemical bonds may result from the electrostatic force of attraction between atoms with opposite charges or through the sharing of electrons, as in covalent bonds. If one looks into an atom, a nucleus is surrounded by an electron cloud. These electrons form the different vibrational and rotational states of the molecule.

Figure 4: The Morse potential (blue) curve and the harmonic oscillator potential (green) curve.

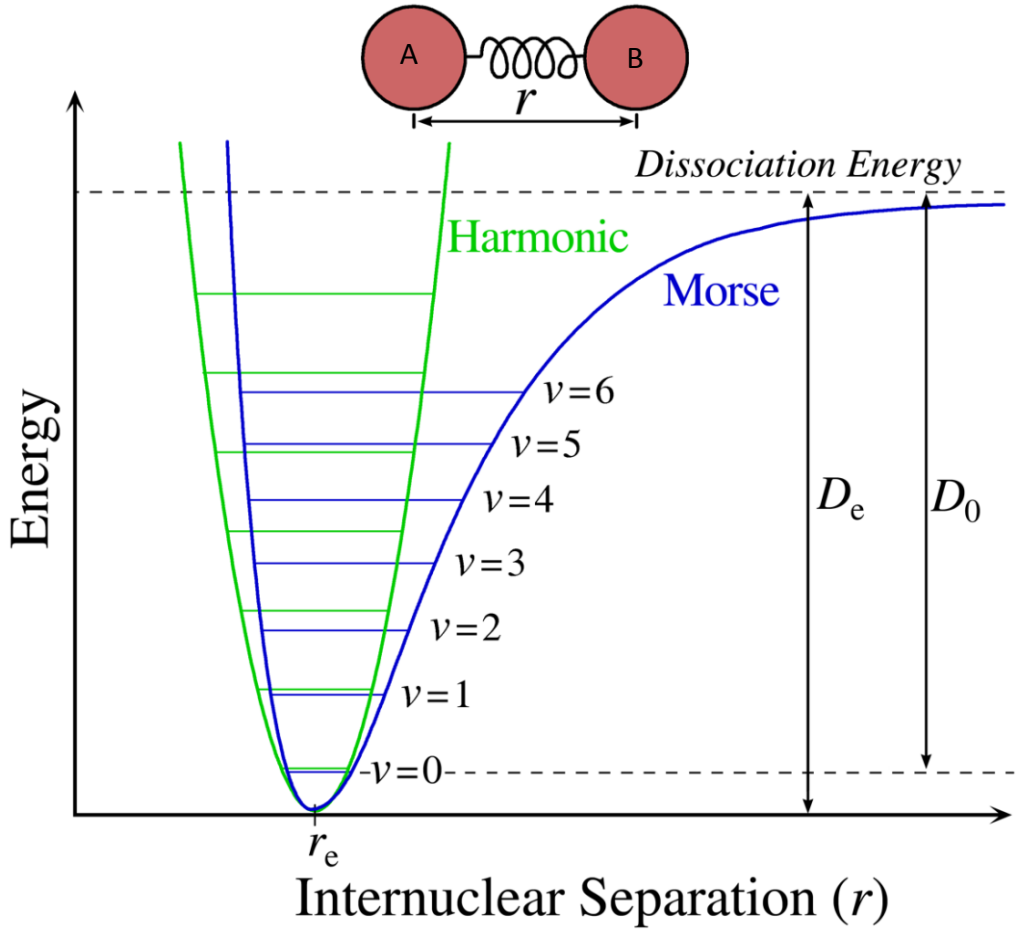


Figure 4 shows a sketch of a typical electronic state of a diatomic molecule. (A diatomic molecule is a molecule that only has two atoms—of the same or different molecule elements. The most common diatomic molecules are hydrogen [H_2] and oxygen [O_2], which are said to be homonuclear. Otherwise, if a diatomic molecule consists of two different atoms, such as carbon monoxide [CO] or nitric oxide [NO], the molecule is said to be heteronuclear [6]. The blue line in the plot, called a Morse curve, provides an institutional view of the state of the molecule. The y-axis represents the potential energy of the molecule system, and the x-axis is the separated distance between two nucleuses. When the distance between two nuclei increases, to when each atom is principally free, the system energy approaches a steady-state value. As the distance decreases, the force between two atoms is dominated by the-attraction force, although the attraction force decrease faster than increasing speed from the repulsive force. When the attraction and the repulsion equal each other, the system energy reaches its lowest point. If these atoms keep getting closer and closer, the system energy rises steeply, while the nucleus-nucleus repulsion starts to increase rapidly and dominates over the attraction. The point where the repulsion force equals the attraction is the position where the molecule system forms a chemical bond. However, not every energy state in the curve is a quantum state. Based on the quantum mechanics, as the nuclei constantly oscillate around the equilibrium position between the "potential walls" of Morse potential, the energy of this vibration is quantized and described by a series of vibrational wavefunctions with their quantum numbers (v = 0, 1, 2, ...)where the ground state (v = 0) is the lowest possible energy a molecule can have. At room temperature, the majority of molecules are in the lowest-energy vibrational state, but not all of them; there are still a small number of molecules that occupy the higher vibrational state. In statistics, the probability distribution of the molecules in each state can be calculated by the Maxwell-Boltzmann function.

Since the vibrational energy at the bottom of the molecule resembles that of the harmonic oscillator, one can treat the chemical bonds approximately as springs connecting nuclei obeying Hooke's Law. The top of Figure 4 shows a model of this approach. Each ball, labeled A or B, is linked by a spring. With this approach, by applying Hooke's Law, the relationship between the vibrations frequencies with, the mass of vibrational atoms and the force constant can be found for a diatomic molecule:

$$\omega = \frac{1}{2\pi c}\sqrt{\frac{K}{\mu}} \qquad (1.2)$$

In equation 1.2, *c* is the speed of light, v is the oscillating frequency of the system, *K* is the force constant of the bond between atom A and B, and μ is the reduced mass of atoms A and B, given by equation 1.3:

$$\mu = \frac{M_A M_B}{M_A + M_B} \qquad (1.3)$$

Here, M_A,M_B represents the mass of atoms A and B.

In such a model, the energy of each state can be represented by:

$$E_\upsilon = \frac{h}{2\pi}(\upsilon + \frac{1}{2})\omega \qquad (1.4)$$

υ = 0 1 2 …, and the vibrational energy of the molecule system can be quantized. From equations 1.1 and 1.2, one can see that the lighter the atoms, the higher the frequency. Thus, a C-H bond vibration's frequency around 9×10^8 Hz is higher than that of a C-I vibration at 1.5×10^8Hz ($\mu_H > \mu_I$). Furthermore, the force constant can be treated as the strength of the spring (chemical bond). The stronger the bond, the higher the frequency.

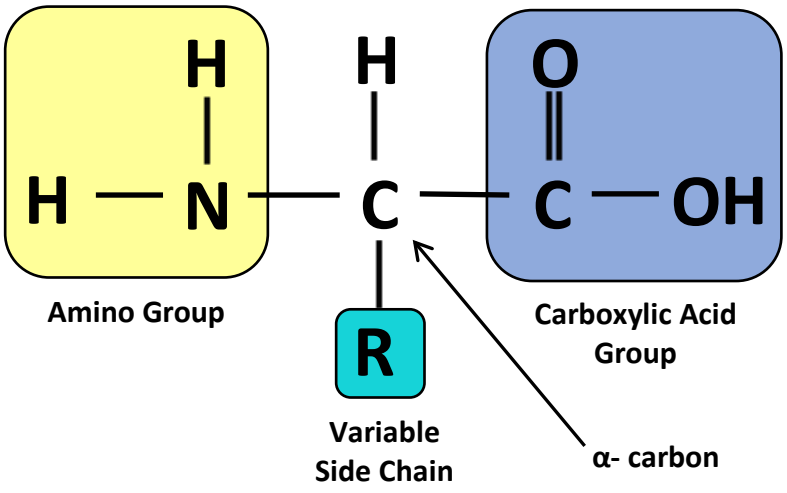
This approximation provides visualization for the vibrational energy state, but one should be aware that this approximation is not entirely the same as in a real diatomic molecule's energy state.

In equation 1.4, by calculating the energy difference between υ and υ +1, the gap between energy levels of the harmonic oscillator is evenly spaced. However, the real bond subject to the Morse curve–quantized energy state is lower than that of the harmonic oscillator, and the space for each energy state becomes smaller as the frequency (ω) increases.

2.3 Amino Acids

Amino acids are the building blocks of proteins. An amino acid is a small molecule that contains an amino group (-NH_2), a carboxyl group (-COOH), a side chain (side group), and a hydrogen atom. All these groups connect to a single carbon atom at the center of the molecule, as

Figure 5: Amino acid general structure.



shown in Figure 5. There are 20 different kinds of amino acids found on earth. Each amino acid has a special side group that shows different chemical properties and has special peak patterns in the Raman spectrum. For example, the simplest amino acid, glycine, whose side group is a

hydrogen atom, shows strong-intensity bands at 894 cm^{-1} and 1,327 cm^{-1} in a solid state. These bands can be considered the fingerprint pattern of glycine.

2.4 Hartree-Fock Method

The Hartree-Fock (HF) method is an approximation to approximately treat an interacting fermion system with an effective single-particle problem. The HF theory solves the time-dependent electronic Schrodinger equation that after including the Born-Oppenheimer approximation. In atomic units, with *r* defining electron position and *R* defining nuclear degrees of freedom, the electronic Schrodinger equation is

$$\left[-\frac{1}{2}\sum_i \nabla_i^2 - \sum_{A,i}\frac{Z_A}{r_{Ai}} + \sum_{A>B}\frac{Z_A Z_B}{R_{AB}} + \sum_{i>j}\frac{1}{r_{ij}}\right]\Psi(r;R) = E_{el}\Psi(r;R) \qquad (1.5)$$

To simplify equation 1.5, we define a one-electron operator *h*,

$$h(i) = -\frac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{r_{iA}} \qquad (1.6)$$

and a two-electron operator $v(i,j)$:

$$v(i,j) = \frac{1}{r_{ij}} \qquad (1.7)$$

The electronic Hamiltonian and the electronic Schrodinger equation can be rewritten as:

$$H_{el} = \sum_i h(i) + \sum_{i<j} v(i,j). \qquad (1.8)$$

The inter-nucleus potential energy V_{NN} is not included in equation 1.8. We neglect the term since it is just a constant for the fixed set of nuclear coordinates. Now the electronic Schrodinger equation becomes:

$$H_{el}\Psi(r;R) = E_{el}\Psi(r;R) \qquad (1.9)$$

The essential idea of the HF theory is based on the assumption that we have solved the electronic Schrodinger equation for hydrogen, which has only one electron. If we add one more electron to the hydrogen system, we suppose that the total electronic wavefunction $\Psi(r_1; r_2)$ describing the motions of the two electrons would just be the product of two hydrogen atom wavefunctions (orbitals), $\Psi(r_1)\Psi(r_2)$.So if we expand the two-electron system to a multi-electron system, our wavefunction is represented as:

$$\Psi_{HP}(x_1, x_2, \dots, x_N) = \chi_1(x_1)\chi_2(x_2) \dots \chi_N(x_N). \quad (1.10)$$

$\chi_N(x_N)$ is the spin orbital of the number Nth electron.

Apparently, this assumption fails to satisfy the antisymmetry principle, which states that a wavefunction describing fermions should be antisymmetric with respect to the interchange of any set of space-spin coordinates. Therefore, we need to introduce Slater Determinants.
A Slater Determinant is a determinant of spin orbitals as shown below:

$$\Psi = \frac{1}{\sqrt{N!}} \begin{bmatrix} \chi_1(x_1) & \cdots & \chi_N(x_1) \\ \vdots & \ddots & \vdots \\ \chi_1(x_N) & \cdots & \chi_N(x_N) \end{bmatrix} \quad (1.11)$$

This form satisfies the antisymmetry requirement for any orbitals and is a more sophisticated statement of the Pauli Exclusion Principle, which is a consequence of the antisymmetry principle. Now that we have a form for the wavefunction and a simplified notation for the Hamiltonian, we can start to calculate the molecular orbitals.

First, the energy of this N-body system is given by the usual quantum mechanical expression:

$$E_{el} = \langle \Psi | H_{el} | \Psi \rangle \quad (1.12)$$

For symmetric energy expressions, we can apply the variational theorem, which states that the energy is always an upper bound to the true energy. Hence, we can obtain better approximate wavefunctions Ψ by varying their parameters until we minimize the energy within the given functional space. Therefore, the correct molecular orbitals are those that minimize the electronic energy $E_{el.}$ The molecular orbitals can be expanded as a linear combination of a set of given basis functions and named the "atomic orbital" basis set.

Now, we rewrite the HF energy E_{el} in terms of integrals of the one- and two-electron operators:

$$E_{HF} = \sum_i \langle i|h|j\rangle + \frac{1}{2}\sum_{ij}[ii|jj] - [ij|ji] \qquad (1.13)$$

Here the first term is the one-electron integral:

$$\langle i|h|j\rangle = \int dx_1 \chi_i^*(x_1) h(r_1) \chi_j(x_1) \qquad (1.14)$$

The a two-electron integral is

$$[ij|ji] = \int dx_1 dx_2 \chi_i^*(x_1)\chi_j(x_1)\frac{1}{r_{12}}\chi_k^*(x_2)\chi_l(x_2) \qquad (1.15)$$

In the next step, we minimize the HF energy expression with respect to changes in the orbitals $\chi_i \rightarrow \chi_i + \delta\chi_i$. We set $\delta E_{HF}[\{\chi_i\}] = 0$ to try to get the minimum energy value with respect to a small change to χ_i, and working through some algebra, we eventually arrive at the HF equations defining the orbitals:

$$h(x_1)\chi_i(x_1) + \sum_{j\neq i}\left[\int dx_2 \left|\chi_j(x_2)\right|^2 r_{12}^{-1}\right]\chi_i(x_1) - \sum_{j\neq i}\left[\int dx_2 \chi_j^*(x_2)\chi_i(x_2) r_{12}^{-1}\right]\chi_j(x_1) = \epsilon_i\,\chi_i(x_1) \quad (1.16)$$

Where ϵ_i is the energy eigenvalue associated with orbital χ_i.

The HF equations can be solved numerically. From this equation, we can see that the solutions ϵ_i actually depend on the orbitals; therefore, we cannot solve this equation directly. Instead, we need to guess some initial orbitals and then refine our guesses iteratively through the HF method until it reaches the steady state. For this reason, HF is called a self-consistent-field (SCF) approach.

2.5 Basis sets

A basis set in the computational simulation is a set of functions (also called "basis functions") that consists in linear combinations to represent molecular orbitals. Most of these functions are typically atomic orbitals centered on atoms but can theoretically be any function. In order to describe the electronic states of molecules, we construct wavefunctions for the electronic states by using molecular orbitals. For mathematical representation, a function for a molecular orbital is constructed as below:

$$\Psi_i = \sum_j c_{ij}\varphi_j \quad (1.17)$$

Ψ_i is a linear combination of other functions, and φ_j provides the basis for representing the molecular orbital.

The ultimate goal for scientists is to create a description of electrons in molecules that enables the calculation of the properties of a multi-body system.

In present-day computational chemistry, quantum chemical calculations are usually performed using a finite set of basis functions. For a molecular simulation, parameters of the basis functions and the coefficients in a linear combination can be optimized in terms of the Variational Theorem to produce a SCF for the electrons. According to the Variational Theorem, the calculated energy is always higher than true energy. Therefore, optimization means that the ground-state energy calculated with the wavefunction is minimized with respect to the variation in the parameters and coefficients defining the function. As a result, if we can find the coefficients with the lowest energy of the wave function, we can get the closest value to the ground-state energy.

Depending on the different basis functions used to form the wavefunctions, there are serval types of basis sets. Three basis sets are provided below as an example.

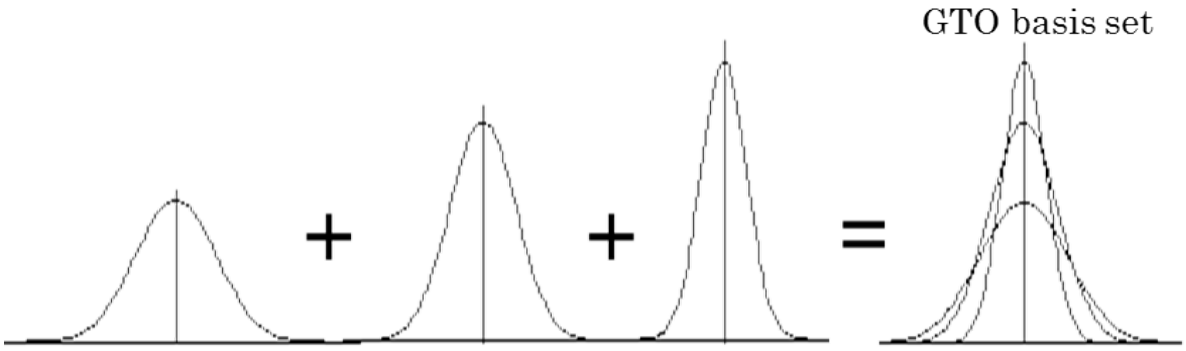
The first type is called a Gaussian Orbital, which consists of a set of Gaussian functions representing the atomic orbital of a molecule. The computation of the integrals is greatly simplified by using Gaussian-type orbitals (GTOs) for the basis functions.

A Gaussian basis function has the form shown in Equation 1.17.

$$G_{nlm}(r,\theta,\psi) = N_n r^{n-1} e^{-\alpha r^2} Y_l^m(\theta,\psi) \qquad (1.18)$$

Note that in all the basis sets, only the radial part of the orbital changes, and the spherical harmonic functions are used in all of them to describe the angular part of the orbital.

Figure 6: Fixed linear combination of Gaussian functions to construct GTO basis.



As shown in Figure 6, GTO basis sets are constructed from fixed linear combinations of Gaussian functions. Gaussian basis sets are identified by abbreviations such as N-MPG*. N is the number of Gaussian primitives used for each inner-shell orbital. The hyphen indicates a split basis set, where the valence orbitals are double zeta. M indicates the number of primitives that form the large zeta function (for the inner valence region), and P indicates the number that forms the small zeta function (for the outer valence region). G identifies the set as being Gaussian. The addition of an asterisk to this notation means that a single set of Gaussian 3d polarization functions is included.

For example, 3-21G means each inner shell is a linear combination of three primitives, and each valence shell is constructed with two sizes of basis function (Two GTOs for contracted valence orbitals; One GTO for extended valence orbitals). Accordingly, there are total of nine functions in a 3-21G basis set.

The second type of basis set is aug-cc-pvdz, also called ACCD. These are Dunning's correlation-consistent basis sets, introduced by T.H. Dunning [7, 8]. They have had redundant functions removed and have been rotated in order to increase computational efficiency.

Polarized basis sets (POLs) were developed by Sadlej et al. [9]. They were designed to improve the calculation of first- and second-order molecular properties. They consist of a

standard double-zeta GTO basis with a set of extra functions derived from derivatives of the outer valence function of the original set.

2.6 Chemical Simulation Codes

2.6.1 GAMESS

Quantum chemistry computer codes are used in computational chemistry to implement the methods of Quantum Chemistry. Most of these programs include the HF method and some post-HF methods. Some of them might use density functional theory (DFT), molecular mechanics (MD), or semi-empirical quantum chemistry methods. The programs include both open-source and commercial software. Most of them are large, often containing several separate programs, and have been developed over many years.

The open-source software GAMESS [10, 11] is famous for general *ab initio* Quantum Chemistry computation. Briefly, GAMESS can compute SCF wavefunctions from RHF, ROHF, UHF, GVB, and MCSCF. Correlation corrections to these SCF wavefunctions include configuration interaction, second-order perturbation theory, and coupled-cluster approaches, as well as the DFT approximation. Excited states can be computed by CI, EOM, or TD-DFT procedures. Nuclear gradients are available for automatic geometry optimization, transition state searches, or the reaction path following. Computation of Hessian energy permits the prediction of vibrational frequencies with IR or Raman intensities. Solvent effects may be modeled by discrete effective fragment potentials or continuum models such as the Polarizable Continuum Model. Numerous relativistic computations are available, including infinite-order two-component scalar relativity corrections, with various spin-orbit coupling options. The Fragment Molecular Orbital Method permits use of many of these sophisticated treatments to be used on very large systems by dividing the computation into small fragments.

A variety of molecular properties, from simple dipole moments to frequency-dependent hyperpolarizabilities may be computed. Many basis sets are stored internally, together with effective core potentials or model core potentials, so that essentially the entire periodic table can be considered [12].

Using GAMESS calculate the polarizabilities of a molecule, we can determine the corresponding Raman spectrum.

2.6.2 GAUSSIAN

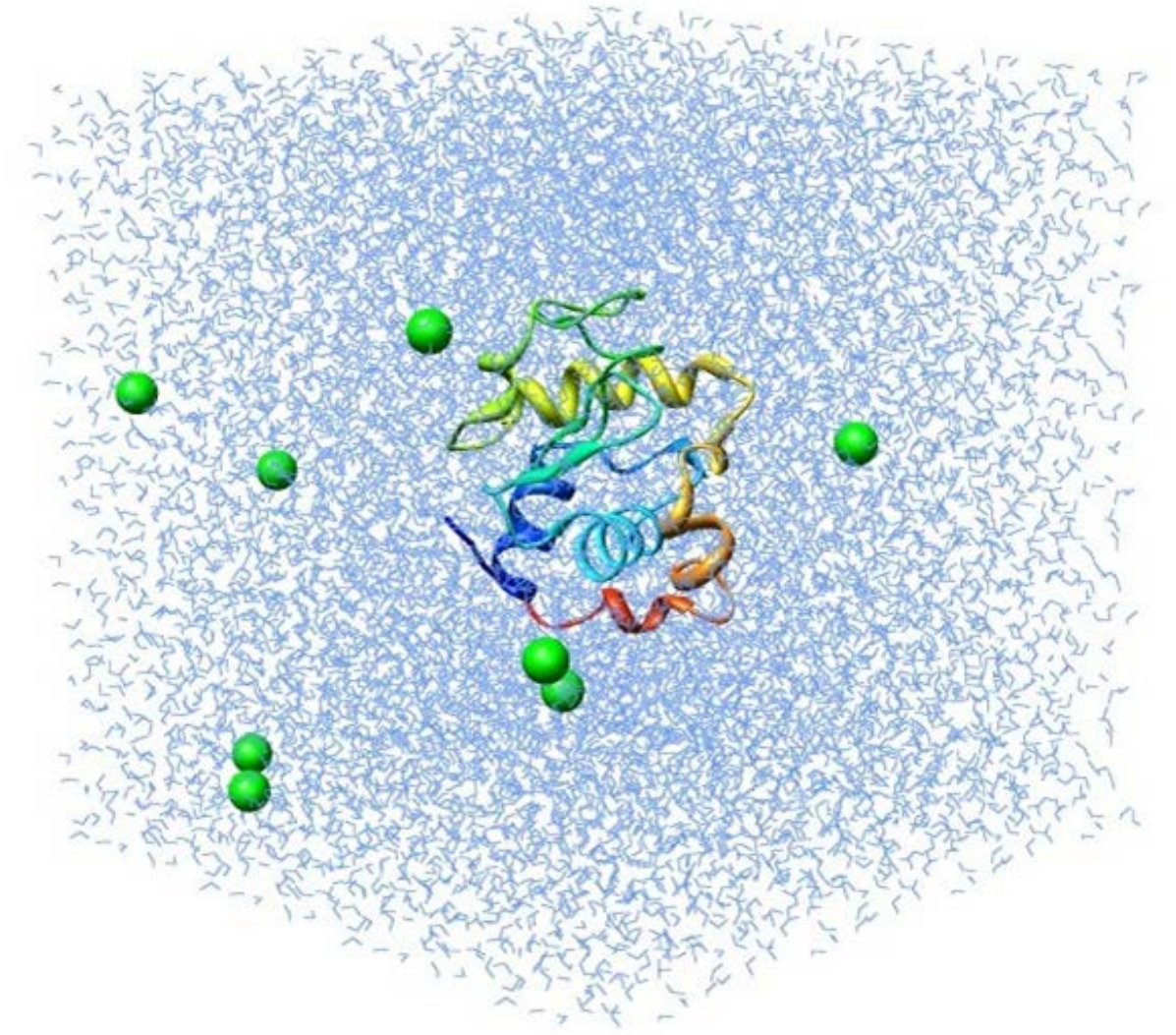
The other powerful quantum computational software is called GAUSSIAN. Right now, the latest version of GAUSSIAN is Gaussian 16. As the most widespread commercial quantum computational software, Gaussian 16 provides a wide-ranging suite of the most advanced modeling capabilities available. It can be used to investigate real-world chemical problems in all of their complexity, even on modest computer hardware. Starting from the fundamental laws of quantum mechanics, Gaussian 16 predicts the energies, molecular structures, vibrational frequencies, and molecular properties of compounds and reactions in a wide variety of chemical environments. Gaussian 16's models can be applied to both stable species and compounds that are difficult or impossible to observe experimentally whether due to their nature (e.g., toxicity, combustibility, or radioactivity) or the inherently fleeting nature (e.g., short-lived intermediates and transition structures). Besides the functions mentioned above, Gaussian 16 can predict a variety of spectra in both the gas phase and in solution, including IR and Raman, and spin-spin coupling constants, vibrational circular dichroism (VCD), Raman optical activity (ROA), and resonance Raman. It can also perform an anharmonic analysis for IR, Raman, VCD, and ROA spectra [13].

2.6.3 GROMACS

The last computational simulation software mainly performs molecular dynamics. It is called GROningen MAchine for Chemical Simulations (GROMACS), and it is a versatile package used to perform molecular dynamics, i.e. simulate the Newtonian equations of motion for systems with hundreds to millions of particles. It is primarily designed for biochemical molecules like proteins, lipids, and nucleic acids that have many complicated bonded interactions. However, since GROMACS is extremely fast at calculating non-bonded interactions (which usually dominate simulations), many groups also use it for research on non-biological systems, e.g. polymers. It supports all the usual algorithms one expects from a modern molecular dynamics implementation. Its code can be run in parallel computation, using either the standard MPI communication protocol, or via "Thread MPI" library for single-node workstations. The package includes a fully automated topology builder for proteins and even multimeric structures. The building blocks are available for the 20 standard and some modified amino acid residues, the four nucleotide and four deoxynucleotide residues, several sugars and lipids, and some special groups and several small molecules [14].

In Figure 7 below, a lysozyme in water's molecular dynamics is simulated by MD method using GROMACS. The molecule in green and yellow at the center of the box is the lysozyme molecule, and the small blue triangles represent the water molecules. This process simulates the evolution of a lysozyme molecule interacting with water molecules over the duration of 1 ns.

Figure 7: Illustration of a simulation of lysozyme-water system

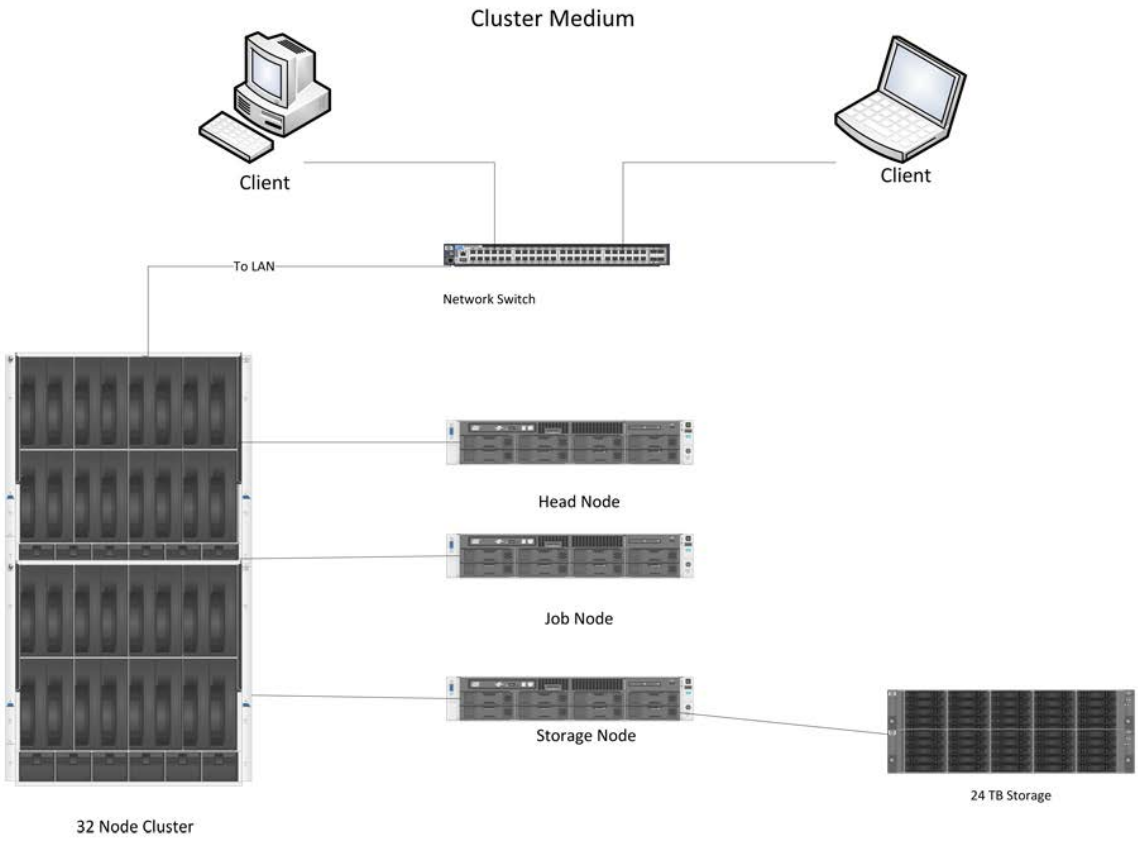


2.7 High-Performance Computing Cluster

Simulation of biological molecules often involves enormous data, and needs to consume significantly large computational sources. The single personal computer cannot provide enough computational sources. In order to run the molecular dynamics simulation, we need to found a powerful platform. The HPC cluster is a kind of computer which can provide powerful computational abilities. It consists of hundreds to tens of thousands multi-core processors. It allows scientists and engineers to solve complex science, engineering, and business problems using code that requires high-bandwidth networking and very high computing capabilities. A HPC cluster consists of many cores and processors, large amounts of memory, high-speed interprocessor communication networking, and large data storages—all shared across many rack-mounted servers. With to these characteristics, a HPC cluster is usually used to run computationally intense tasks such as the simulation of physical phenomena for studying climate change and galaxy formation.

Figure 8 shows the architecture of a HPC cluster with 32 nodes. As we can see, there are several parts. First, clients can remotely log in to the cluster through the network switch. Once they log in to the cluster, they can send a task to the head node, whose main function is to distribute the task to several jobs and ask the job node to do the computation. During the computation, data are stored in the storage node, and once the task is done, the result is sent to the head node; the client can download the results from the head node. This is the processes for utilizing the HPC cluster.

Figure 8: Basic components of a HPC cluster.



Chapter III: METHODS

3.1 The GAMESS

The theoretical Raman vibrational spectra of methane, water, water dimer, and glycine molecules were performed using a restricted HF method on an Intel® Xeon E5 high-performance computer using the GAMESS-US 2017 [10] program package.

The whole simulation process can be divided into three parts: 1) generate the input files for each object molecule, 2) upload the input file to the cluster and run the simulation, and 3) download the results from the cluster and analyze the data with graphic user interface (GUI) software on a personal computer.

In the input file preparation part, we need to generate input files for methane, water, and glycine. The input file is a script with several groups to describe parameters for the simulation process. We used some basic groups in our simulation which consists of four parts:(1) the control group described the job control parameters; and (2) the system group took into account the time and memory allocated to the job; and (3) the basis group described basis sets to be used for the simulation system; and (4) the data group described geometry using a set of Cartesian coordinate data. To obtain the Cartesian coordinate structural data for molecules, we drew the molecular structures with the GUI software called Avogado and output the Cartesian data for each atom. This structure was considered the initial structure data to calculate the geometry optimization (RUNTYP = OPTIMIZE) using the closed-shell (SCFTYP = RHF) method and the three different basis sets. According to Quinet's work [15, 16], the simulated Raman spectrum can be calculated by derivatives of frequency-dependent polarizability with respect to atomic Cartesian coordinates. We then replace the initial structure data with by the optimized geometries' structure data, and our input files were ready for simulations.

We transferred the input file to our cluster and submitted the job using a gms command. The basis sets 3-21G, ACCD, and Sadlef were used in the calculations, keeping the SCF density convergence criterion at 10^{-6} Hartree/Bohr. During the computational process, the head node distributed tasks to compute nodes and made sure to use the advantage of the parallel algorithm to reduce computation time. Subject to the complexity of the objective molecule, the computation time can take from several minutes up to a week. Once the simulations were completed, the output data were output to a file with a ".log" extension.

We downloaded the output file from our cluster to the personal computer and ran a cross-check with the file in the cluster to inspect the completeness of the file. After making sure the data were complete, we imported these data to MacMolPlt and Matlab for analysis. MacMolPlt is an open-source, cross-platform (Mac OS X, Linux, and Windows) GUI for preparing, submitting, and visualizing input and output for the GAMESS quantum chemistry package. Features include a graphical molecule builder, GAMESS input generation, animation of output, and visualization of molecules, normal modes, orbitals, and other properties. All 3-D illustrated molecular models in this paper were made using this GUI.

By following the steps, the simulated Raman vibrational spectra of methane, water, water dimer, and glycine were generated and processed. Results and discussions will be presented in the following chapter.

3.2 The GROMACS

We performed simulations of the of glycine aqueous solution evolution with the GROMACS 1.5.4 software package on our HPC cluster. There are five steps for this simulation: Prepare the topology file; define the box and solvate; add ions; minimize the energy; and run the MD simulation.

The first step was to prepare the topology file for glycine. The topology file with the filename extension “pdb.” could be found on the Research Collaboratory for Structural Bioinformatics (RCSB) website. A topology (topol.top by default) contains all the information necessary to define a molecule within a simulation, which includes non-bonded parameters (atom types and charges) and bonded parameters (bonds, angles, and dihedrals). By executing the pdb2gmx command, the force field (the all-atom OPLS force field) was written to the topology.

After the topology file is generated, the next step is to limit the system within a space. Defining a box with a periodic boundary condition and filling the box with water molecules. By using the editconf module, we defined a 1nm×1nm×1nm cubic space for filling with water molecules.

After the system having a charged molecule, a tool called “genion” within GROMACS is used to add some ions to our system. This step was necessary to make the net charge of the system becomes zero; otherwise, the later iterative process would add net charge infinitely, which would not happen in the real system.

Before we began to run the molecular dynamic with the solvated, electroneutral system, we needed to ensure that the system had no steric clashes or inappropriate geometry. This process is called energy minimization, which can be done through the GROMACS MD engine, “mdrun.”

After these steps, we performed an MD simulation for a duration time of 1 ns.

The result was transferred back to the personal computer and analyzed using another GUI viewer called PyMOL. PyMOL is a molecular visualization system that can produce high-quality 3-D images of small molecules and biological macromolecules such as proteins. According to

the original author (Warren Lyford DeLano) [17], by 2009, almost a quarter of all published images of 3-D protein structures in the scientific literature were made using PyMOL.

CHAPTER IV: RESULTS AND DISSCUSSION

4.1 Result of CH_4 Simulation

4.1.1 The Geometric Structure

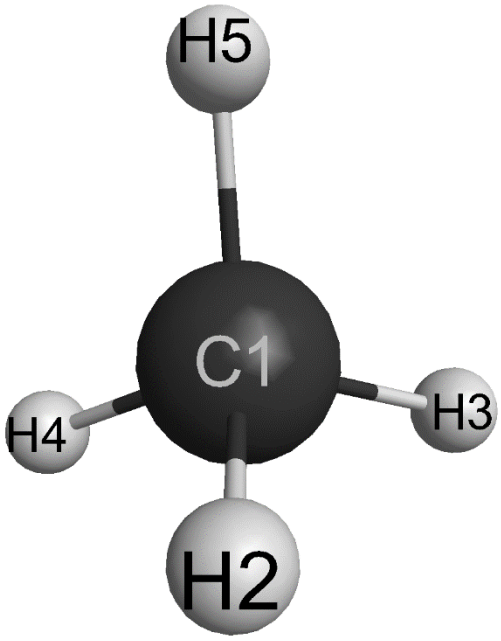
The optimized geometric structures calculated by HF with three different basis sets, 3-21G, ACCD, and Sadlef, are shown in Table 1 below. The corresponding 3-D model of methane is shown in Figure 9.

Table 1: Optimized geometrical parameters of methane at HF with 3-21, ACCD, and Sadlef basis sets and experimental data

	3-21G	ACCD	Sadlef	Expt.[1]
		Bond lengths (Å)		
C1-H2	1.0828455	1.0893858	1.0900152	
C1-H3	1.0828455	1.0893858	1.0900152	
C1-H4	1.0828455	1.0893858	1.0900152	1.092±0.014
C1-H5	1.0828455	1.0893858	1.0900152	
		Bond angles (°)		
H2-C1-H3	109.4734021	109.4727727	109.4728897	
H2-C1-H4	109.4701303	109.4704498	109.470386	
H2-C1-H5	109.4701303	109.4704498	109.470386	
H3-C1-H4	109.4701303	109.4704498	109.470386	110.62±4.47
H3-C1-H5	109.4701303	109.4704498	109.470386	
H4-C1-H5	109.4734021	109.4727727	109.4728897	

1 Experimental length and angle data for C-H bond from Computational Chemistry and Benchmark Data Base NIST.

Figure 9: 3-D model illustrate the structure of methane (CH_4)



The experimental value of the bond lengths and angles of methane is the average C-H bond taken from the Computational Chemistry Comparison and Benchmark DataBase (CCCBDB). By comparing the optimized parameter with the experimental data, it was found that the optimized bond lengths of C-H fell in the range of 1.0828455–1.0900152 Å for these three basis sets, which is within the range of experimental bond lengths (1.078–1.106 Å). The optimized bond angle of C-H fits in the range of 109.4701303–109.4728897 °, which is also in agreement with the experimental data. Although the values of optimized bond length and bond angles agreed with the experimental values, they were a little bit less than the average value, which are 1.092 Å for length and 110.62 ° for angle which was introduced by the fact that the HF method of calculation underestimates the bond length and angle, as shown by Lee [18]. Our optimized data show that the bond length for each basis set was identical, although the bond angle was not. In the table 1, it was found that the bond angles H2-C1-H3 and H4-C1-H5 were

always a little larger than the rest of the angles for each basis set. Hence we concluded that although there was some difference between the calculated and experimental values, the optimized geometrical parameters can well reproduce the real methane molecule, and they are the basis used in the following discussion.

4.1.2 Fundamental Modes

Methane formed by five atoms has nine different normal modes of vibration. Since the HF method did not take into consideration the anharmonicity during the calculation process [19], the calculated frequency for each basis set was overestimated by comparing it with the experimental data. In order to improve the agreement with the observed frequency, we scaled down the calculated frequencies with a scale factor. Table 2 lists the scaled frequencies for methane compared with its experimental frequencies. This table also lists the vibrational assignment for each frequency.

Table 2 shows that after being scaled by the scale factor, the 3-21G basis set is a little off; however, the data for the other two basis sets had a good agreement with the experimental data. The largest deviation using the ACCD and Sadlef basis sets was 18.48 cm^{-1}. Basis set 3-21G, had a slightly larger deviation, around 21.46-75.64 cm^{-1} compared to the experimental data. The scale factors c for the 3-21G, ACCD, and the Sadlef basis sets were calculated using equation $c = \Sigma(\nu_i \times \omega_i)/\Sigma(\omega_i^2)$ where ν_i is the experimentally observed vibrational frequencies, and ω_i is the theoretical vibrational frequencies. It gives the results for three basis sets in 0.9085, 0.9255, and 0.9238, respectively.

Table 2: Methane theoretical values and experimental results

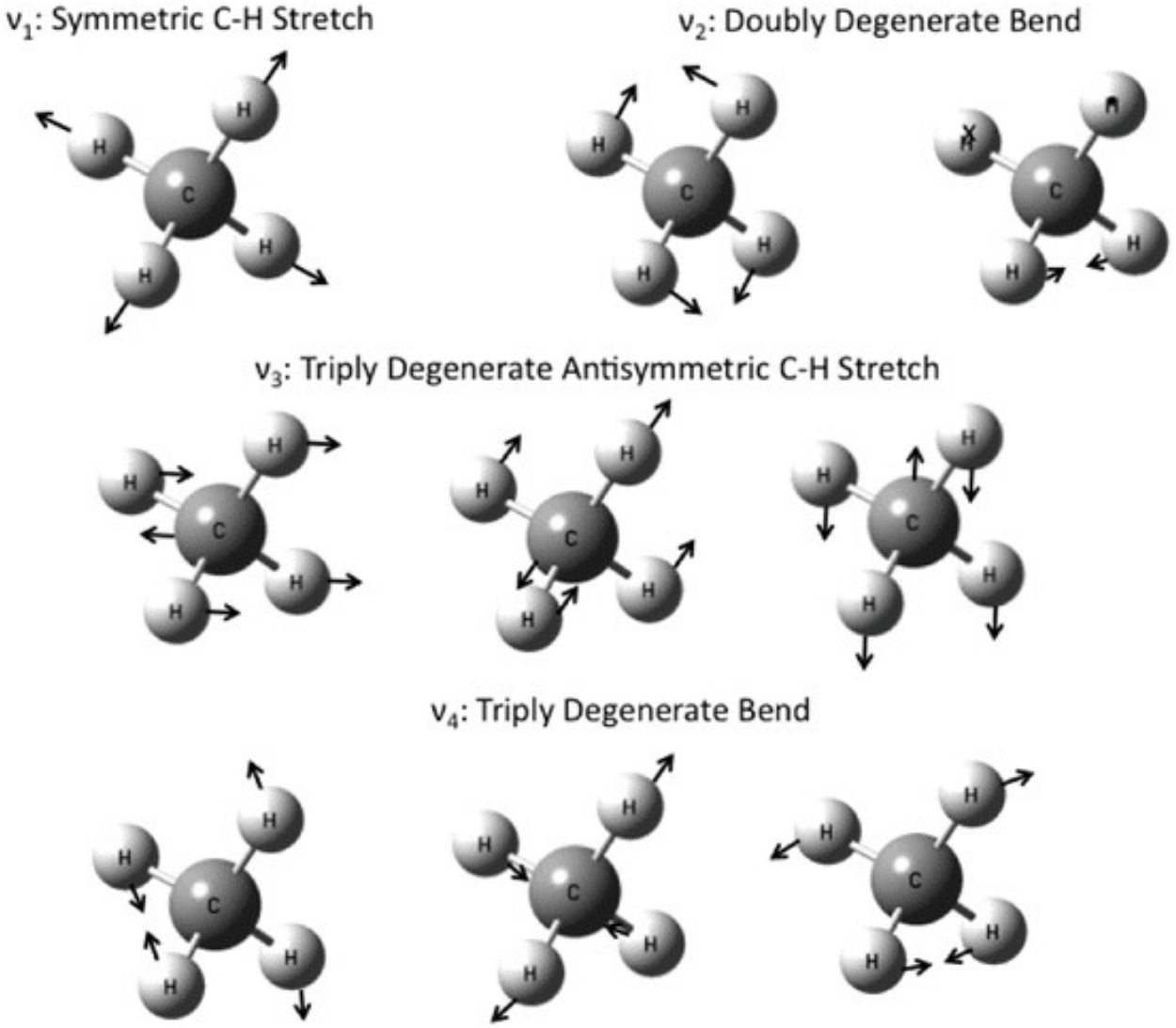
Mode No.	Basis sets									Expt. Data[1]	Vibrational assignments
	3-21			ACCD			Sadlef				
	Calc. freq.	Scaled freq.	Calc. Raman peak	Calc. freq.	Scaled freq.	Calc. Raman peak	Calc. freq.	Scaled freq.	Calc. Raman peak		
1	1520.2	1381.04	9.28	1423.6	1317.47	0.03	1430.5	1321.52	0.12	1306.0[2]	Triply Degenerate Bend (v_4)
2	1739.8	1580.54	78.38	1637.6	1515.52	7.78	1657.1	1530.86	8.67	1534.0[3]	Doubly Degenerate Bend(v_2)
3	3187.3	2895.54	122.17	3152.7	2917.67	226.85	3147.8	2908.00	241.09	2917.0[3]	Symmetric C-H Stretch(v_1)
4	3280.7	2980.39	177.50	3266.1	3022.62	160.44	3271.7	3022.46	157.21	3019.0[2]	Triply Degenerate Antisymmetric C-H Stretch(v_3)

1 Experimental vibrational data for methane from Computational Chemistry and Benchmark Data Base NIST.

2 Vibrational Intensities in Infrared and Raman Spectroscopy " WB Person, G Zerbi, ed. Elsevier, Amsterdam,1982.

3 Shimanouchi, T. , Tables of Molecular Vibrational Frequencies, Consolidated Volume 1, NSRD NBS-39.

Figure 10: Fundamental vibrational modes of methane (CH_4)



In Figure 10, the nine vibrational modes of methane are distributed as one symmetric C-H stretching (v_1), two degenerate bendings (v_2), three degenerate antisymmetric C-H stretchings (v_3), and three degenerate bending. From data of Shimanouchi [20], there are four types of vibrational modes of CH_4. However, Jourdannequ's [21] work shows that for the Dyad region, with a range from 850 to 2,000 cm^{-1}, the vibrational frequency near 1,530 cm^{-1} dominates. This also coincides with our calculated result, where the Raman intensity around a frequency of 1,306 cm^{-1} is almost close to zero in ACCD and Sadlef basis sets.

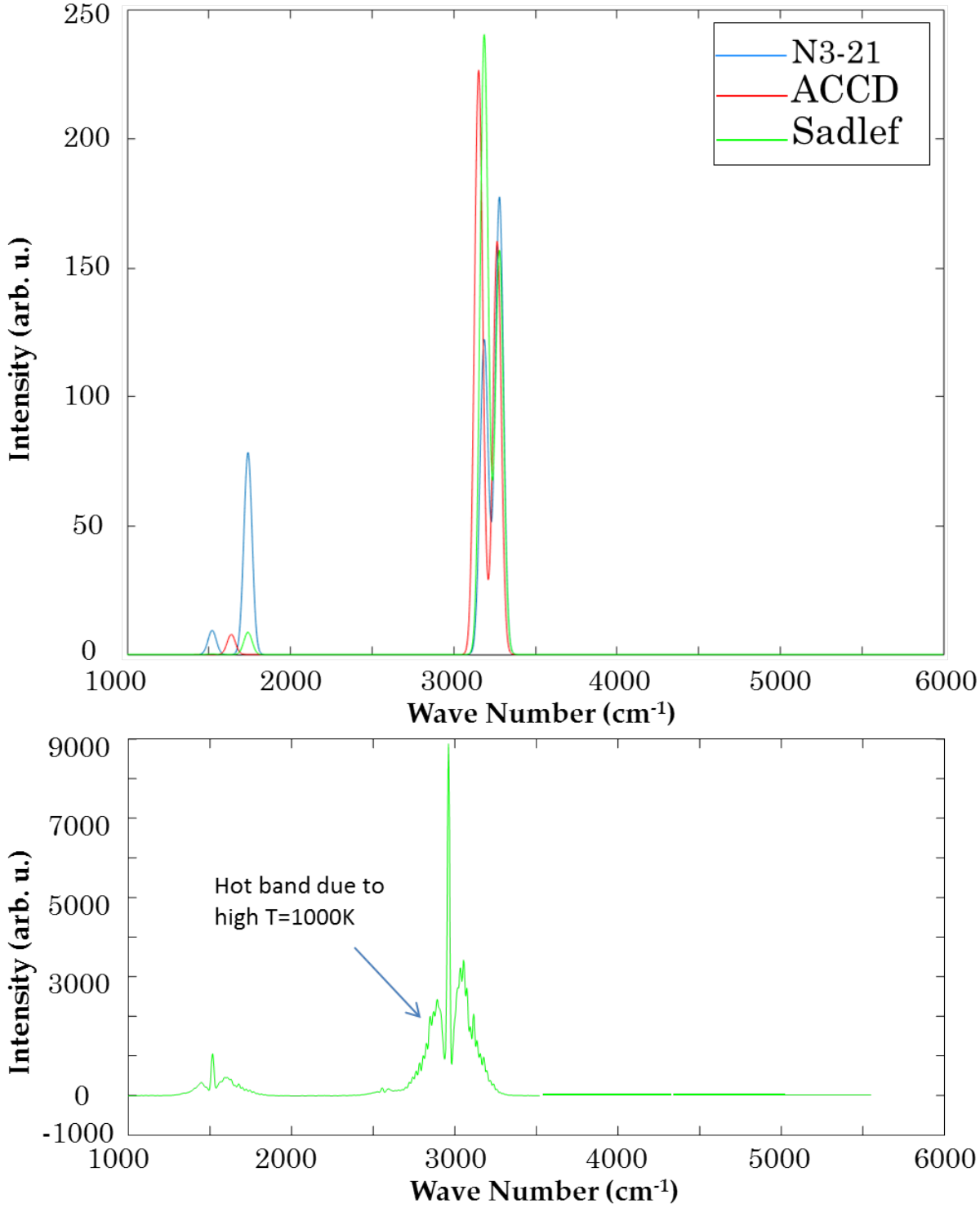
4.1.3 Simulated Raman Spectrum

Using the raw data shown above, a Raman spectrum (Figure 11) is obtained and plotted by using Matlab with a linewidth of 25cm^{-1} for each basis. In Figure 11, three simulated Raman spectra are plotted with different basis sets in three colors. The shape of the blue line, which belongs to the 3-21G basis set, is different from the red line, which belongs to ACCD, and the green plot, which belongs to the Sadlef basis set. Around 1,500 cm^{-1}, there is a small peak only displayed with the 3-21G basis. This band theoretically belongs to the triply degenerate bending vibrational mode at 1,380 cm^{-1}. However, in the experimental data, as shown below, this vibration was very weak. 3-21G also has a sharp peak at a frequency of 1,739.8 cm^{-1}. Comparing with the other two basis sets, we found that the spectra of ACCD and Sadlef are almost the same except at a low frequency of around 1,500–1,600 cm^{-1}.

The bottom part of Figure 11 below shows the experimental Raman spectrum at a temperature of 1000 K, which has a hot band around 2,900 cm^{-1}. The hot band lies in the region where the intensity of the Raman spectrum proportionally increases as the environmental temperature rises which is why there is an intense peak around 2,900 cm^{-1} in the experimental data. Comparing the shape of the calculated and experimental spectra, we found that the calculated spectrum has the same three strong peaks as the experimental spectrum, but the frequencies of three strong lines were all higher.

This situation occurred because of the method of calculation we used. *Ab initio* harmonic vibrational frequencies are typically larger than the fundamentals observed experimentally, since in the theoretical treatment anharmonicity effects are in general overlooked [22]. Anthony introduces a scale factor to correct the overestimated harmonic vibrational frequencies. The scale factor is relatively uniform, only varying for different basis sets.

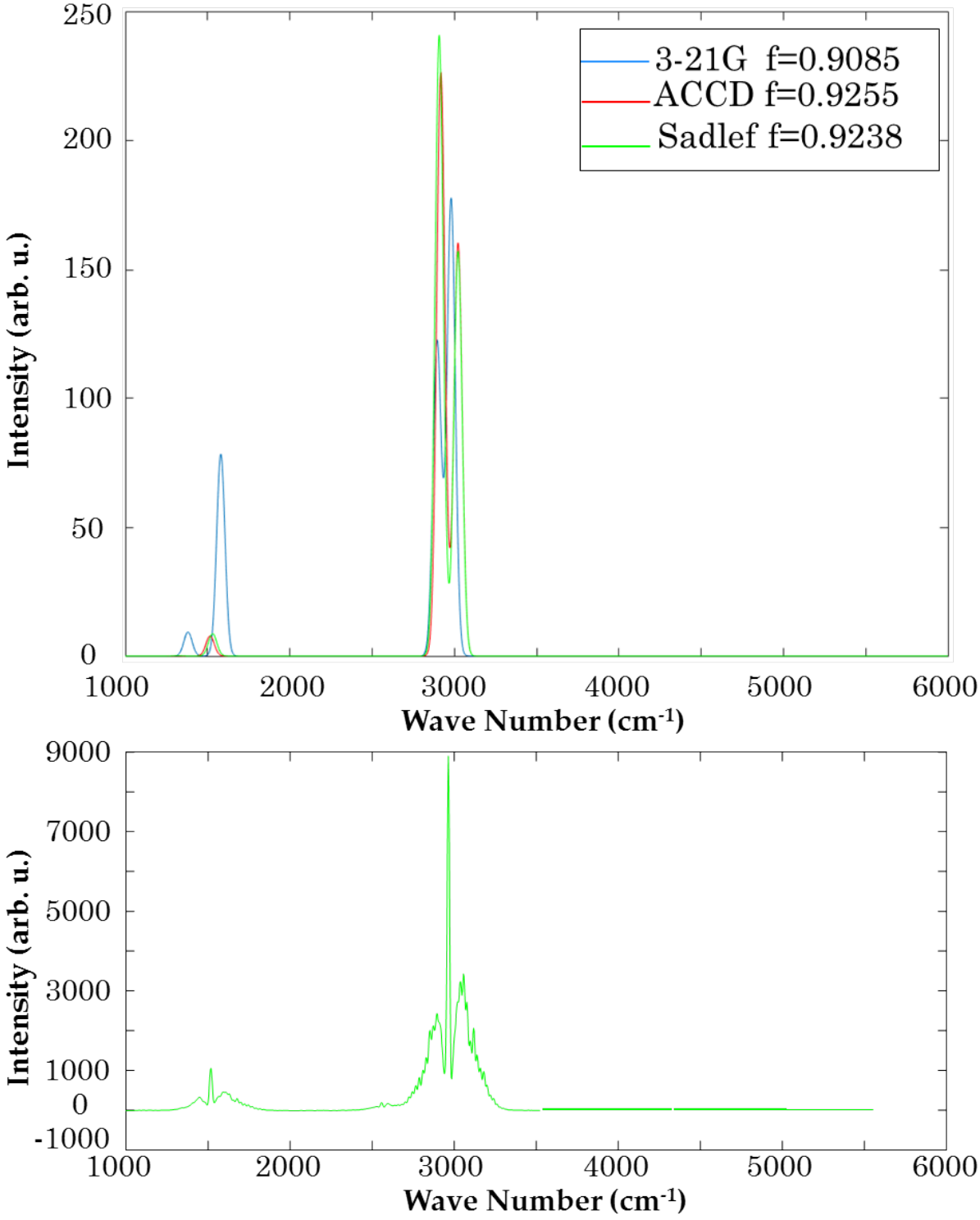
Figure 11: Unscaled simulated Raman spectrum for methane and published Raman spectrum for methane.



The method of calculating the scale factor for this study has been discussed in the theory part above. After applying the corresponding scale factor to our basis set data, we compared each calculated vibrational frequency to the experimental result [21]. The Raman spectra are replotted in Figure 12. Table 2 shows the frequency difference between the exponential and calculated frequencies before and after being scaled. From this table, we can see that the ACCD and Sadlef basis sets have good agreement with the observed data. We calculated the average error for the Sadlef basis set as 8.23 cm^{-1} and for ACCD as 8.71 cm^{-1}, comparing to 185.6 cm^{-1} and 176.5 cm^{-1} before applying the scale factor. However, basis set 3-21G did not match well with the observed data, especially at low frequencies.

We replotted the scaled Raman spectra and compared it to the experimental methane Raman spectrum. In Figure 12, it shows that at a low frequency, the 3-21G basis set has two peaks, neither of which match with the experimentally observed peak. At a high frequency of around 3,000 cm^{-1}, 3-21G still has a good agreement with the experimental data, just slightly lower than the observed frequency. Sadlef and ACCD almost overlap each other, and both have good agreements with the experimental spectrum for both the bending and stretching modes after applying the scale factor.

Figure 13: Simulated Raman spectrum for methane after scaling (top) and the published results for methane (bottom)



4.2 Result of H_2O Simulation

The reason why we chose water as one of the research objects is that a water molecule has the simplest structure of all the molecules. There are just three vibrational modes for a single water molecule. Fewer vibrational modes mean it is easier to analyze and assign a peak to each mode. Since the water molecule just has two O-H bonds, it is also easier to analyze its properties and more clearly determine which peak belongs to the O-H bond.

4.2.1 The Geometric Structure

The optimized geometric structure calculated by the HF method with three different basis sets, 3-21G, ACCD, and Sadlef, are listed in Table 3 and compared with the experimental data. The corresponding 3-D model is illustrated in Figure 13.

Table 3: Optimized geometrical parameters of water at HF with 3-21, ACCD, and Sadlef basis sets and experimental data

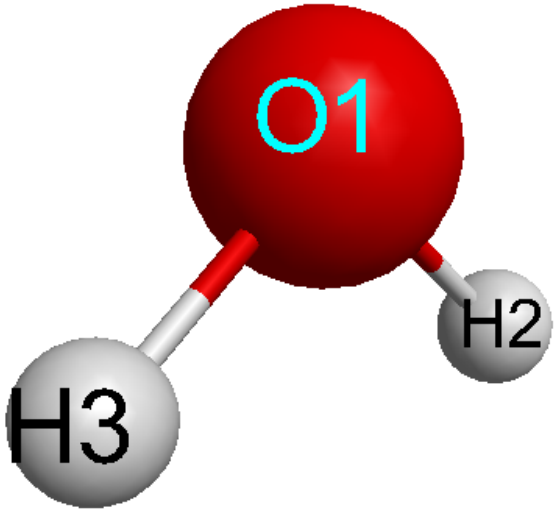
	3-21G	ACCD	Sadlef	Expt.[1]
Bond lengths(Å)				
O1-H2	0.9666589	0.9436271	0.9449685	0.979±0.032
O1-H3	0.9666587	0.9436270	0.9449684	
Bond angles(°)				
H2-O1-H3	107.6889913	105.9309726	105.8507257	107.89±4.82

1 Experimental length and angle data for O-H bond from Computational Chemistry and Benchmark Data Base NIST.

The experimental bond length and bond angle are the average O-H values taken from CCCDB [23]. In table 3, the bond length of O-H falls in the range of 0.9436276–0.9666589 Å. The bond lengths for these three basis sets are within the range of experimental bond lengths (0.947–1.011 Å). The optimized bond angles of O-H also fall in the range of 105.8507257–

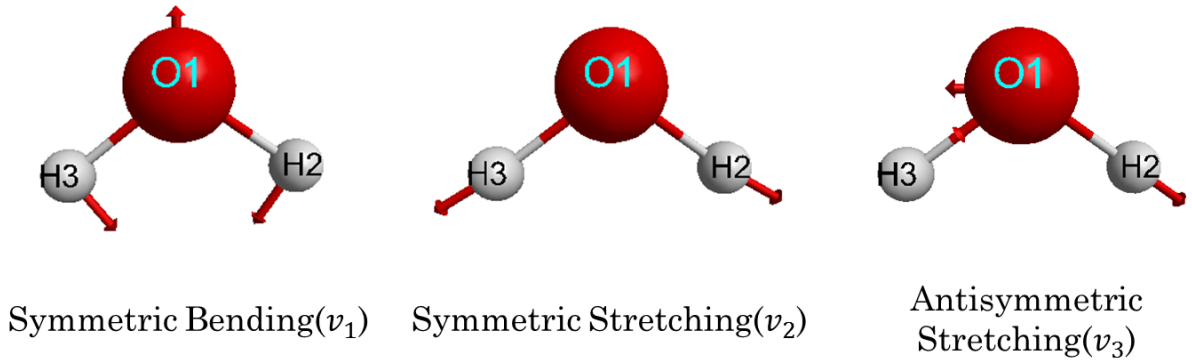
107.6889913°, which is in good agreement with the experimental data [24]. In Figure 13, the 3D model shows that water molecule consists of two Hydrogen atoms (H3 and H2) and one Oxygen atoms (O1) connecting by two covalent bonds (H3-O1 and H2-O1). It is well known that for the ideal H_2O molecule, the H3-O1 and H2-O1 bond length is identical. However, the calculated data shows that the two lengths are slightly different. The largest difference is around 0.000002 Å, which can be ignored. Based on these comparisons, the optimized geometrical parameters will be able to reproduce water molecule's geometrical properties well.

Figure 14: 3-D model illustrates the structure of water (H_2O)



4.2.2 Fundamental Vibrational Modes

Figure 15: Fundamental vibrational modes of water (H_2O)



Since a water molecule consists of one oxygen atom and two hydrogen atoms, with the estimation of 3N-6 (with N as the number of atoms) for vibrational modes, there are three fundamental vibrational modes: one symmetric stretching (v_1), one symmetric bending (v_2), and one anti-symmetric stretching (v_3). There fundamental vibrational modes are illustrated in Figure 14.

Table 4: Water molecule theoretical values and experimental results

Mode No.	Basis sets									Expt. Data[1]	Vibrational assignments
	3-21			ACCD			Sadlef				
	Calc. freq.	Scaled freq.	Calc. Raman peak	Calc. freq.	Scaled freq.	Calc. Raman peak	Calc. freq.	Scaled freq.	Calc. Raman peak		
1	1799.3	1634.66	11.57	1831.6	1695.15	1.14	1819	1680.40	1.83	1595	Symmetric bending (v_1)
2	3812.4	3463.57	97.24	3807.6	3523.93	99.28	3789.9	3501.11	97.47	3657	Symmetric Stretching(v_2)
3	3945.9	3584.85	44.79	3901.9	3611.21	29.09	3878.8	3583.24	30.33	3756	Antisymmetric Stretching(v_3)

1 Experimental vibrational data for methane from Computational Chemistry and Benchmark Data Base NIST.

4.2.3 Simulated Raman Spectrum

Table 4 above lists the calculated frequency, scaled frequency, and experimental frequency [24, 25] for a single water molecule in three different basis sets, where the vibrational assignments for each mode are given at the end. By comparing each scaled frequency to its experimental frequency, it was not hard to find that for each basis set, Mode 2 had a stronger Raman intensity than the other two modes. This occurred because the symmetric stretching vibrational mode is the main vibrational mode in a water molecule.

From these raw data, a simulated unscaled Raman spectrum and scaled Raman spectrum for each basis set were replotted in Figure 15. By comparing the scaled and unscaled Raman spectra, we found the effect of the scale factor on the water molecule and chose the most reasonable basis set for the simulated water molecule.

Before applying the scale factor, the three basis sets showed almost the same intensity and frequency at around 3,800 cm^{-1}. However, for the range from 3,900 cm^{-1} to 4,000 cm^{-1}, 3-21G had a higher frequency and higher intensity than the other two basis sets. For the region around 1,600 cm-1, the three basis sets had similar frequencies, and the 3-21G produces the highest intensity.

To reduce the impacts of the anharmonicity on our simulation results, we applied the scale factor to each basis and plotted the scaled simulated Raman spectrum. We also introduced the experimental spectrum as Figure 16 [26] for liquid water to compare with our simulated data.

Figure 16: The unscaled simulated Raman spectrum and scaled simulated Raman Spectrum for water

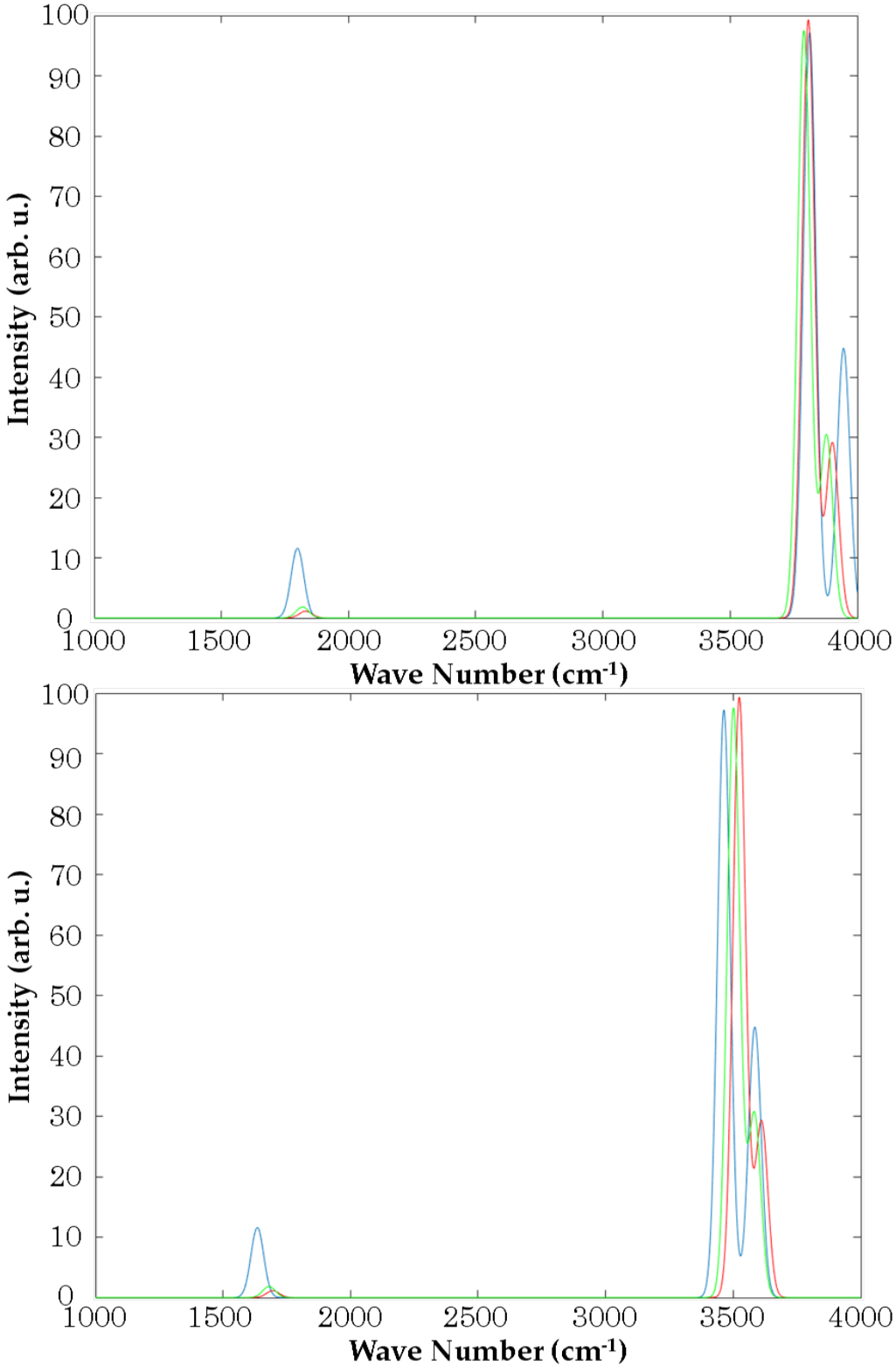
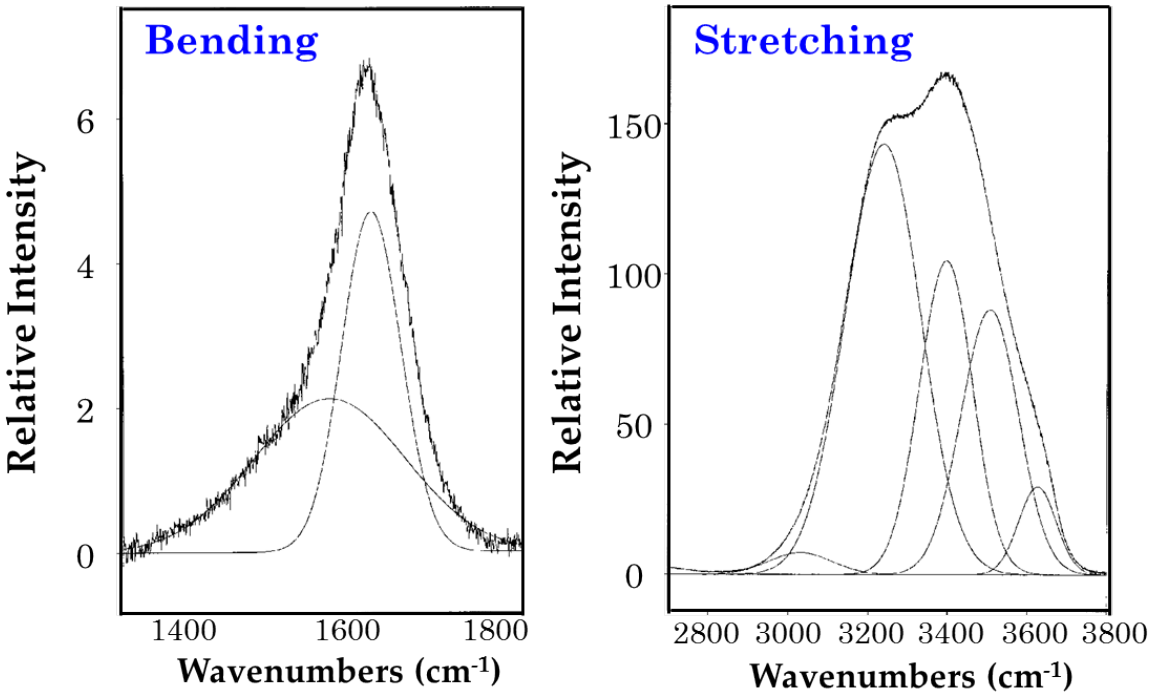


Figure 17: Published Raman Spectrum for water range from 1400-1800cm^{-1} and 2800-3800cm^{-1}



In the region from 1,400 to 1,600 cm^{-1}, 3-21G showed good agreement with the experimental data. Although the two peaks from the experimental data in the range of 3,000–3,600 cm^{-1} also appeared in our simulated results for all basis sets, the shapes of the results did not match. The linewidth for the experimental result, which covered a range of almost 800 cm^{-1}, was much wider than our results. This was because the H_2O molecules in the aqueous environment couple with each other by the intermolecular hydrogen bond [27, 28]. To take the inter-molecular interaction into account, we added a simulation of a water dimer.

4.2.4 Intermolecular Effect

A system of two water molecules interacting through a hydrogen bond is called a water dimer.

Figure 18: 3-D model illustrate the structure of water dimer (H_2O-H_2O)

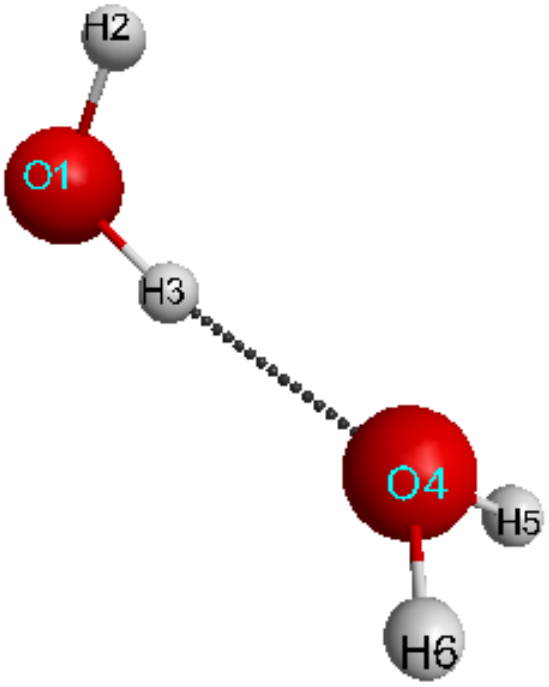


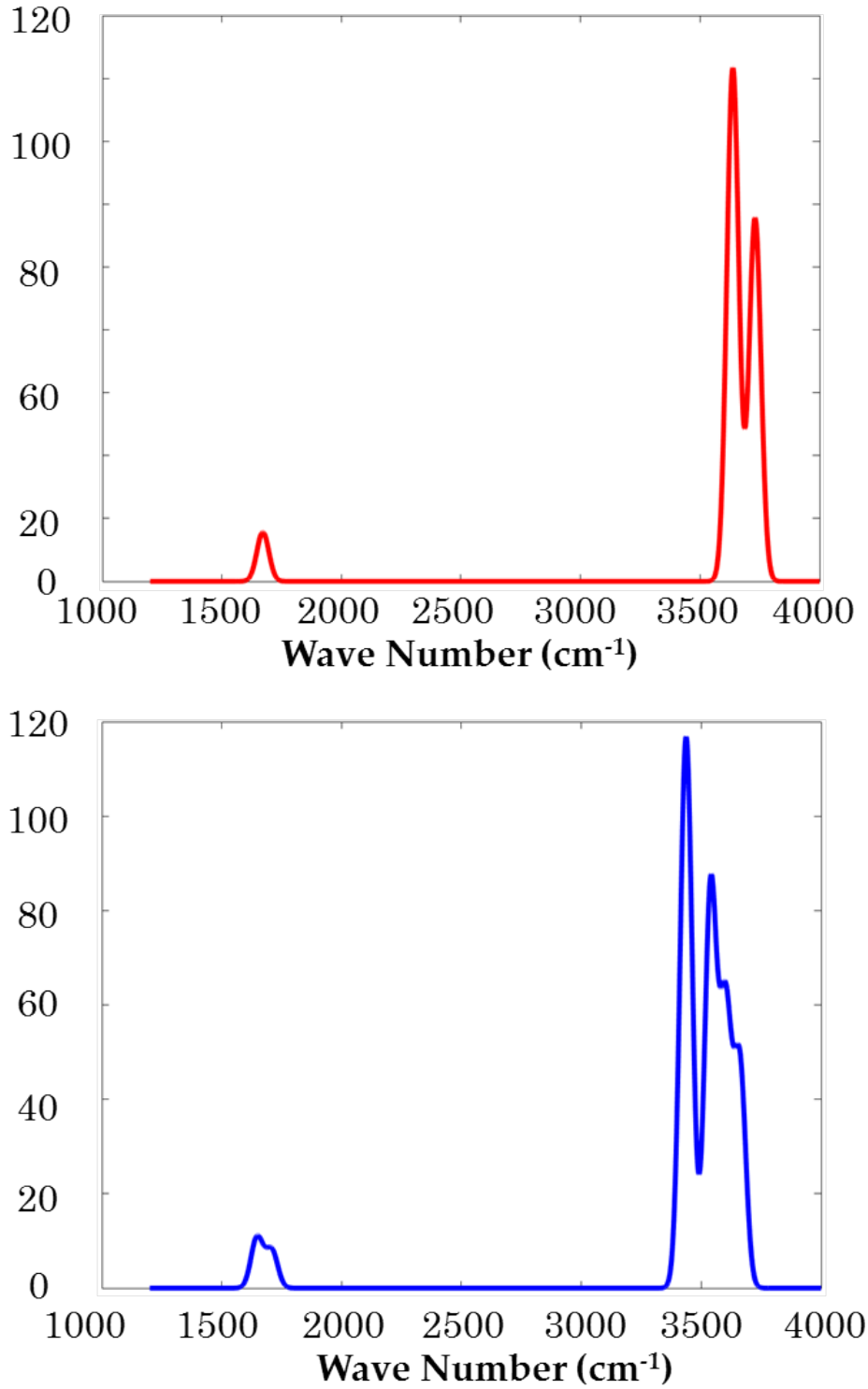
Figure 17 shows a 3-D model for the water dimer structure. In this model, the black dashed line between atoms H3 and O4 represents the hydrogen bond in the water dimer. This intermolecular action occurs normally in liquid water. In a system with more than one water molecule, the water molecule couples with another water molecule through this hydrogen bond, and forms a cluster. Here we consider a water dimer our target base for two reasons. First, a water dimer is less expensive for computation since it consists of just two atoms. Second, by analyzing the simulated Raman spectrum of a water dimer and comparing it to that of a single molecule, we can illustrate the effects of molecule clustering on Raman spectrum.

Figure 18 shows two spectra; the red line is the simulated Raman spectrum for a single H_2O molecule, and the blue line is the simulated Raman spectrum for a water dimer. The 3-21G basis set is used in both of these simulated spectra.

In the bending region, which is around 1,600 cm^{-1}, the linewidth of the water dimer peak increased from around 100 to 200 cm^{-1}. For the stretching region, the linewidth of the water dimer was close to 300 cm^{-1}, compared to 200 cm^{-1} for a single water molecule.

In both regions of the water dimer spectrum, we found a trend of the linewidth increasing as we added the effect of the hydrogen bond. Therefore, we can conclude that if we include more water molecules in a system, the effect of intermolecular coupling will continues to broaden the linewidth and demonstrate better agreement with the experimental data.

Figure 19: Simulated Raman Spectrum for single water molecule (above) and Simulated Raman Spectrum for water dime (bottom)



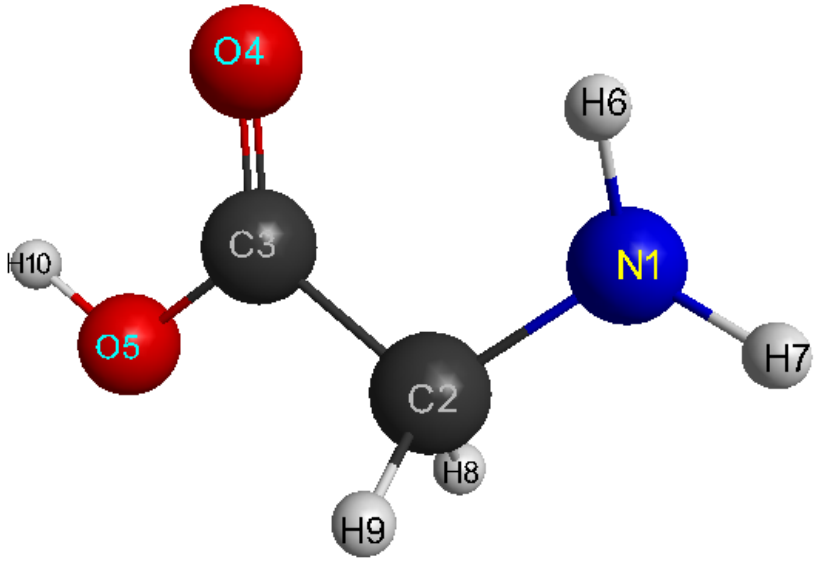
4.3 Result of Glycine Simulation

Amino acids are the basic “building blocks” of proteins and play an important physiological role in all species. To determine the whether there are proteins in a new material, it is necessary to study the Raman spectra of amino acids. Glycine is the simplest amino acid, so it is the first choice for our study.

4.3.1 The Geometric Structure

The basic structure of amino acids consists of three functional groups: one carboxyl group (-COOH), one amine group (-NH_2), and an R group. All these groups connect to a center carbon atom. The R group is specific to each amino acid. For glycine, the R group is just a hydrogen atom. This is why it is the simplest amino acid. Figure 19 illustrates the basic structure of glycine.

Figure 20: 3-D model illustrates the structure of Glycine ($C_2H_5NO_2$)



4.3.2 Fundamental Vibrational Modes

Because Glycine has 10 atoms, the fundamental vibrational modes can be calculated by 3N-6 (with N as the number of atoms). It has 24 fundamental vibrational modes. According to the work done by Santosh Kumar [29], 13 bands can be found in the Raman spectrum, at 497, 602, 697, 893, 1,033, 1,323, 1,410, 1,508, 1,567, 1,667, 2,123, 2,930, and 3,050 cm^{-1}. These bands are listed in Table 5. In Table 5, all 24 calculated modes are listed for the three basis sets, including the calculated frequency and the corresponding intensity. By comparing our calculated data with the experimental data, it was found that all three basis sets have good matches with the experimental data. Out of these 13 bands, two have intense peaks at around 894 and 1,327 cm^{-1}. The band with the peak at 894 cm^{-1} is assigned to the vibration of NH_2 twisting and COOH bending. The other peak, at 1,327 cm^{-1}, is assigned to the vibration of three functional groups: NH_2 twisting, CH_2 twisting, and COOH bending. From the simulated data plot, we can see that Raman peaks at a wavenumber beyond 3,000 cm^{-1} are stronger than those from 100 to 3,000cm^{-1}. Unfortunately, we couldn't find published experimental data beyond 3,000 cm^{-1}. Although, our simulated data below 3,000 cm^{-1} was in good agreement with the observed data for both intensity and frequency.

Table 5: Glycine molecule theoretical values and experimental results

Mode No.	Basis sets									Expt. Data[1]	Vibrational assignments
	3-21			ACCD			Sadlef				
	Calc. freq.	Scaled freq.	Calc. Raman peak	Calc. freq.	Scaled freq.	Calc. Raman peak	Calc. freq.	Scaled freq.	Calc. Raman peak		
1	105.9	96.21	0.44	102.9	95.23	0.1	104.7	96.72	0.09	-	-
2	160.6	145.91	4.23	215.0	198.95	0.15	210.3	194.28	0.15	-	-
3	298.4	271.10	0.32	304.2	281.54	0.18	304.7	281.48	0.18	-	-
4	495.6	450.25	2.62	500.3	463.03	1.26	504.4	465.96	1.27	497	COO^{-}bend+CH_2 bend
5	555.5	504.67	5.34	562.2	520.32	1.31	561	518.25	1.24	-	-
6	620.9	564.09	6.97	669.2	619.34	0.97	672.9	621.63	0.9	602	COOH bend+ NCCO bend
7	691.7	628.41	2.12	715.9	662.57	1.26	718.6	663.84	1.23	697	CH_2 twist+ COOH twist
8	701	636.86	1.19	894.6	827.95	1.65	909.2	839.92	4.17	-	-
9	878.9	798.48	9.8	913.5	845.44	10.05	918.3	848.33	7.49	894	NH_2 twist+ COOH bend
10	1121.7	1019.06	2.24	1121.4	1037.86	0.59	1124.9	1039.18	0.59	-	-

11	1166.7	1059.95	4.41	1185.6	1097.27	4.44	1190.4	1099.69	4.19	1033	C-N stretch+ C-C vibration
12	1220.1	1108.46	0.99	1254.9	1161.41	1.18	1267.1	1170.55	1.09	-	-
13	1305.7	1186.23	6.28	1328.2	1229.25	4.21	1337.9	1235.95	4.03	1327	NH_2twist+CH_2 twist+ COOH bend
14	1377.8	1251.73	14.69	1378.5	1275.80	5.43	1383	1277.62	5.40	-	NH_2twist+CH_2 twist
15	1461.8	1328.05	3.4	1442.6	1335.13	1.36	1453.1	1342.37	1.25	1410	CH_2 scissoring
16	1568.7	1425.16	2.9	1588.3	1469.97	1.62	1599.5	1477.62	1.52	-	-
17	1670.5	1517.65	18.27	1602.7	1483.30	5.42	1608.4	1485.84	5.61	1508	CH_2 bending
18	1821.0	1654.38	5.25	1764.2	1632.77	1.74	1777	1641.59	1.75	-	-
19	1968.1	1788.02	4.81	1989.6	1841.37	7.44	2007.5	1854.53	7.66	1667	C=O stretching
20	3142.8	2855.24	107.35	3129.9	2896.72	115.53	3129.9	2891.40	118.24	2930	Antisymmetric CH_2 stretch
21	3258.2	2960.07	85.10	3241.1	2999.64	77.3	3233.9	2987.48	78.83	3050	Antisymmetric CH_2 stretch
22	3721.2	3380.71	82.56	3743.2	3464.33	100.47	3750.8	3464.99	101.43	-	-

23	3835.1	3484.19	56.32	3833.4	3547.81	42.41	3841.1	3548.41	42.15	-	-
24	3873.7	3519.26	115.79	4110.2	3803.99	71.99	4109.1	3795.99	72.58	-	-

1. Experimental vibrational data for Glycine from Santosh Kumar, Amareshwar K. Rai et al, “Vibrational spectrum of glycine molecule”.

4.3.3 Simulated Raman Spectrum

Using the simulated data in table 5, we plotted the Raman spectrum for three different basis sets in three different colors, shown in Figure 20. In Figure 20, ACCD and Sadlef almost have the same line shape and overlap with each other. The wavenumber simulated by the 3-21G basis set was slightly smaller compared to results from the other two basis sets; except ~ 894 and 1,327 cm^{-1}, where 3-21G showed two strong peaks. Furthermore, 3-21G also showed one intense peak around 1,377.8 cm^{-1}. This peak occurred in all three basis sets, but 3-21G showed the most intense intensity. From the table 5, we found that this peak belongs to the NH_2 twist and CH_2 twist vibration. We will carry out further research to figure out why this peak disappears from in the previous published experimental data.

Figure 21: Simulated Raman Spectrum for Glycine with different basis sets

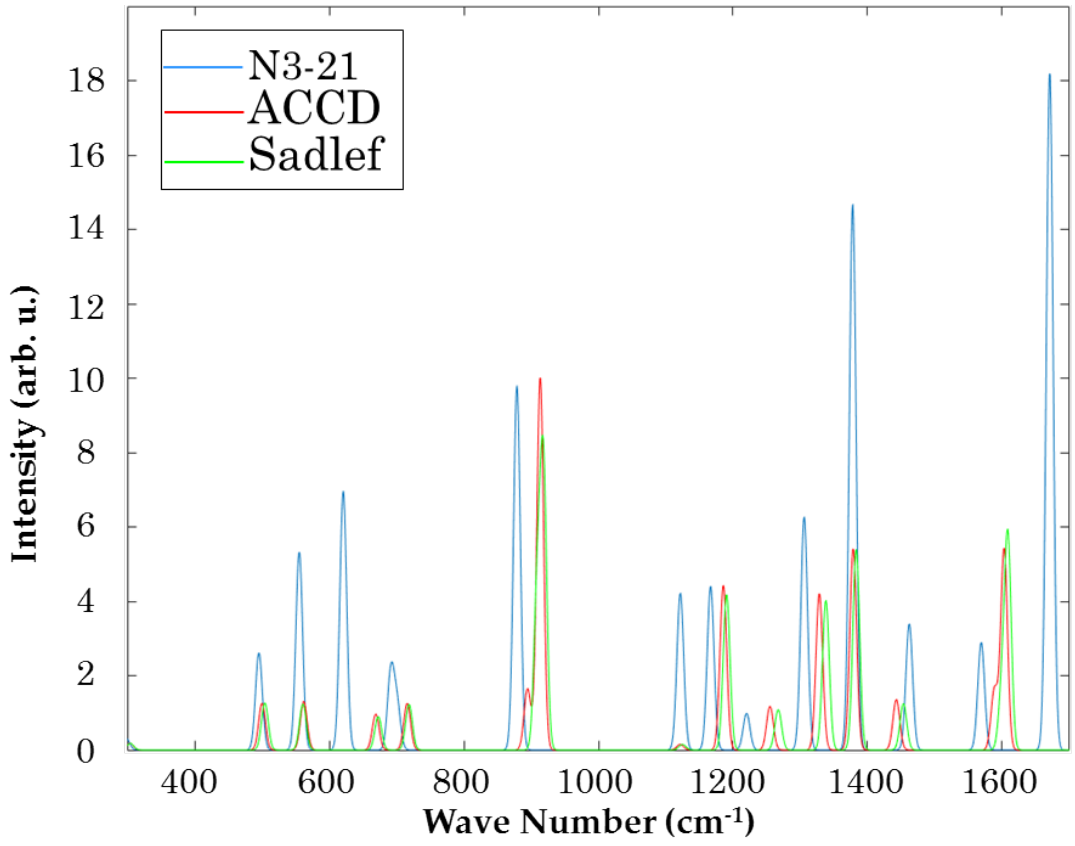


Figure 22: Simulated Raman Spectrum with ACCD basis set (top) and published Raman Spectrum (bottom) for Glycine

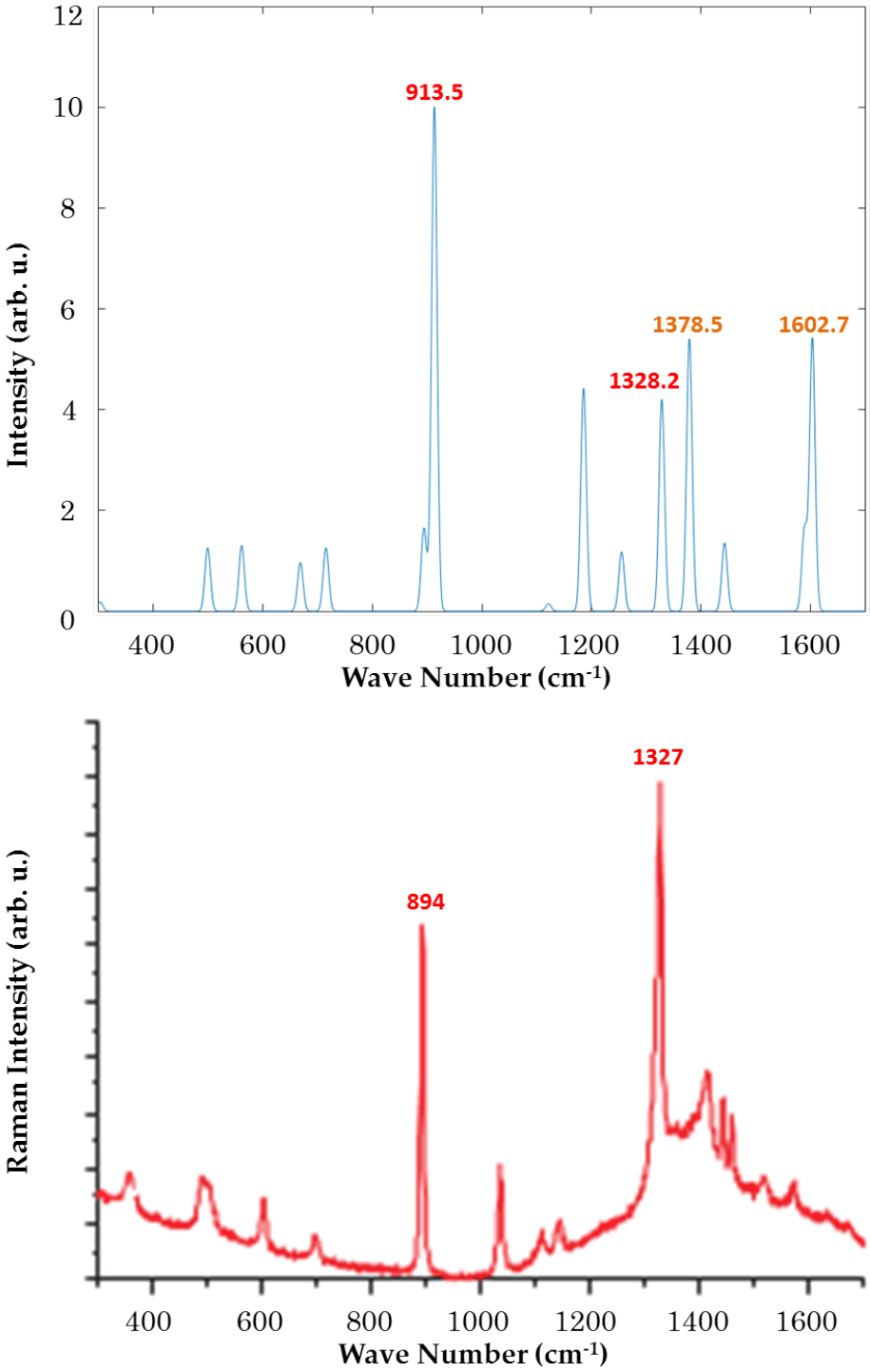


Figure 21 shows a further comparison between our simulated Raman spectrum using the ACCD basis set and the experimental Raman spectrum collected by Zhu, et al [30]. The Raman shift (wavenumber) that contains the majority of the Raman bands is known as the fingerprint region. The fingerprint region for glycine is from 700 to 1,700 cm^{-1}. In this region, the strong bands at 894 and 1,327 cm^{-1} can be considered glycine markers. One can clearly see these two markers in the experimental data. In our calculated results, these two intense peaks, located at 913.5 and 1,328.2 cm^{-1}, matched very well with the published data.

We found two extra modes from our simulation results. The peaks of these two extra modes are located at 1,378.5 and 1,602.7 cm^{-1}, which did not appear in the published results. The peak at 1,378.5 came from the NH_2 twist and CH_2 twist motion, and the peak at 1,602.7 cm^{-1} was from CH_2 bending. This might have occurred because consideration of the intermolecular molecule effect was neglected. Since these experimental results were taken from a sample of solid glycine formed by multiple glycine molecules, the bond formed between glycine molecule and its neighbor glycine molecule might smooth out the existences of these two modes.

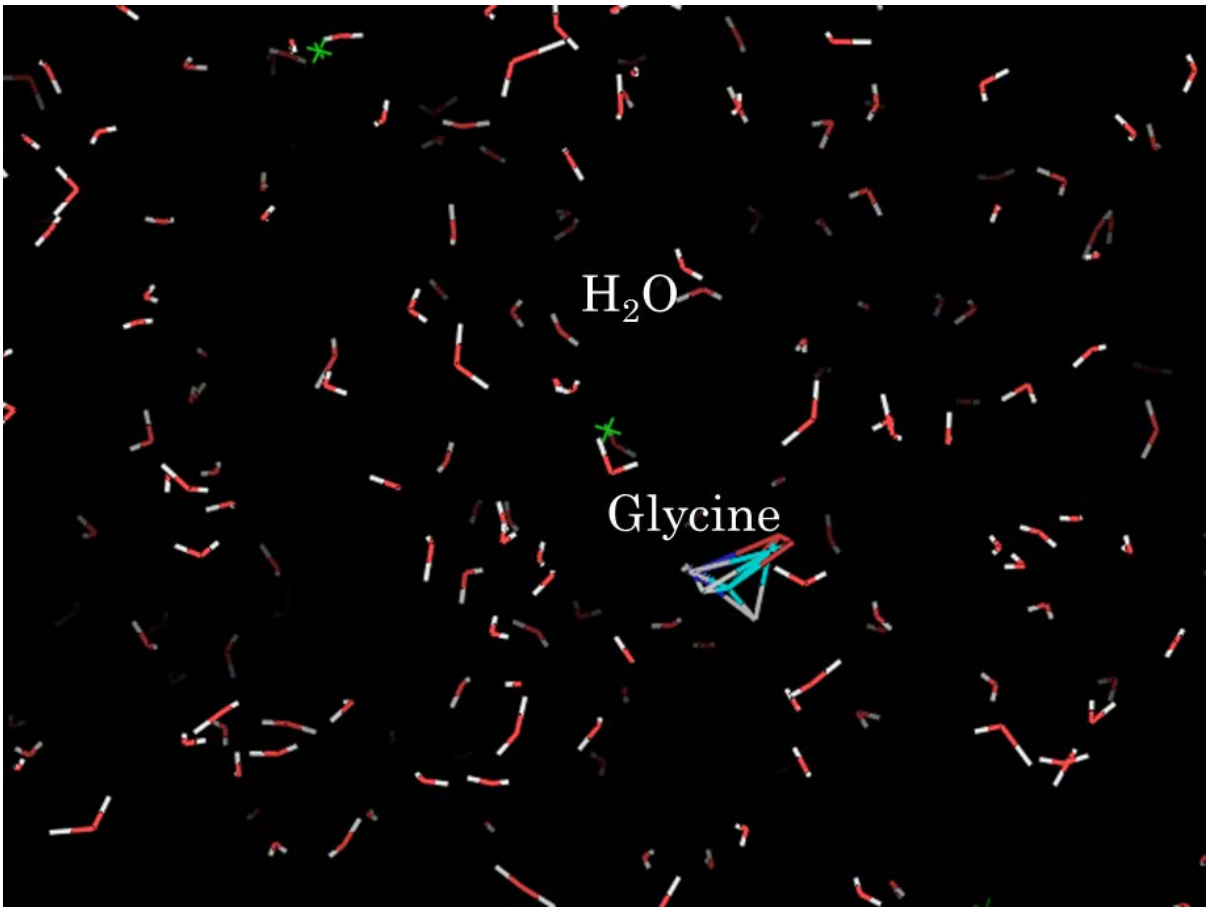
4.4 Result of Glycine Aqueous Solution Simulation

Biological organisms are usually affected by environmental factors such as incident photons, surrounding water molecules in solution, and neighbor chemical bonding. To characterize the responses of these organic matters to their environmental factors is challenging because high structure resolution and an ultra-fast time resolution often are required. Fortunately, computational simulations provide a cheaper way to understand these interactions. The solvent properties of amino acids are determined by the interaction with water molecules surrounding them, i.e. amino acids with nonpolar substituents are said to be hydrophobic (water-hating). Amino acids with polar R groups that form hydrogen bonds with water are classified as

hydrophilic (water-loving). In our simulation, we placed a glycine molecule in a water environment to test the solvent properties of glycine.

Figure 22, illustrates the simulation of a glycine aqueous solution system. The white-and-blue molecule at the center is the glycine molecule. The red-and-white molecules surrounding the glycine are water molecules. The green particle is the ion used to balance the electrons of the whole system. This simulation can perform the evolution of glycine interacting with water molecules in solution.

Figure 23: Illustration of Glycine aqueous solution system



After simulation, we analyzed the output data to extract useful information and illustrate the interaction process.

Figure 24: Plotting of RMSD (nm) of glycine solution respect with Time(ns)

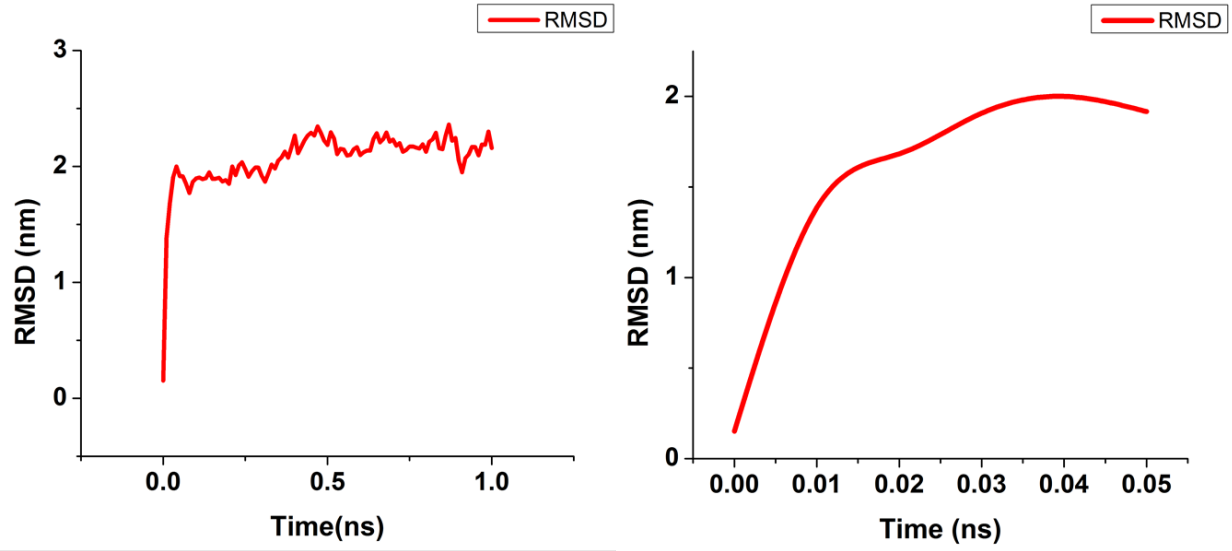


Figure 23 shows the plot of a root-mean-square deviation (RMSD) versus time for the glycine-water system. The graph on the left is the RMSD for the interaction within 1 ns, and the right plot is zoomed in on the time range of 0–0.05 ns. RMSD is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins.

From the plot in Figure 23, it was found that the RMSD of the system increased sharply in the first 50 ps and after that, it stayed in the vicinity of 2 nm. This means that the glycine water system reached an equilibrated state within 50 ps, after which the total energy for this system remained constant.

The analysis of the RMSD indicates that the glycine water system reached the equilibrated state in a very short time. Based on that analysis, we realize that the solvent properties of glycine are hydrophobic (water-hating), since the RMSD needed longer than 50 ps to form the hydrogen bond between water and the glycine molecule. Since the system reaches its minimum energy state in a short time, it means no interactions happen between biological organisms and their neighbor molecules

CHAPTER V: DISCUSSION

5.1 Results Discussion

In this work, we created simulated Raman spectra for methane (CH_4), water (H_2O), and glycine ($C_2H_5NO_2$) on our HPC cluster. It is found that simulated Raman spectrum results are in good agreement with experimental data and published results. Different basis sets, 3-21G, ACCD, and Sadlef, were tested to find their impacts on the simulation results and found the best basis sets with these small molecules.

The HF method we used to calculate the simulated Raman spectra introduce a frequency shift with our calculated results, because it treated the chemical bond vibration in the system as a harmonic oscillator. To correct this deficiency, we introduced a scale factor to take into account the anharmonicity effect and renormalized the calculated results. After applying the scale factor, we compared the scaled Raman spectra with experimental and published spectra and showed consistencies between the two results.

In the next step, we optimized the geometric structures of these small molecules by minimizing the total system energy and studied the effect of optimization on the Raman-simulated results.

For the H_2O molecule, we found that the hydrogen bond interaction among water molecules broadens the linewidth and shifts the frequency. In order to verify this effect, we simulated a Raman spectrum of a water dimer and compared with the simulated Raman spectrum of a single H_2O molecule.

For a basic amino acid (glycine), by comparing our simulated Raman spectrum and the published result, we found that the simulated results have a good agreement for wavenumbers below 1,310 cm^{-1}. However, at higher wavenumbers, we found two new high-frequency bending

modes that are not shown in the published results. Further study is needed to determine why these two peaks were missing from the previous experimental results.

In the last part, we simulated the evolution of glycine in a water solution using the GROMACS code. The trajectory of the glycine aqueous solution was investigated, and RMSD data were extracted from this simulation to prove the hydrophobic nature of the R group in the glycine molecule.

5.2 Further Improvements

In the next steps, several further improvements are needed to move forward our current calculations. First, a cross-check and benchmark test need to be performed using both GAMESS and GAUSSIAN code packages on both small molecules and larger macromolecules. Since for the Raman spectrum simulation, our result came from a single simulation software, a cross-check by GAUSSIAN is necessary. It is worthwhile to run a cross-check with the same targets to make sure of the accuracy of our results.

To study the water cluster effect on the Raman spectrum, we will include more water molecules in the simulation system instead of a water dimer. By including more water molecules, we can verify the inter-molecule effects in water.

Since some simulations did not fully run in parallel with MPI on our cluster, optimizing the performance of our HPC cluster is also our next goal.

Further, we will develop experiments to test simulated results with larger macromolecular biochemical molecules, such as lipids, amino acids, or DNA. These experiments can serve as references for us to move our simulation into complex materials.

5.3 Prospects

With the rapid growth of modern technology, the research on amino acids, proteins, lipids, and DNA will move more into the molecular level. After we finish the enhancements described above, we will continue our work on Raman spectroscopy to study DNA structure modifications under different environmental changes. An interaction simulation of DNA with small molecules and small molecule drugs will be performed, with the aim of creating new medicine for specific DNA genes. Raman spectroscopy has the advantages of fast analysis speed, low sample concentration, no pretreatment of sample, non-destructive sample, and high sensitivity etc. It has become a powerful tool in many fields of biology. Especially after the laser Raman spectroscopy, surface enhanced Raman spectroscopy, Fourier transform Raman spectroscopy were created, Raman spectroscopy has been developing rapidly in material structure analysis and material qualitative and quantitative analysis. Future research trends in the study of Raman spectroscopy include: the interaction of DNA with enzymes, drugs; the weak interaction between biological macromolecules; the study of molecular structures; the identification of enzymes and substrates, and the quality control based on protein so on. It is believed that Raman spectroscopy combined with other instruments will play a greater role in molecular biology, immunology, biochemistry, biophysics, food, pharmacy, environment and energy. In general, Raman spectroscopy can serve as a powerful tool for studying a variety of molecular structure and the mechanism of interaction between substances.

REFERENCES

1. Petroff, O. A. (2002). Book Review: GABA and Glutamate in the Human Brain. *The Neuroscientist* 8(6), 562-573.
2. Smekal, A. (1923). The quantum theory of dispersion. *Naturwissenschaften* 11, 873-875
3. Raman, C. V., & Krishnan, K. S. (1928). The Optical Analogue of the Compton Effect. *Nature*, 121(3053), 711-711.
4. Welsh, H. L., Crawford, M. F., Thomas, T. R., & Love, G. R. (1952). Raman Spectroscopy Of Low Pressure Gases And Vapors. *Canadian Journal of Physics*, 30(5), 577-96.
5. Porto, S. P., &Wood, D. L. (1962). Ruby Optical Maser as a Raman Source. *Applied Optics* 1(S1), 139.
6. Ditchfield, R., Hehre,W. J., & Pople, J. A. (1971). Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* 54(2), 724-28.
7. Dunning, T. H. (1989). Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics* 90(2), 1007-023.
8. Kendall, R. A., Dunning, T. H., & Harrison, R. J. (1992). Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of Chemical Physics* 96(9), 6796-806.
9. Sadlej, A. J. (1988). Medium-size polarized basis sets for high-level correlated calculations of molecular electric properties. *Collection of Czechoslovak Chemical Communications* 53(9), 1995-2016.
10. Schmidt, M.W. et al (1993). General Atomic and Molecular Electronic Structure System, *Journal of Computational Chemistry* 14(11), 1347-1363.
11. Gordon, M. S., & Schmidt, M. W. (2005). Advances in electronic structure theory. *Theory and Applications of Computational Chemistry*, 1167-189.
12. Summary of GAMESS' Capabilities, http://www.msg.ameslab.gov/gamess/capabilities.html
13. About Gaussian 16, (2017). http://gaussian.com/g16main/
14. About Gromacs, (2016). http://www.gromacs.org/About_Gromacs
15. Quinet, O., & Champagne, B. (2001). Time-dependent Hartree–Fock schemes for analytical evaluation of the Raman intensities. *The Journal of Chemical Physics* 115(14), 6293-299.
16. Quinet, O., & Champagne, B. (2002). Analytical time-dependent Hartree–Fock schemes for the evaluation of the hyper-Raman intensities. *The Journal of Chemical Physics* 117(6), 2481-488.
17. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.
18. Lee, S.Y., & Boo, B.H. (1996). Molecular Structures and Vibrational Spectra of Pyrrole and Carbazole by Density Functional Theory and Conventional ab Initio Calculations, *The Journal of Physical Chemistry* 100, 15073-15078
19. Lee, T. J., Martin, J. M., & Taylor, P.R. (1995). An accurate ab initio quartic force field and vibrational frequencies for CH4 and isotopomers. *The Journal of Chemical Physics* 102(1), 254-261.

20. Shimanouchi, T. (1972). Tables of molecular vibrational frequencies, consolidated volume 1.
21. Jourdanneau, E., Chaussard, F., Saint-Loup, R., Gabard, T., & Berger, H. (2005). The methane Raman spectrum from 1200 to 5500cm−1: A first step toward temperature diagnostic using methane as a probe molecule in combustion systems. *Journal of Molecular Spectroscopy* 233(2), 219-30.
22. Scott, A. P., & Radom, L. (1996). Harmonic Vibrational Frequencies: An Evaluation of Hartree−Fock, Møller−Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *The Journal of Physical Chemistry* 100(41), 16502-6513.
23. Huber, K. P., & Herzberg, G. (1979) Constants of diatomic molecules. *Molecular Spectra and Molecular Structure*, 8-689.
24. Hoy, A. R., & Bunker, P.R. (1979). A precise Solution of the Rotation Benning Schrodinger Equation for a Triatomic Molecule with Application to the Water Molecule *J. Molecular Spectroscopy* 74, 1-8.
25. Shimanouchi, T. (1972). Tables of molecular vibrational frequencies. Washington: National Bureau of Standards; for sale by the Supt. of Docs., *U.S. Govt. Print. Off*.
26. CareyGerald, D. M. & Korenowski, M. (1998). Measurement of the Raman Spectrum of liquid water, *The Journal of Physical Chemistry* 108, 2669.
27. Bosma, W. B., Fried, L. E. & Mukamel, S. (1993). Simulation of the intermolecular vibrational spectra of liquid water and water clusters, *The Journal of Chemical Physics* 98, 4413-4421.
28. Auer, B.M., & Skinner, J.L. (2008). IR and Raman spectra of liquid water: Theory and interpretation, *The Journal of Chemical Physics*128, 224511.
29. Kumar, S., Amareshwar, K. R. & Singh, V.B. et al (2005). Vibrational spectrum of glycine molecule, *Spectrochimica Acta Part A* 61, 2741-2746.
30. Zhu G., Zhu, X., Fan, Q., & Wan, X. (2011). Raman spectra of amino acids and their aqueous solutions, *Spectrochimica Acta Part A* 78 1187-1195.