Calculating Infrared Spectra

# of Proteins and Other Organic Molecules

Based on Normal Modes

by

Adam J. Mott

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved April 2012 by the
Graduate Supervisory Committee:

Peter Rez, Chair
S. Banu Ozkan
John Shumway
Michael Thorpe
Sara Vaiana

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

The goal of this theoretical study of infrared spectra was to ascertain to what degree molecules may be identified from their IR spectra and which spectral regions are best suited for this purpose. The frequencies considered range from the lowest frequency molecular vibrations in the far-IR, terahertz region (below $\sim 3\,\mathrm{THz}$ or $100\,\mathrm{cm}^{-1}$) up to the highest frequency vibrations ($\lesssim 120\,\mathrm{THz}$ or $4000\,\mathrm{cm}^{-1}$). An emphasis was placed on the IR spectra of chemical and biological threat molecules in the interest of detection and prevention. To calculate IR spectra, the technique of normal mode analysis was applied to organic molecules ranging in size from 8 to 11 352 atoms. The IR intensities of the vibrational modes were calculated in terms of the derivative of the molecular dipole moment with respect to each normal coordinate. Three sets of molecules were studied: the organophosphorus G- and V-type nerve agents and chemically related simulants (15 molecules ranging in size from 11 to 40 atoms); 21 other small molecules ranging in size from 8 to 24 atoms; and 13 proteins ranging in size from 304 to 11 352 atoms. Spectra for the first two sets of molecules were calculated using quantum chemistry software, the last two sets using force fields. The "middle" set used both methods, allowing for comparison between them and with experimental spectra from the NIST/EPA Gas-Phase Infrared Library. The calculated spectra of proteins, for which only force field calculations are practical, reproduced the experimentally observed amide I and II bands, but they were shifted by $\approx +40\,\mathrm{cm}^{-1}$ relative to experiment. Considering the entire spectrum of protein vibrations, the most promising frequency range for differentiating between proteins was ~ 600–$1300\,\mathrm{cm}^{-1}$ where water has low absorption and the proteins show some differences.

For my parents, Joseph and Beverly Mott

ACKNOWLEDGEMENTS

First and foremost, I owe a debt of gratitude to Peter Rez for extending so much of his time and effort on mentorship. He proposed this project, successfully applied for the funding, and offered help and guidance at every stage. The Defense Threat Reduction Agency (DTRA) funded this project, which paid my stipend as a research assistant for 2½ years and allowed me to attend two conferences to present my work. My father, Joseph Mott, has been a constant source of support in my life. He never pressured me into a particular career path and always supported me in my education. The value that my father and mother, Beverly, placed on education was influential in me attending college, and later, graduate school. My grandfather, Henry Hurst, encouraged an early interest in science—especially astronomy—which eventually led me to study physics. Since summer 2011, Fernanda Camacho Alanis has been a tremendous source of love and encouragement to me; she has lifted my spirits when I have felt discouraged and pushed me toward the finish line. At the beginning of this project, when I knew very little about computational chemistry, I relied heavily on the expertise of Justin Spiriti to help me set up calculations using CHARMM and Gaussian. Without Justin, the learning curve would have been much steeper, and I am grateful for his help. Saravana Prakash Thirumuruganandham worked heavily on molecular dynamics simulations for this project. Scott Menor helped me understand the format of the Hessian matrix output by CHARMM and taught me how to submit job scripts to ASU's Saguaro computer cluster. Jill Kolp and Araceli Vizcarra helped with administrative aspects such as my stipend, tuition and health insurance waivers, and travel reimbursements. I would like to thank the members of my supervisory committee for feedback and constructive criticism.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Due to the potential threat of terrorist attacks using chemical or biological weapons, there is considerable interest in reliable detection and identification of chemical and biological threat agents in order to prevent their production, distribution, and use. One important analytical tool that may be of use for these purposes is infrared (IR) spectroscopy. Since IR spectroscopy is sensitive to molecular vibrations, and a molecule's set of vibrational frequencies and associated IR intensities are unique to that molecule, a molecule's IR spectrum provides a "signature" or "fingerprint" that can aid identification.

The purpose of this dissertation is to evaluate to what extent IR spectra may be used to uniquely identify chemical and biological threat molecules varying in size from small molecules, such as the highly toxic organophosphorus nerve agents (18 to 42 atoms), up to large molecules, such as anthrax toxin protein ($\sim 10^4$ atoms). This is a theoretical work in that all original results (IR spectra) presented here were calculated based on various computational techniques. Wherever possible, I have attempted to make reference to experimental IR spectra in the published literature.

A molecule's emission or absorption of radiation in the IR region of the electromagnetic spectrum is primarily due to molecular vibrations, not overall rotations of the molecule or transitions between electronic states. As the atomic nuclei are displaced from their equilibrium positions by vibration of the molecule, the distribution of electric charge in the molecule changes accordingly. In classical electromagnetic theory, charges undergoing oscillations in position will emit radiation at the same frequency as the oscillation.

The main method for predicting molecular vibrations is normal mode analysis. This method considers vibrations involving small displacements of the atomic nuclei away from their equilibrium positions. In order to do normal mode analysis, one must have a way to calculate the total potential energy of the molecule as a function of the atomic displacements. More specifically, it is necessary to know all the mixed second (partial) derivatives of the potential energy with respect to the atomic displacements before one can proceed with this analysis.

For small molecules ($\lesssim$ 50 atoms) it is possible to directly calculate the total potential energy of the molecule (and hence also its derivatives) based on quantum theory using quantum chemistry software. Examples of software to perform these *ab initio* calculations include Gaussian 03 (Frisch *et al.*, 2003) and GAMESS (Schmidt *et al.*, 1993; Gordon and Schmidt, 2005). Both of these programs allow the user to calculate the total energy of a given molecule in a given geometric configuration and to optimize the geometric configuration to minimize the total energy for the purpose of determining the equilibrium positions of the atomic nuclei. Once these equilibrium positions are known, the software can calculate the second derivatives of the energy with respect to the atoms' displacements from their equilibrium positions. The molecule's normal modes of vibration are then found by diagonalizing the matrix of mass-weighted second derivatives. Finally, the IR intensity associated with a given normal mode of vibration may be calculated based on the change in the molecule's electric dipole moment when the atomic nuclei are displaced according to that normal mode.

For larger molecules such as proteins (*e.g.*, 6000 atoms), *ab initio* calculations are not practically possible with present-day computers because they would require too much time. In this case, it is necessary to switch to using approximate potential functions ("force fields") that have been developed for the purpose of performing molecular dynamics simulations. In this dissertation I have made use of the molecular dynamics software CHARMM (Brooks *et al.*, 2009) for its force field that is specific to proteins, and for its ability to calculate the needed second derivatives of the potential with respect to the atomic displacements.

Regardless of the method used to calculate the potential energy—*ab initio* or force field—once the normal mode frequencies and associated IR intensities are known, the IR spectrum may be simulated by giving the IR lines finite width through convolution with a Lorentzian peak profile (or any other chosen peak function).

A few words should be said on the units of frequency commonly used in IR spectroscopy. Since the frequency ν and wavelength λ of light in vacuum are related by $\lambda\nu = c$, where $c = 2.997\,924\,58 \times 10^{10}\,\text{cm/s}$ is the speed of light in vacuum, this can be rearranged as $\nu/c = 1/\lambda$. The "wave number" is defined as the inverse of the wavelength in cm, and often frequencies are reported in wave numbers (cm^{-1}) rather than Hz:

$$\nu[\text{cm}^{-1}] \equiv \frac{1}{\lambda[\text{cm}]} = \frac{\nu[\text{Hz}]}{2.997\,924\,58 \times 10^{10}} \ . \tag{1.1}$$

Since $1\,\text{THz} = 10^{12}\,\text{Hz}$, the previous equation may be rearranged to give a convenient conversion factor between frequencies in terahertz and cm^{-1},

$$\begin{aligned}\nu[\text{cm}^{-1}] &= \frac{100}{2.997\,924\,58} \times \nu[\text{THz}] \\ &\approx 33.356 \times \nu[\text{THz}] \ ,\end{aligned} \tag{1.2}$$

so $1\,\text{THz} \approx 33\,\text{cm}^{-1}$. Table 1.1 lists some frequencies and corresponding wavelengths relevant for IR and THz (far-infrared) spectroscopy. Terahertz spectroscopy usually considers frequencies in the range 0.03–3 THz (1–100 cm^{-1}). The highest frequency molecular vibrations, which involve displacements of hydrogen atoms, occur at frequencies of $\sim 4000\,\mathrm{cm}^{-1}$ (~ 120 THz).

Table 1.1: Conversions between frequencies in THz and cm^{-1} and corresponding wavelengths of light in vacuum in cm and µm in the range that is of interest for infrared and far-infrared spectroscopy.

ν[THz]	ν[cm^{-1}]	λ[cm]	λ[µm]
0.03	1	1	10^4
1	33	0.030	300
2	67	0.015	150
3	100	0.010	100
30	1000	0.001	10
60	2000	5×10^{-4}	5
90	3000	3.3×10^{-4}	3.3
120	4000	2.5×10^{-4}	2.5

The structure of this dissertation will be as follows: In Chapter 2, I will review the theory of small oscillations and normal mode analysis. In Chapter 3, I will apply normal mode analysis to the organophosphorus nerve agents and similar small molecules using quantum chemistry software to calculate IR spectra. In Chapter 4, I will attempt to bridge the gap between *ab initio* calculations and force field based calculations by comparing some small molecules' IR spectra derived from both methods to experimental spectra. In Chapter 5, I will use force fields as a basis for performing normal mode analysis and calculating IR spectra for a set of proteins.

Chapter 2

THEORY

Molecular vibrations are primarily responsible for the emission and absorption of radiation in the terahertz and infrared regions of the electromagnetic spectrum. Hence, the most straightforward approach to predicting the THz/IR spectrum of a molecule involves the calculation of the molecule's normal modes of vibration. The classical theory of small vibrations is described by Goldstein *et al.* (2002, chap. 6) and Wilson *et al.* (1955, chap. 2). Also, Krimm and Bandekar (1986) and Bahar *et al.* (2010) give excellent overviews in matrix form. In the classical view, a molecule of N atoms is treated as a collection of point masses. The kinetic energy of the molecule is then

$$T = \frac{1}{2}\sum_{i=1}^{N} m_i \left[\left(\frac{d\Delta x_i}{dt}\right)^2 + \left(\frac{d\Delta y_i}{dt}\right)^2 + \left(\frac{d\Delta z_i}{dt}\right)^2 \right], \tag{2.1}$$

where m_i is the mass of the ith atom and Δx_i, Δy_i, and Δz_i are the atomic displacements away from their equilibrium positions in Cartesian coordinates. The notation becomes simpler with the use of the mass-weighted displacements:

$$\begin{aligned} q_1 &= \sqrt{m_1}\,\Delta x_1, \\ q_2 &= \sqrt{m_1}\,\Delta y_1, \\ q_3 &= \sqrt{m_1}\,\Delta z_1, \\ q_4 &= \sqrt{m_2}\,\Delta x_2, \end{aligned} \tag{2.2}$$

and so on. Then the kinetic energy can be rewritten as

$$T = \frac{1}{2}\sum_{i=1}^{3N} \dot{q}_i^2 \ . \tag{2.3}$$

Expand the potential energy as a power series up to second order in the mass-weighted displacements:

$$V = V_0 + \sum_{i=1}^{3N} \left(\frac{\partial V}{\partial q_i} \right)_0 q_i + \frac{1}{2} \sum_{i=1}^{3N} \sum_{j=1}^{3N} \left(\frac{\partial^2 V}{\partial q_i \, \partial q_j} \right)_0 q_i \, q_j + \dots \qquad (2.4)$$

Here the "0" subscripts mean that the derivatives are to be evaluated with the molecule in its equilibrium configuration. This theory considers only small vibrations. By "small" vibrations it is meant that the atoms' displacements from their equilibrium positions are small enough that the potential energy may be adequately approximated to be a quadratic function of the atomic displacements. The higher order ("anharmonic") terms in the expansion of the potential are assumed to be vanishingly small and are therefore ignored.

Since the dynamics of the molecule are determined from the interatomic forces and these forces are calculated from the derivatives of the potential with respect to the atomic displacements, any constant value may be added to the potential energy without affecting the dynamics. Thus it is acceptable to set $V_0 = 0$. Furthermore, the equilibrium geometry of the molecule is its minimum-energy configuration; ideally, this is the molecular geometry for which the molecule's potential energy is at its *global* minimum. For the purposes of this discussion, it is only necessary that the potential energy is at a *local* minimum. At any local minimum of the potential energy, it must be true that $\partial V / \partial q_i = 0$ for $i = 1, 2, \dots, 3N$.

2.1 Hessian matrix

With these considerations, eq. 2.4 may be simplified to

$$V = \frac{1}{2} \sum_{i,j=1}^{3N} H_{ij} \, q_i \, q_j \ , \qquad (2.5)$$

where

$$H_{ij} = \left(\frac{\partial^2 V}{\partial q_i \, \partial q_j} \right)_0 . \tag{2.6}$$

In mathematics, a matrix of all possible combinations of second partial derivatives of a multivariable function is called the Hessian matrix (or simply the Hessian) of that function. In this dissertation, the function with which the Hessian is associated will always be a molecule's potential energy function. The Hessian is a square matrix, and it is symmetric ($H_{ji} = H_{ij}$) since the order in which one takes partial derivatives does not matter.

The matrix defined by eq. 2.6 has dimensions $3N \times 3N$ because there are $3N$ mass-weighted Cartesian displacement coordinates q_i. Common descriptors of this matrix include the mass-weighted Hessian matrix, mass-weighted second derivative matrix, and the mass-weighted force constant matrix. This is to specify that the derivatives that form the matrix elements are with respect to the mass-weighted Cartesian displacement coordinates. Some authors are even more specific by saying "*root*-mass-weighted" since the displacements in eq. 2.2 are weighted by the square roots of the atoms' masses. The specificity is to avoid confusion with the *non*-mass-weighted second derivative matrix, $(\partial^2 V / \partial \Delta_{i,k} \, \partial \Delta_{j,l})_0$, in which the derivatives are with respect to the original $3N$ Cartesian displacement coordinates Δx_1, Δy_1, Δz_1, Δx_2, *etc.* Here the subscripts i, j index the atoms from 1 to N while k, l refer to the Cartesian directions with $1 = x$, $2 = y$, and $3 = z$. This can be converted to the mass-weighted Hessian of eq. 2.6 by dividing by the square roots of the atomic masses:

$$H_{I,J} = \frac{1}{\sqrt{m_i \, m_j}} \left(\frac{\partial^2 V}{\partial \Delta_{i,k} \, \partial \Delta_{j,l}} \right)_0 , \tag{2.7}$$

where

$$\begin{aligned} I &= 3(i-1)+k \ , \\ J &= 3(j-1)+l \ . \end{aligned} \tag{2.8}$$

These matrix indices I, J assume that the $3N$ atomic displacements are ordered as $\Delta x_1, \Delta y_1, \Delta z_1,\ \Delta x_2, \Delta y_2, \Delta z_2,\ \ldots,\ \Delta x_N, \Delta y_N, \Delta z_N$ (the same order as in eq. 2.2).

2.2 Coupled equations of motion

Using eqs. 2.3 and 2.5 for the kinetic and potential energies of the molecule, one can write down the equations of motion using the Lagrangian formulation of mechanics. For this, one needs to compute the generalized forces $(-\partial V/\partial q_i)$. It's worth noting that there is a bit of subtlety involved in computing these derivatives because the sum in eq. 2.5 involves terms that are either explicitly linear $(H_{ij}\, q_i\, q_j)$ or quadratic $(H_{ij}\, q_i^2)$ in a given generalized coordinate q_i, and the cases need to be handled separately. The sum in eq. 2.5 may be rewritten as two sums—one for terms involving off-diagonal elements of $\mathbf{H}$, and another for terms involving diagonal elements:

$$V = \frac{1}{2}\left[\sum_{i\neq j} H_{ij}\, q_i\, q_j + \sum_{i=j} H_{ij}\, q_i^2\right] . \tag{2.9}$$

The Lagrangian is

$$L = T - V = \frac{1}{2}\sum_{i=1}^{3N} \dot{q}_i^2 - \frac{1}{2}\left[\sum_{i\neq j} H_{ij}\, q_i\, q_j + \sum_{i=j} H_{ij}\, q_i^2\right] . \tag{2.10}$$

The Lagrangian equation of motion for a given coordinate q_k is

$$\frac{\partial L}{\partial q_k} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) . \tag{2.11}$$

Inputting the expression for the Lagrangian and simplifying results in

$$-\sum_{j=1}^{3N} H_{kj}\, q_j = \ddot{q}_k \tag{2.12}$$

as the equation of motion associated with coordinate q_k. There are $3N$ such equations of motion (coupled second order linear differential equations) that must be solved simultaneously to find the trajectories $q_k(t)$. For a normal mode of vibration, all $3N$ coordinates simultaneously undergo harmonic oscillations at the mode frequency ν (angular frequency $\omega = 2\pi\nu$), so one expects a solution of the form

$$q_k = a_k\, e^{-i\omega t}, \tag{2.13}$$

where it is understood that q_k is the *real part* of the complex number. The complex coefficients a_k determine both the amplitudes and phases of the oscillations. Substituting this expression back into eq. 2.12 gives

$$\sum_{j=1}^{3N} H_{kj}\, a_j = \omega^2 a_k\ . \tag{2.14}$$

There are $3N$ such equations, one for each coefficient a_k. Together, they comprise a matrix equation

$$\mathbf{H}\,\mathbf{a} = \omega^2\,\mathbf{a}\ , \tag{2.15}$$

in which $\mathbf{a}$ is the column vector of coefficients $a_1, a_2, \ldots, a_{3N}$. For a square matrix $\mathbf{M}$, λ is said to be an eigenvalue and $\mathbf{v}$ an eigenvector of the matrix if they satisfy the equation $\mathbf{M}\,\mathbf{v} = \lambda\,\mathbf{v}$. Hence, the square of the angular frequency, ω^2, in eq. 2.15 is an eigenvalue of the mass-weighted Hessian matrix, and the vector of coefficients, $\mathbf{a}$, is the corresponding eigenvector. The eigenvector specifies each atom's amplitude and direction of oscillation when the molecule is vibrating in that specific normal mode at that normal mode's frequency. The angular frequency of the normal mode is easily obtained from the square root of its eigenvalue, $\omega = \sqrt{\omega^2}$, from which the frequency of vibration in Hz is $\nu = \omega/2\pi$.

The matrix $\mathbf{H}$ will have a total of $3N$ eigenvalues ω^2 and corresponding eigenvectors $\mathbf{a}$, but not all of these solutions will correspond to vibrational modes. Even among the set of eigenvalues and eigenvectors corresponding to vibrational modes, some of the mode frequencies may not be unique.

2.3 Zero eigenvalues

In general, this approach will yield several eigenvectors corresponding to an eigenvalue of $\omega^2 = 0$. These eigenvectors do not describe vibrational modes; rather, they describe constant velocity (zero frequency) translations or constant angular velocity rotations of the molecule as a whole. The molecule's center of mass may move at constant velocity in three independent directions (for example, in the x, y, or z directions), and the molecule may rotate around three different axes (for example, its three principal axes of rotation). The exception is a linear molecule such as CO_2 for which it doesn't make sense to speak of a rotation about the axis running through the collinear atoms; there are only two independent rotations in this case. Thus a non-linear molecule will have six zero-frequency modes (three translations and three rotations), whereas a linear molecule will have five (three translations and two rotations). Therefore, a non-linear molecule of N atoms will have $3N - 6$ normal modes of vibration; a linear molecule will have $3N - 5$. That is not to say that the vibrational frequencies of these modes will all necessarily be unique. Depending on the symmetry of the molecule there may be repeated ("degenerate") values of ω^2 that satisfy eq. 2.15, in which case there will be multiple eigenvectors (modes of vibration) associated with a single vibrational frequency.

Alternatively, it is possible to formulate the problem in such a way that the zero-frequency translation and rotation solutions do not arise. Since there are really only $3N - 6$ vibrational degrees of freedom (for a non-linear molecule), the set of $3N$ Cartesian coordinates contains six extra degrees of freedom beyond what is necessary for the vibration-only problem. The extra degrees of freedom may be eliminated by using "internal" coordinates in place of Cartesians, in which case the atoms are located relative to other atoms in the molecule by bond lengths, bond angles, and dihedral angles. In computational chemistry, the data file that describes the locations of all the atoms in a molecule in this way is called a Z-matrix. For example, the configuration of a water molecule can be specified using only three internal coordinates: the lengths of the two oxygen-hydrogen bonds and the angle between those bonds. Notice that such a coordinate system is oblivious to overall translation or rotation of the molecule. In general, any minimal set of internal coordinates that locates every atom in the molecule relative to another atom will involve $3N - 6$ coordinates. When using such a coordinate system, there will be no zero-frequency solutions to the eigenvalue problem that correspond to overall translations or rotations of the molecule. However, in this case the expression for the molecule's kinetic energy will not only have diagonal terms (containing $\dot{q}_i^2$) as in eq. 2.3, but also off-diagonal terms (containing $\dot{q}_i \, \dot{q}_j$). This results in an eigenvalue equation that is somewhat different from that shown in eq. 2.15. Goldstein *et al.* (2002, chap. 6) give a more general treatment of the problem of vibrations in any set of coordinates.

2.4 Negative eigenvalues

When one uses a computer program to determine the minimum-energy equilibrium geometry of a molecule, the mass-weighted Hessian matrix in terms of Cartesian atomic displacements from equilibrium, and the eigenvalues and eigenvectors of this matrix, the resulting six lowest eigenvalues will likely be *close to* but not exactly zero. This is due to the computer's finite precision. But if the six lowest frequency eigenvalues differ significantly from zero, this is an indication that the geometry of the molecule needs to be further optimized in order to obtain a minimum-energy, equilibrium configuration prior to calculating the Hessian matrix. Often if the molecular geometry has not been properly optimized, the first six eigenvalues will come out as large negative numbers. Remember that the eigenvalues are the squares of angular frequencies, so negative eigenvalues are nonsensical. This theory breaks down if the molecule is not in a minimum-energy geometry because in that case $\partial V/\partial q_i \neq 0$ in eq. 2.4 and thus eq. 2.5 will not accurately describe the potential energy. Furthermore, the small oscillations of the atoms' positions should be centered on their equilibrium, minimum potential energy positions, so it doesn't make sense to perform this normal mode analysis with the molecule in any other configuration besides its equilibrium one.

2.5 Properties of the eigenvalues and eigenvectors

It can be shown that the eigenvalues of any real, symmetric matrix (like $\mathbf{H}$) must be real numbers. Furthermore, as a result of the fact that the potential energy as defined in eq. 2.5 will always be greater than or equal to zero, it can be shown that the eigenvalues of $\mathbf{H}$ must be greater than or equal to zero.

The elements of any given eigenvector $\mathbf{a}_i$ will all have the same complex phase, meaning that all the atoms in the molecule will pass through their equilibrium positions at the same time, pass through their maximum displacements at the same time, and so on. Since the elements of an eigenvector have the same complex phase, that arbitrary phase can be chosen such that all the elements are real numbers. Hence all the eigenvectors can be chosen to be real. A complicated motion of the molecule involving more than one simultaneous normal mode vibration may be described as a linear combination of real eigenvectors with the appropriately chosen complex amplitudes to allow for the possibility that the different normal modes have different relative phases.

Furthermore, since eq. 2.15 can be multiplied on either side by any constant value, it is seen that there is an overall indeterminacy in the normalization of an eigenvector $\mathbf{a}_i$. That is, if $\mathbf{a}_i$ is a valid eigenvector, then so are $2\mathbf{a}_i$, $-5\mathbf{a}_i$, and so on. To remedy this indeterminacy, the eigenvectors are usually normalized to have $\mathbf{a}_i^T\,\mathbf{a}_i = 1$. Even then there is an indeterminacy by overall sign, since the vector $-\mathbf{a}_i$ will satisfy the normalization condition just as well as $\mathbf{a}_i$; to get around the sign indeterminacy, one may choose the first element of each eigenvector to have positive sign.

Also, the eigenvectors have the important property that they are mutually orthogonal. That is, for two different eigenvectors $\mathbf{a}_i$ and $\mathbf{a}_j$, it is always true that $\mathbf{a}_i^T\,\mathbf{a}_j = 0$. In the case that an eigenvalue is degenerate (having more than one eigenvector associated with it), orthogonal eigenvectors may be constructed through a Gramm-Schmidt orthogonalization process.

If one forms a matrix $\mathbf{A}$ whose columns are the normalized eigenvectors, then $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, the identity matrix (*i.e.*, $\mathbf{A}^T = \mathbf{A}^{-1}$). The orthonormal matrix $\mathbf{A}$ is said to diagonalize the matrix $\mathbf{H}$. That is, when one forms the matrix product $\mathbf{A}^T\mathbf{H}\mathbf{A}$, the resulting matrix has nonzero elements only along its diagonal, and those diagonal elements are the eigenvalues of $\mathbf{H}$. Furthermore, the normal mode eigenvectors may be used to define a new set of coordinates called the normal coordinates, $Q_1, Q_2, \ldots, Q_{3N}$. In terms of the original set of coordinates $q_1, q_2, \ldots, q_{3N}$ from eq. 2.2, the normal coordinates are obtained from

$$\mathbf{Q} = \mathbf{A}^T\mathbf{q} \,, \tag{2.16}$$

where $\mathbf{Q}$ and $\mathbf{q}$ are column vectors containing the $3N$ coordinates from each of the two coordinate sets. The reverse transformation (to convert from the Q_i's to the q_i's) is

$$\mathbf{q} = \mathbf{A}\mathbf{Q} \,. \tag{2.17}$$

2.6 Absorption of Electromagnetic Radiation

Imagine an electromagnetic (EM) wave passing through a molecule. An individual photon of the EM radiation can either be completely absorbed by the molecule or scattered off of it. In the scattering case, if the photons scatter inelastically off the molecule (Raman scattering), the resulting frequency distribution of the scattered photons is called a Raman spectrum. A scattered photon can have either lower or higher energy than the incident photon depending on whether the molecule jumped to a higher or lower energy state than it started in.

This is in contrast to the absorption case. When a photon is absorbed by the molecule, the molecule jumps to a higher energy state. Hence the EM wave will lose energy as it travels through a collection of molecules because some of its energy will be transferred to the molecules. The change in quantum state of the molecules could be a jump to a higher energy electronic state.

Or it could be to a higher energy state of translation, rotation, or vibration. For EM radiation in the IR and THz range, the absorption of photons is mainly due to exciting the vibrational states of the molecules. Predicting the frequency dependence of the absorption—that is, the IR/THz spectrum—requires knowledge of the vibrational motions of the molecules (either from normal mode analysis or from molecular dynamics trajectories) as well as the distribution of charge within the molecules. The charge distribution is crucial because it is through the charges that the oscillating electric field of the EM wave exerts forces on parts of the molecule, driving the molecule to oscillate at the same frequency as the EM radiation.

The vibrational response of a molecule to EM radiation is best illustrated by a classical treatment of a single bound charge. This is described by Griffiths (1999, section 9.4.3) and Jackson (1999, section 7.5). A point particle having charge e and mass m is driven to oscillate by a plane EM wave having frequency $\omega/2\pi$. At the location of the particle, the electric field of the EM wave is given by the real part of $\mathbf{E}(t) = \hat{\mathbf{x}}\, E_0\, e^{-i\omega t}$. The particle is bound to the origin ($x = 0$) by a harmonic potential, $V(x) = \frac{1}{2}m\omega_0 x^2$, so that its natural frequency of vibration is ω_0. The particle's motion is damped by a force that is proportional to its velocity, $F_{\text{damping}} = -m\gamma\dot{x}$, with damping constant γ. The motion of the particle in this driven, damped harmonic oscillator system is given by the real part of $x(t) = x_0\, e^{-i\omega t}$ with the complex factor

$$x_0 = \frac{1}{\omega_0^2 - \omega^2 - i\gamma\omega}\frac{eE_0}{m} \tag{2.18}$$

specifying both the amplitude and the phase of the particle's oscillation. The oscillating charge results in an oscillating electric dipole moment that is the real part of $\mathbf{p}(t) = e\,x(t)\,\hat{\mathbf{x}}$. Comparing this to

$$\mathbf{p}(t) = \epsilon_0 \gamma_p \mathbf{E}(t) \tag{2.19}$$

defines an expression for the complex polarizability γ_p (which ends up having dimensions of volume and is not to be confused with the damping constant γ). If instead of a single point charge bound harmonically to a site, there are many such oscillators distributed throughout a volume, then the complex polarization vector $\mathbf{P}$ (the real part of which is the net dipole moment per unit volume) for this medium is the sum of all the individual dipoles $\mathbf{p}(t)$ divided by the volume. This expression will look like

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}\ , \tag{2.20}$$

which defines an expression for the complex electric susceptibility χ_e. Then the complex dielectric constant is

$$\epsilon/\epsilon_0 = 1 + \chi_e\ . \tag{2.21}$$

Griffiths uses this in a wave equation for the electric field in this medium,

$$\nabla^2 \mathbf{E} = \epsilon \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}\ , \tag{2.22}$$

with a trial solution

$$\mathbf{E}(z,t) = \hat{\mathbf{x}}\, E_0 \exp\left[i\left(kz - \omega t\right)\right] \tag{2.23}$$

to derive a complex wave number

$$k = \sqrt{\epsilon\mu_0}\,\omega = \frac{\omega}{c}\sqrt{1+\chi_e}\ . \tag{2.24}$$

The real part of k (call it k_r) leads to the usual traveling wave, but the imaginary part (k_i) results in an exponential attenuation of the amplitude:

$$\mathbf{E}(z,t) = \hat{\mathbf{x}}\, E_0\, e^{-k_i z} e^{i(k_r z - \omega t)}\ . \tag{2.25}$$

The power per unit area carried by the EM wave as it propagates in the $+z$ direction is given by the magnitude of the Poynting vector, which is proportional to E^2. After traveling a distance z, the wave's intensity has been attenuated by the factor $e^{-2k_i z}$ due to transfer of energy to the oscillating charges. Comparing this to the Beer–Lambert law,

$$I(z) = I_0 \, e^{-\alpha z} \ , \tag{2.26}$$

the absorption coefficient α, which is a function of the frequency ω of the EM radiation, is identified as being

$$\alpha(\omega) = 2k_i(\omega) \ . \tag{2.27}$$

For the oscillators spread throughout a volume V and having possibly different charges e_i, masses m_i, natural frequencies ω_i (different spring constants), and damping constants γ_i, the absorption coefficient is equal to

$$\alpha(\omega) \approx \frac{\omega^2}{V\epsilon_0 c} \sum_i \frac{e_i^2 \gamma_i}{m_i \left[\left(\omega_i^2 - \omega^2\right)^2 + \gamma_i^2 \omega^2 \right]} \ . \tag{2.28}$$

The approximation used by Griffiths to derive this expression is

$$\sqrt{\epsilon/\epsilon_0} = \sqrt{1+\chi_e} \approx 1 + \frac{1}{2}\chi_e \text{ for } |\chi_e| \ll 1 \tag{2.29}$$

in eq. 2.24, which is to say that the molecule is considered to be in the gas phase so that the dielectric constant ϵ/ϵ_0 is close to 1, its value in a vacuum. Alternatively, an exact relationship between the real and imaginary parts of ϵ/ϵ_0 and k can be derived from writing $k = k_r + ik_i$, squaring to get $k^2 = k_r^2 - k_i^2 + 2ik_r k_i$, and comparing this with eq. 2.24, which says that $k^2 = \mu_0 \epsilon \omega^2 = (\omega^2/c^2)(\epsilon/\epsilon_0)$. This results in two equations in two unknowns that allow one to solve for k_r and k_i in terms of $\operatorname{Re}\epsilon/\epsilon_0$ and $\operatorname{Im}\epsilon/\epsilon_0$:

$$\begin{aligned} k_r^2 - k_i^2 &= \frac{\omega^2}{c^2} \operatorname{Re}\epsilon/\epsilon_0 \\ 2k_r k_i &= \frac{\omega^2}{c^2} \operatorname{Im}\epsilon/\epsilon_0 \end{aligned} \tag{2.30}$$

Once k_i has been solved for, the absorption coefficient $\alpha(\omega)$ can be calculated using 2.27.

Extending this analysis from isolated point charges to a distribution of charges in a molecule, it will no longer be generally true that the induced oscillating dipole moment is aligned with the direction of the electric field oscillation that is causing it. In this case eqs. 2.19 and 2.20 become matrix equations with the complex polarizability γ_p and susceptibility χ_e now being 3×3 matrices:

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} \tag{2.31}$$

$$\begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} = \begin{pmatrix} \chi_{11} & \chi_{12} & \chi_{13} \\ \chi_{21} & \chi_{22} & \chi_{23} \\ \chi_{31} & \chi_{32} & \chi_{33} \end{pmatrix} \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix} \tag{2.32}$$

The subscripts refer to the Cartesian directions with $1 = x$, $2 = y$, and $3 = z$. E_1, E_2, and E_3 are the x, y, and z components of the amplitude $\mathbf{E}_0$ in $\mathbf{E}(t) = \mathbf{E}_0\, e^{-i\omega t}$. The difference between the dipole moment $\mathbf{p}$ and the polarization $\mathbf{P}$ is that $\mathbf{p}$ is a microscopic quantity (the induced dipole of a single molecule) whereas $\mathbf{P}$ is a macroscopic quantity (the net dipole moment per unit volume of a material). In a region of volume V, the two are related by

$$\mathbf{P} = \frac{1}{V} \sum_{i=1}^{N_{\text{mol}}} \mathbf{p}_i = \frac{N_{\text{mol}}}{V} \langle \mathbf{p} \rangle \ . \tag{2.33}$$

where N_{mol} is the number of molecules in the volume and $\langle \mathbf{p} \rangle$ is the average dipole moment of a single molecule. For simplicity, consider the volume to be that of a single molecule (for a large biological macromolecule like a protein, it's not so ridiculous to speak of the volume of a single molecule). In that case, $\mathbf{P} = \mathbf{p}/V$. The effect of different orientations of the molecule will be handled later.

The elements of the susceptibility tensor in eq. 2.32 may be written in terms of the molecule's normal modes of vibration:

$$\chi_{k,l} = \frac{1}{V\epsilon_0} \sum_{n=7}^{3N} \frac{1}{\omega_n^2 - \omega^2 - i\gamma_n\omega} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{e_i \, A_{I,n} \, A_{J,n} \, e_j}{m_i^{1/2} \, m_j^{1/2}} \ . \tag{2.34}$$

The subscripts i, j, k, l, I, J have the same meanings as in eq. 2.8: i, j index the atoms from 1 to N; k, l specify one of the three Cartesian directions; and I, J combine the atom number and direction into a single index. The subscript n indexes the $3N - 6$ normal modes of vibration from 7 to $3N$ in the order of increasing frequency, omitting the first six modes because they are assumed to be zero-frequency modes corresponding to overall translations or rotations of the molecule. The mass of atom i is m_i, and e_i is its effective partial charge. $\mathbf{A}$ is the matrix introduced in the previous section whose columns are the normal mode eigenvectors. $A_{I,n}$ is the element of the nth normal mode eigenvector corresponding to atom i and direction k.

A single susceptibility value χ_e that is an average over the possible orientations of the molecule may be calculated from the trace of the susceptibility tensor:

$$\chi_e = \frac{1}{3} \sum_{k=l=1}^{3} \chi_{k,l} = \frac{1}{3} \left(\chi_{11} + \chi_{22} + \chi_{33} \right) \ . \tag{2.35}$$

Using this χ_e with the approximation of eq. 2.29 in eq. 2.24 to calculate the complex wavenumber k, taking its imaginary part, and multiplying by two (eq. 2.27) gives the absorption coefficient

$$\alpha(\omega) \approx \frac{1}{3V\epsilon_0 c} \sum_{n=7}^{3N} \frac{\gamma_n \omega^2}{\left(\omega_n^2 - \omega^2\right)^2 + \gamma_n^2 \omega^2} \sum_{k=l=1}^{3} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{e_i\, A_{I,n}\, A_{J,n}\, e_j}{m_i^{1/2}\, m_j^{1/2}} \ . \qquad (2.36)$$

Whenever I report a calculated IR/THz spectrum in this dissertation, I am really referring to the frequency-dependent absorption coefficient. Ignoring constant pre-factors, this expression can be summarized as

$$\alpha(\omega) \propto \sum_{n=7}^{3N} S_n(\omega) \times I_n \qquad (2.37)$$

where the line strength I_n gives the intensity (maximum depth) of the absorption line due to normal mode n:

$$I_n = \sum_{k=l=1}^{3} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{e_i\, A_{I,n}\, A_{J,n}\, e_j}{m_i^{1/2}\, m_j^{1/2}} \ . \qquad (2.38)$$

It will be shown later (eq. 2.52) that there is an equivalent but computationally more efficient way of calculating this quantity.

The line-shape function $S_n(\omega)$ only depends on the frequency ω of the EM wave, the frequency ω_n of the nth normal mode, and the damping constant γ_n associated with that mode:

$$S_n(\omega) = \frac{\gamma_n \omega^2}{\left(\omega_n^2 - \omega^2\right)^2 + \gamma_n^2 \omega^2} \ . \qquad (2.39)$$

This function describes the frequency dependence of the resonant absorption of EM radiation at frequencies near one of the natural resonant frequencies (normal modes) of the molecule. The resonant absorption is strongest when the frequency ω of the radiation matches one of the normal mode frequencies ω_n of the molecule. Thus the line-shape function is peaked at $\omega = \omega_n$, with the height of the peak being $S_n(\omega_n) = 1/\gamma_n$. Note that the line-shape function has the desirable property that the height of the peak does not depend on the frequency. Thus the same line shape can be used for peaks at various frequencies, with their heights being set by multiplication by the I_n in eq. 2.37. The width of the peak—that is, the range of radiation frequencies that can be absorbed by mode n—is proportional to the damping constant γ_n. The peak's area is independent of γ_n since its width is proportional to γ_n while its height is $1/\gamma_n$.

2.6.1 Lorentzian line shape

It can be shown that the absorption line shape of eq. 2.39 is approximately a Lorentzian function. The denominator is equal to

$$\omega_n^4 - 2\omega_n^2\omega^2 + \omega^4 + \gamma_n^2\omega^2 \ . \tag{2.40}$$

If the ω^2 in the numerator is then written as ω^{-2} in the denominator, then the denominator becomes

$$\omega_n^4\omega^{-2} - 2\omega_n^2 + \omega^2 + \gamma_n^2 \ . \tag{2.41}$$

Let $\omega = \omega_n + \delta = \omega_n(1 + \delta/\omega_n)$. Then $\omega^2 = \omega_n^2 + 2\omega_n\delta + \delta^2$. Using a Taylor expansion up to second order in δ/ω_n,

$$\omega^{-2} = \omega_n^{-2}\left(1 + \frac{\delta}{\omega_n}\right)^{-2} \approx \omega_n^{-2}\left[1 - 2\frac{\delta}{\omega_n} + 3\left(\frac{\delta}{\omega_n}\right)^2\right] . \tag{2.42}$$

Substituting these expressions for ω^2 and ω^{-2}, the denominator becomes

$$\omega_n^2\left[1 - 2\frac{\delta}{\omega_n} + 3\left(\frac{\delta}{\omega_n}\right)^2\right] - 2\omega_n^2 + \left(\omega_n^2 + 2\omega_n\delta + \delta^2\right) + \gamma_n^2 \ , \tag{2.43}$$

which simplifies to $4\delta^2 + \gamma_n^2$. Recalling that $\delta = \omega - \omega_n$, the line-shape function becomes

$$S_n(\omega) \approx \frac{\gamma_n}{4\left(\omega - \omega_n\right)^2 + \gamma_n^2} = \frac{1}{2}\frac{\frac{1}{2}\gamma_n}{\left(\omega - \omega_n\right)^2 + \left(\frac{1}{2}\gamma_n\right)^2} \ . \tag{2.44}$$

Up to a normalization factor, this is identical to the Lorentzian function

$$L_n(\omega) = \frac{1}{\pi} \frac{\frac{1}{2}\Gamma}{(\omega - \omega_n)^2 + \left(\frac{1}{2}\Gamma\right)^2} \;, \tag{2.45}$$

which has been normalized so that its integral over all frequencies is 1. As can be seen from setting $\omega - \omega_n = \pm\frac{1}{2}\Gamma$, the full width of the Lorentzian function at half its maximum value (FWHM) is Γ. Since $S_n(\omega)$ is approximately a Lorentzian function, the damping constant $\gamma_n \approx \Gamma$, the FWHM of the peak. This proves the earlier statement that the width of the peak is proportional to γ_n.

Fig. 2.1 compares the line shape $S_n(\omega)$ to a Lorentzian function $L_n(\omega)$. Both peaks are centered at $\omega_n = 500\,\mathrm{cm}^{-1}$ and have the same peak width, $\gamma_n = \Gamma = 100\,\mathrm{cm}^{-1}$. The Lorentzian is a good approximation to $S_n(\omega)$ and near ω_n the two functions are practically indistinguishable, although noticeable differences appear in the tails of the peaks. The Lorentzian function is exactly symmetric about $\omega = \omega_n$, as eq. 2.45 is insensitive to a change of sign of the quantity $\omega - \omega_n$. Also, the Lorentzian function doesn't go to zero at zero frequency. This is in contrast to $S_n(\omega)$, which is slightly asymmetric and goes to zero in the limit of zero frequency. In practice, either of these two line shapes can be used in eq. 2.37 without making much difference in the final convolved spectrum $\alpha(\omega)$.

An important question is what should the damping constants γ_n be for the various normal modes? Currently, I don't have a good way of predicting the γ_n from theory. For the IR/THz spectra calculated in this dissertation, I have set γ_n equal to the same constant line width (FWHM) Γ for all modes.

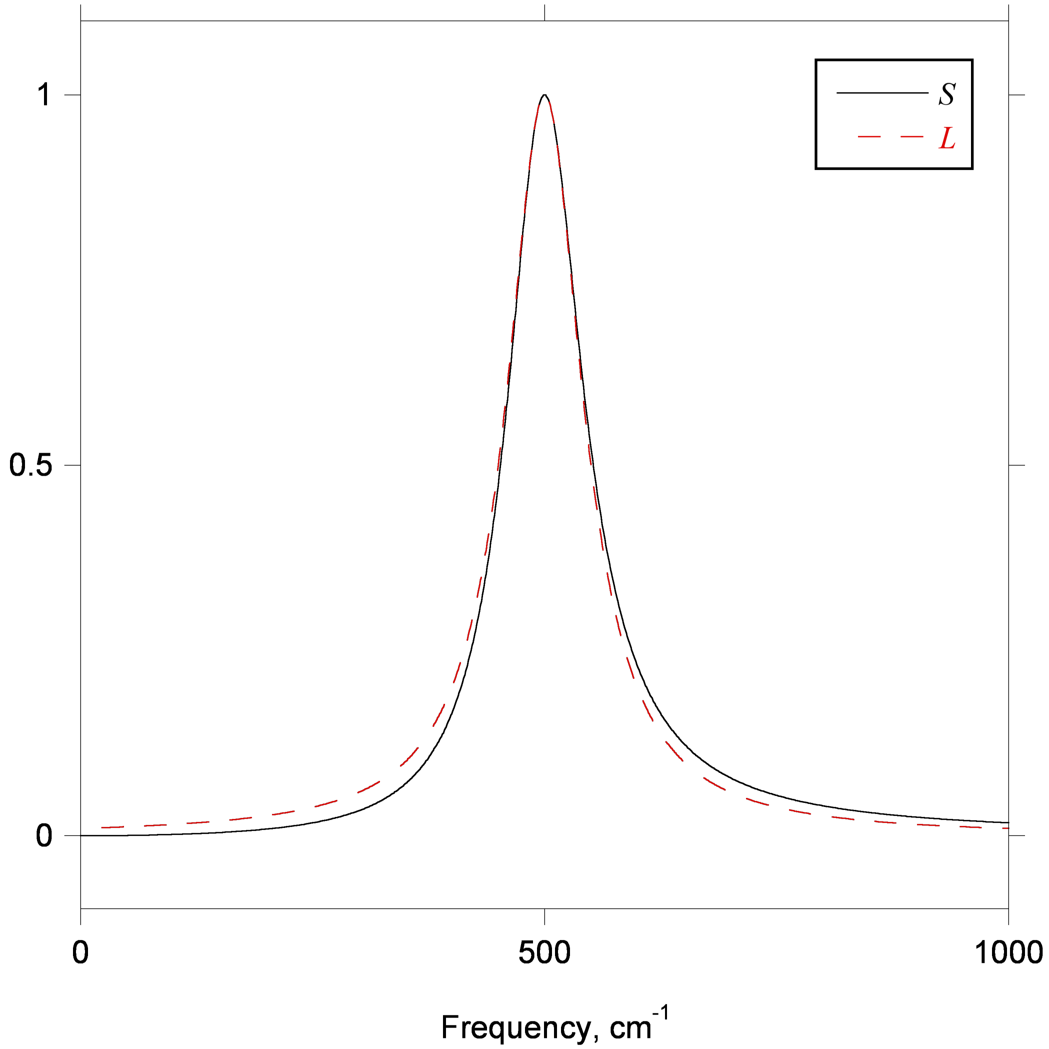

Figure 2.1: Comparison of the line shape $S_n(\omega)$ with a Lorentzian function $L_n(\omega)$. The peaks are centered at $\omega_n = 500\,\mathrm{cm}^{-1}$ and their widths (FWHM) are set by $\gamma_n = \Gamma = 100\,\mathrm{cm}^{-1}$.

2.6.2 *Equivalence to dipole-derivative method*

In papers that calculate IR spectra of molecules from normal modes, it is standard practice to calculate the IR intensity of normal mode n from the square of the derivative of the molecule's dipole moment with respect to normal coordinate Q_n,

$$I_n \propto \left|\frac{\partial \mathbf{p}}{\partial Q_n}\right|^2 = \sum_{k=1}^{3} \left(\frac{\partial p_k}{\partial Q_n}\right)^2 \ , \tag{2.46}$$

where subscript k specifies the Cartesian directions with $1 = x$, $2 = y$, and $3 = z$. To show that eq. 2.46 is equivalent to eq. 2.38, use the chain rule to expand the derivative $\partial p_k / \partial Q_n$ in terms of the original mass-weighted Cartesian displacement coordinates (the q's of eq. 2.2),

$$\frac{\partial p_k}{\partial Q_n} = \sum_{i=1}^{N} \frac{\partial p_k}{\partial q_I} \frac{\partial q_I}{\partial Q_n} \ , \tag{2.47}$$

where subscript I combines the atom index i and direction k according to eq. 2.8. (Mathematically, this sum should have $3N$ terms to include all the q_I's, but clearly the k component of the dipole only depends on the displacements of the N atoms in the k direction.) Formally, one may write each one of the original coordinates, q_I, in terms of the $3N$ normal coordinates as

$$q_I = \frac{\partial q_I}{\partial Q_1} Q_1 + \frac{\partial q_I}{\partial Q_2} Q_2 + \ldots + \frac{\partial q_I}{\partial Q_{3N}} Q_{3N} \ . \tag{2.48}$$

By comparing this with eq. 2.17, one sees that

$$\frac{\partial q_I}{\partial Q_n} = A_{I,n} \ . \tag{2.49}$$

The contribution to the k component of the molecule's dipole moment by displacement q_I of atom i (mass m_i, charge e_i) is given by $e_i \times q_I / \sqrt{m_i}$, where dividing by the square root of the atom's mass is necessary to convert the mass-weighted coordinate q_I back into an ordinary Cartesian displacement (reverse of eq. 2.2). Therefore,

$$\frac{\partial p_k}{\partial q_I} = \frac{e_i}{\sqrt{m_i}} \; . \tag{2.50}$$

Using eqs. 2.50 and 2.49 in 2.47 gives

$$\frac{\partial p_k}{\partial Q_n} = \sum_{i=1}^{N} \frac{e_i}{\sqrt{m_i}} A_{I,n} \; . \tag{2.51}$$

This quantity is then squared and summed over the three Cartesian directions to get the IR intensity of eq. 2.46:

$$I_n \propto \sum_{k=1}^{3} \left(\sum_{i=1}^{N} \frac{e_i \; A_{I,n}}{\sqrt{m_i}} \right)^2 \; . \tag{2.52}$$

The squared quantity above can be rewritten as

$$\begin{aligned} \left(\sum_{i=1}^{N} \frac{e_i \; A_{I,n}}{\sqrt{m_i}} \right)^2 &= \left(\sum_{i=1}^{N} \frac{e_i \; A_{I,n}}{\sqrt{m_i}} \right) \left(\sum_{j=1}^{N} \frac{e_j \; A_{J,n}}{\sqrt{m_j}} \right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{e_i \; A_{I,n} \; A_{J,n} \; e_j}{\sqrt{m_i} \, \sqrt{m_j}} \; , \end{aligned} \tag{2.53}$$

with renamed indices j and J (with $l = k$ in eq. 2.8) for the second sum, which establishes the equivalence between calculating the IR intensities based on the dipole derivatives (eqs. 2.46, 2.52) and eq. 2.38. However, note that eq. 2.52 with its single sum over the atom index i (N terms) offers a much more efficient algorithm for calculating the IR intensity than eq. 2.38's double sum over the atom indices i and j (N^2 terms). Since the intensities will need to be calculated for all $3N - 6$ vibrational modes, the number of operations or time required for this calculation scales as N^2 if eq. 2.52 is used, or N^3 if eq. 2.38 is used. This difference in run time becomes ever more significant as the number of atoms N in the molecule increases. Thus eq. 2.52 effectively replaces the mathematically equivalent but computationally less efficient eq. 2.38.

The theoretical basis for calculating IR intensities based on dipole derivatives comes from using quantum mechanics to calculate the transition rate of a harmonic oscillator going from one of its quantized energy states to the next highest one. Derivations are given by Wilson *et al.* (1955); Zerbi (1982); Galabov and Dudev (1996).

2.6.3 IR intensities from quantum chemistry calculations

In the case of *ab initio* quantum chemistry calculations, instead of representing the atoms as discrete effective partial charges, one considers the nuclei to be fixed point charges surrounded by a continuous electron cloud. One can no longer calculate the IR intensities using discrete charges as in eq. 2.38 or 2.52, but the dipole-derivative method of eq. 2.46 is still valid, provided that one can find an appropriate means of calculating $\partial \mathbf{p}/\partial Q_n$. An efficient way to do this was described by Komornicki and Jaffe (1979). Since the energy of a dipole in an electric field $\mathbf{E}$ is given by $U = -\mathbf{p}\cdot\mathbf{E}$, the x component of the dipole vector is given by $p_x = -\partial U/\partial E_x$, and similarly for the y and z components. Taking the derivative of this with respect to one of the nuclear displacement coordinates X,

$$\frac{\partial p_x}{\partial X} = -\frac{\partial}{\partial X}\frac{\partial U}{\partial E_x} = -\frac{\partial}{\partial E_x}\frac{\partial U}{\partial X}\ , \tag{2.54}$$

where in the last step the order of differentiation was reversed. The derivative $\partial/\partial E_x$ may be obtained from considering finite differences of the quantity $g_X \equiv \partial U/\partial X$ (which is the gradient of the molecule's total energy with respect to the nuclear displacements) subject to two different (small) external field strengths, $E_x = \pm\delta$. Thus all three components of the derivative of the dipole with respect to nuclear displacements can be obtained from six energy-gradient operations (Komornicki and Jaffe, 1979). Then these derivatives with respect to nuclear displacements ($\partial p_x/\partial X$) can be transformed into derivatives with respect to normal coordinates ($\partial p_x/\partial Q_n$) in the same way as eq. 2.47, from which point eq. 2.46 can be used to calculate the IR intensities of the normal modes. According to the support staff at Gaussian, Inc.—the makers of the Gaussian quantum chemistry software package—this is how the Gaussian program calculates IR intensities.

Chapter 3

CALCULATED INFRARED SPECTRA OF NERVE AGENTS AND SIMULANTS

The organophosphorus nerve agents are among the most toxic of the chemical warfare agents. These nerve agents have been categorized as two families of molecules: the G-series agents and the V-series agents. In order to study their properties, such as dispersal, often closely related simulant molecules are used in place of the more toxic nerve agents to reduce risks to human health and the environment. There is considerable interest in rapid, reliable detection and identification of chemical warfare agents in order to determine appropriate countermeasures and decontamination procedures. One of the techniques that is useful for identification of chemicals is infrared (IR) spectroscopy.

One approach to identifying an unknown chemical based on its IR spectrum is to record a library of experimental IR spectra for comparison. A complimentary approach is to simulate the IR spectra using first principles quantum chemistry calculations. These calculations allow us to identify the specific atomic motions within a molecule that are responsible for the various IR peaks that are seen in both the experimental and simulated spectra.

There are some published experimental IR spectra of G-series and V-series nerve agents and their common simulants. Published in 1977, the first in the series of "Blue Books" from the Finnish Institute for Verification of the Chemical Weapons Convention (VERIFIN) contains measured IR spectra of several organophosphorus molecules. A collection of interpreted IR spectra of organophosphorus compounds was published by Shagidullin *et al.* (1990). IR

Spectra were measured for VX by Creasy *et al.* (1997) and Sarin by Durst *et al.* (1998). Söderström measured IR spectra of several V-series chemicals in the condensed phase using cryodeposition gas chromatography Fourier transform infrared spectroscopy (GC-FTIR) (Söderström, 1998). Matrix isolation IR spectroscopy of the simulant dimethyl methylphosphonate (DMMP) and its 1:1 hydrogen-bonded complex with water was performed by Ault *et al.* (2004), who also showed a comparison with an *ab initio* calculation of the IR spectrum. The 2005 book chapter by Söderström provides an excellent review and many references on Fourier transform infrared spectroscopy (FTIR) of chemical warfare agents (Söderström, 2005). An IR spectrum of DMMP in the gas phase was measured by Bermudez (2007b) and comparison was made with a calculated spectrum. Gurton *et al.* (2007) used FTIR and flow-through photoacoustics in order to measure optical cross sections of aerosols of four simulants over wavelengths from 3 to 13 µm ($1/\lambda$ from 770 to 3300 cm^{-1}).

Early Hartree–Fock calculations on the structure of O-methyl methylphosphonofluoridate, a homolog of Sarin and Soman, were performed by Ewig and Van Wazer (1985). *Ab initio* calculations have also been used to determine the structure and rotational constants of Sarin and DMMP for interpretation of measured microwave spectra (Walker *et al.*, 2001; Suenram *et al.*, 2002). Bermudez performed calculations of the adsorption of DMMP and Sarin onto various surfaces to compare a simulant and a nerve agent (Bermudez, 2007a,b, 2010). There has been considerable recent interest in using molecular dynamics simulations to study the transport of nerve agents and simulants in aqueous solutions and their permeation through barriers (Vishnyakov and Neimark, 2004; Rivin *et al.*, 2004; Vishnyakov *et al.*, 2011). *Ab initio* calculations were used to develop force fields specific to these molecules. Additionally, based on an experimental IR database, Flanigan (1997) used modeling and simulation to study the limits of remote detection of hazardous clouds, including a simulated cloud of Sarin.

For this chapter, the Gaussian 03 software package (Frisch *et al.*, 2003) was used to perform quantum chemistry calculations of IR spectra of five G-series agents, five V-series agents, and five simulants.

Initial structures were obtained from the PubChem Compound Database (Bolton *et al.*, 2008). Although not commonly used as a simulant, dimethyl fluorophosphate (DMFP) was included in this study due to its similarity with the four simulants; Vishnyakov *et al.* (2011) have made use of DMFP in their development of a force field for molecular dynamics simulations of aqueous solutions of organophosphorus compounds.

Table 3.1 lists the G agents, V agents, and simulants. Hereafter, each molecule will be referred to by its abbreviation (e.g., GB) given in the table. For each of the 15 molecules, the table lists two numerical identifiers: the Chemical Abstract Service (CAS) registry number and the PubChem Compound ID (CID). The molecule's name (e.g., Sarin) is listed if such a name exists; in the case of the simulant molecules, the chemical names (e.g., dimethyl methylphosphonate) are given. In all cases chemical formulae are provided.

Table 3.1: G agents, V agents, and simulants.

Abbreviation	Name	CAS #	CID #	Formula
GA	Tabun	77-81-6	6500	$C_5H_{11}N_2O_2P$
GB	Sarin	107-44-8	7871	$C_4H_{10}FO_2P$
GD	Soman	96-64-0	7305	$C_7H_{16}FO_2P$
GE		1189-87-3	65566	$C_5H_{12}FO_2P$
GF	Cyclosarin	329-99-7	64505	$C_7H_{14}FO_2P$
VE		21738-25-0	65568	$C_{10}H_{24}NO_2PS$
VG	Amiton	78-53-5	6542	$C_{10}H_{24}NO_3PS$
VM	Edemo	21770-86-5	30800	$C_9H_{22}NO_2PS$
VR	Russian VX	159939-87-4	178033	$C_{11}H_{26}NO_2PS$
VX		50782-69-9	39793	$C_{11}H_{26}NO_2PS$
DEMP	diethyl methylphosphonate	683-08-9	12685	$C_5H_{13}O_3P$
DMFP	dimethyl fluorophosphate	5954-50-7	80052	$C_2H_6FO_3P$
DIFP	diisopropyl fluorophosphate	55-91-4	5936	$C_6H_{14}FO_3P$
DIMP	diisopropyl methylphosphonate	1445-75-6	3073	$C_7H_{17}O_3P$
DMMP	dimethyl methylphosphonate	756-79-6	12958	$C_3H_9O_3P$

Each molecule is considered to be in the gas phase, treated as a single isolated molecule. Starting from the molecular structures obtained from the PubChem Compound Database (3D SDF files), the structures were optimized to determine their minimum energy configuration using Gaussian 03 quantum chemistry software (Frisch *et al.*, 2003) with the 6-31+G(d,p) basis set and density functional theory (B3LYP), respectively. After obtaining optimized minimum-energy molecular structures, the vibrational modes of the molecule along with the IR intensities associated with each vibrational mode were calculated by Gaussian 03 based on the same basis set and level of theory as was used in the energy-minimization step. The Gaussian program follows the calculation method of Komornicki and Jaffe (1979) for the integrated IR intensities associated with each vibrational mode. To investigate the effect of the choice of basis set on the calculated spectra, separate calculations were performed for two of the molecules, GB and VX, using the 6-311+G(d,p), cc-pVDZ, and cc-pVTZ basis sets.

The final simulated IR spectrum is constructed from summing Lorentzian spectral line profiles of a specified full width at half maximum (FWHM = $24\,\mathrm{cm}^{-1}$) centered at the frequency of each vibrational mode. This choice for the FWHM was based on the observed peak widths in the NIST/EPA Gas Phase Infrared Library (Stein, 1992), which is a collection of IR spectra measured for a large set of small molecules. This peak width is likely a combination of the intrinsic width of an individual IR line and experimental broadening. This choice for the peak width is therefore in agreement with what could be expected from a measured high-resolution IR spectrum.

The simulated IR spectra of the G agents, V agents, and simulants are shown in Figs. 3.1, 3.2, and 3.3 respectively. The most prominent IR lines fall into two regions: a low-frequency region ~500–1600 cm^{-1} and a high-frequency region ~2800–3200 cm^{-1}. In the intermediate region ~1600–2800 cm^{-1}, these molecules have no IR lines—the exception being GA, which has a weak IR line due to a C≡N stretching mode at 2315 cm^{-1}.

The high-frequency lines are solely due to carbon-hydrogen bond-stretching modes. Among these high-frequency lines, individual C−H stretches occur at the lowest frequencies (2924–2974 cm^{-1}), followed by symmetric C−H stretches within methyl groups at higher frequencies (3023–3053 cm^{-1}) and anti-symmetric C−H stretches within methyl groups at the highest frequencies (3072–3140 cm^{-1}).

In the low-frequency region (~500–1500 cm^{-1}, the strongest lines almost always involve motions of a phosphorus atom relative to its neighbors since the phosphorus atom tends to have the highest (negative) effective charge in these molecules. Here the strongest IR line is often due to anti-symmetric stretching modes of P−O−C segments of the molecules. The frequencies of these lines are given in Table 3.2. In about half of the molecules (GA, GB, GE, VE, VM, VR, VX), this line exists as a single strong peak; in other cases (GD, GF, VG, DEMP, DIFP, DIMP, DMFP, DMMP) this line is split into two peaks of nearly equal intensity. In those molecules where there are two P−O−C combinations in slightly different environments, the antisymmetric P−O−C stretch occurs at two slightly different frequencies. For example, VG has two peaks of nearly equal intensity at frequencies of 1040 and 1063 cm^{-1}, corresponding to anti-symmetric stretches of two different P−O−C groups, as shown in Fig. 3.4.

Another common strong line is due to P=O stretching, which occurs in all the molecules in this study. Fig. 3.5 illustrates a P=O stretching mode of Sarin (GB). The P=O stretching frequency varies considerably with different

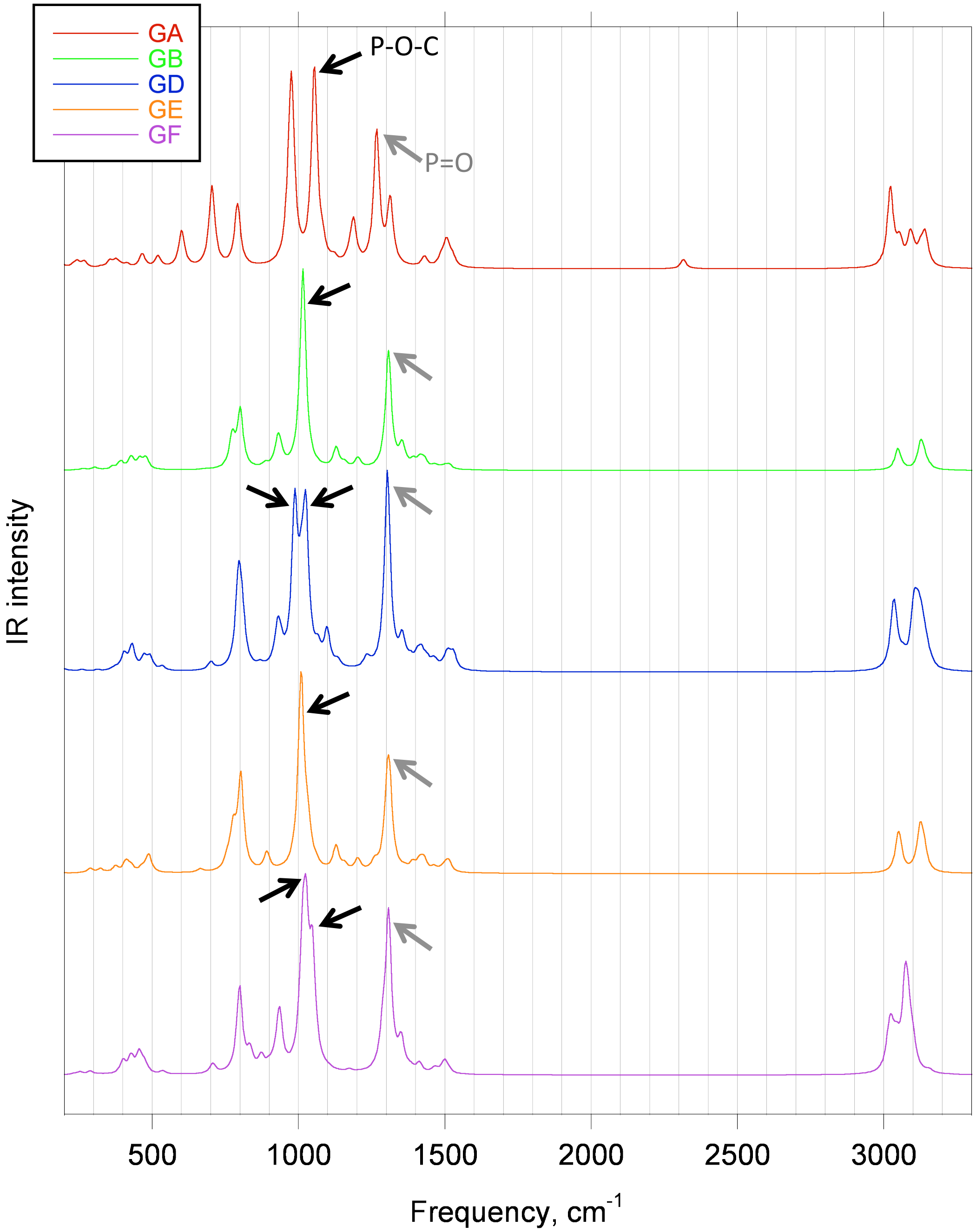

Figure 3.1: Calculated infrared spectra of G-series nerve agents. The spectra have been normalized to have a maximum intensity of 1 and are vertically offset from each other by that same value for clarity. Arrows mark the IR peaks due to bond stretching in P$-$O$-$C and P=O.

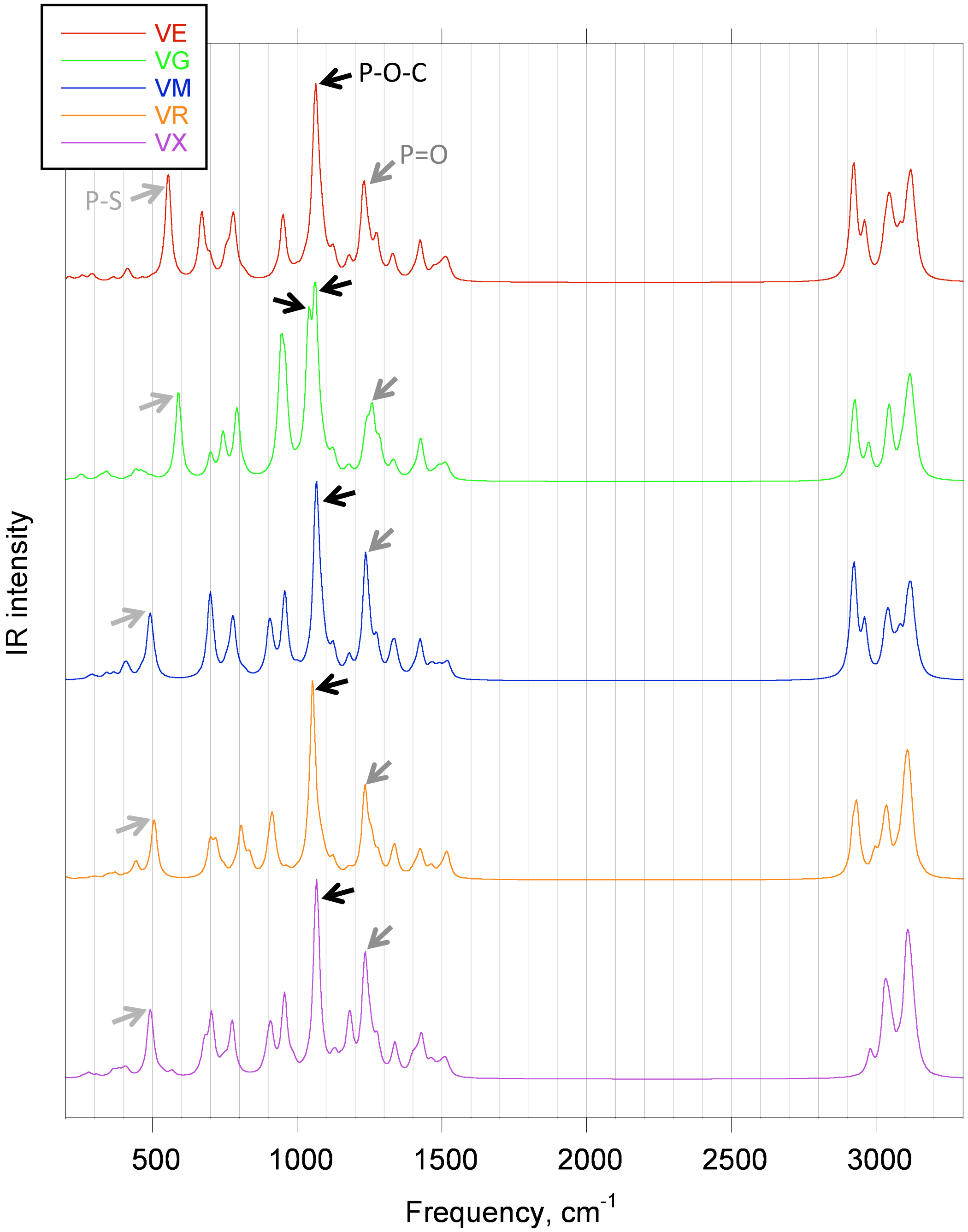

Figure 3.2: Calculated infrared spectra of V-series nerve agents. Arrows mark the IR peaks due to bond stretching in P−O−C, P=O, and P−S.

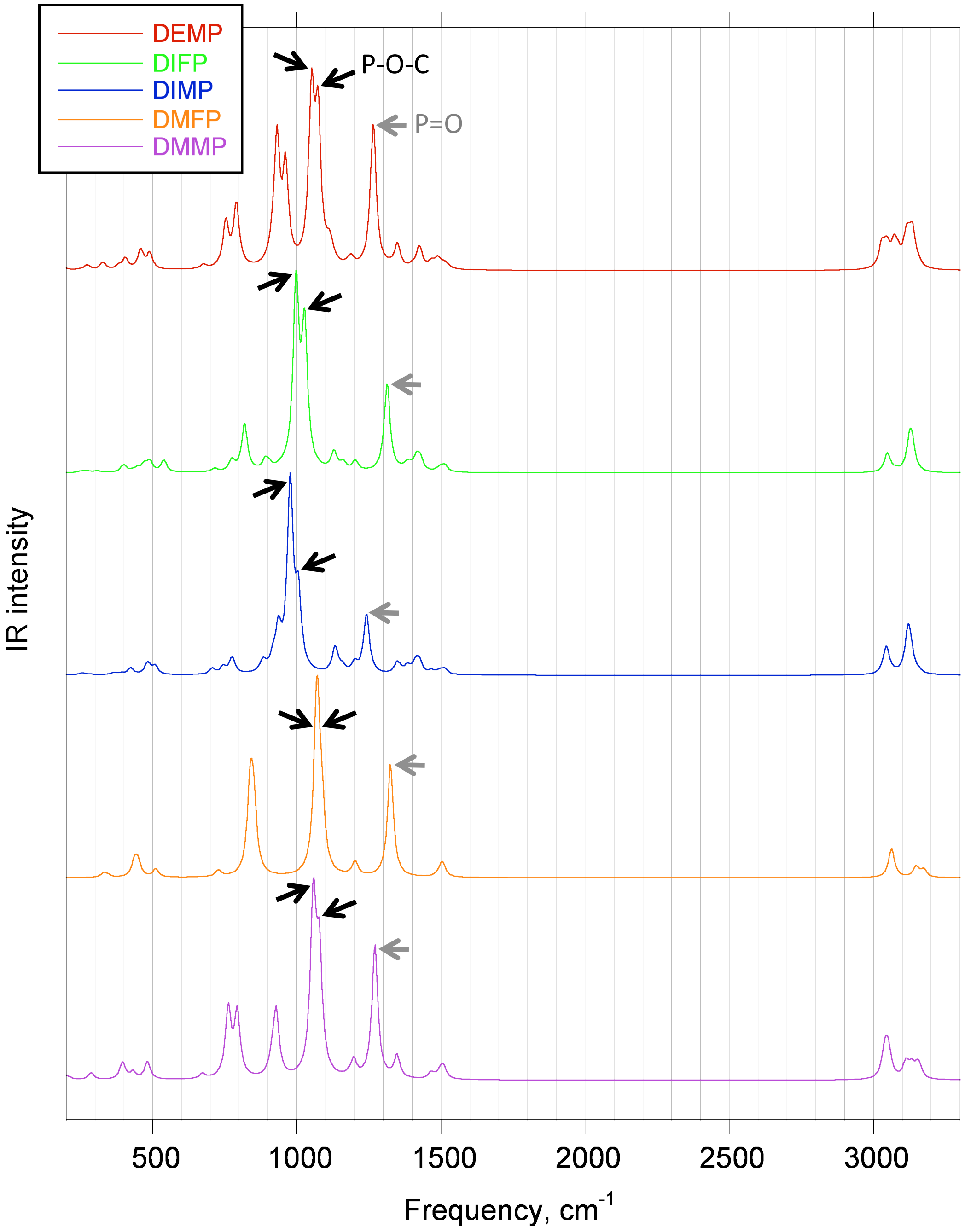

Figure 3.3: Calculated infrared spectra of four simulants and DMFP. Arrows mark the IR peaks due to bond stretching in P−O−C and P=O.

Table 3.2: Anti-symmetric stretches of P−O−C group.

Molecule	Frequency, cm^{-1}
GA	1054
GB	1015
GD	987, 1025
GE	1009
GF	1026, 1048
VE	1065
VG	1040, 1063
VM	1066
VR	1053
VX	1067
DEMP	1051, 1074
DIFP	997, 1027
DIMP	977, 1005
DMFP	1070, 1087
DMMP	1057, 1078

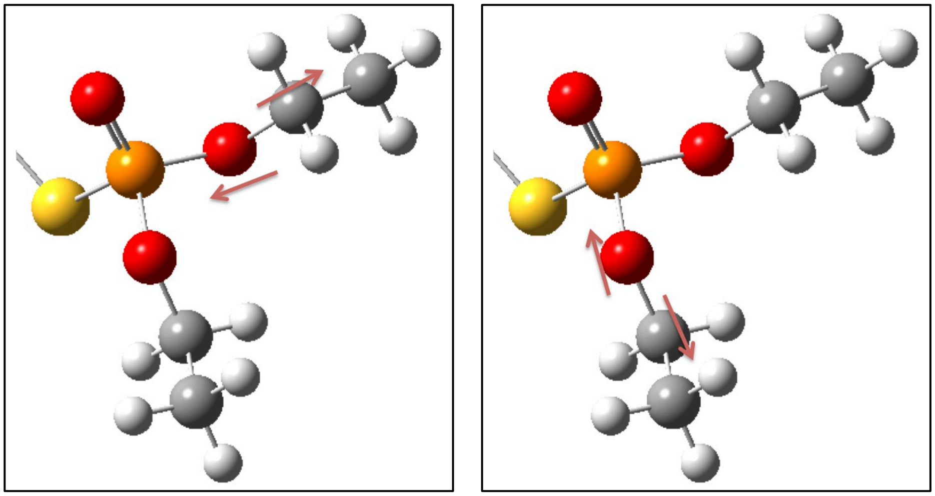

Figure 3.4: Two modes of VG involving anti-symmetric stretches of P−O−C (the portion of the molecule connected to the sulfur atom has been omitted from this diagram). Left image depicts mode with frequency 1040 cm^{-1}; right image depicts mode with frequency 1063 cm^{-1}.

local environments of the phosphorus atom. Across the entire 15-molecule set, the P=O stretching frequency varies from 1228 to 1325 cm^{-1}, but when the molecules are grouped according to the local environment of the phosphorus atom, the P=O stretching frequencies vary less within each subset. For example, the molecules GB, GD, GE, and GF all have their P atom surrounded by O, O, C, and F atoms. Within this subset of four molecules, the P=O stretching frequencies are close together in the range 1304–1312 cm^{-1}. Similar results are found for VE, VM, VR, and VX, which all have P surrounded by O, O, C, and S, with P=O stretching frequencies in the range 1228–1237 cm^{-1}. Table 3.3 gives the frequencies of the P=O stretching modes in all 15 molecules, with the molecules grouped according the local environment of the P atom.

A unique feature of the V agents that differentiates them from the G agents and simulants is the line at ~ 500 cm^{-1} due to P−S stretching. This is because the V agents possess a sulfur atom whereas the G agents and simulants do

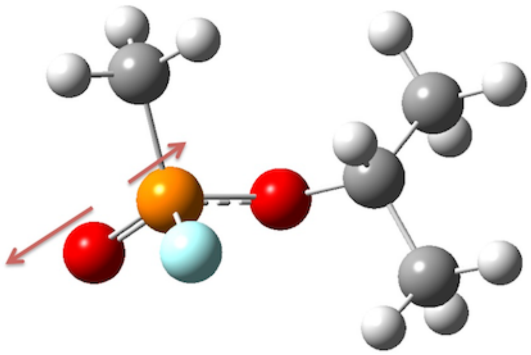

Figure 3.5: P=O stretching mode of Sarin (GB) at 1308 cm^{-1}.

Table 3.3: P=O stretching frequencies. Molecules have been grouped according to the local environment of the phosphorus atom.

Atoms Surrounding P	Molecule	Frequency, cm^{-1}
O,O,C,N	GA	1267
O,O,C,F	GB	1308
	GD	1304
	GE	1312
	GF	1308
O,O,C,S	VE	1228
	VM	1237
	VR	1234
	VX	1232
O,O,O,S	VG	1259
O,O,O,C	DEMP	1265
	DIMP	1242
	DMMP	1271
O,O,O,F	DIFP	1313
	DMFP	1325

not. Like the P=O stretching frequency, the P−S stretching frequency varies considerably with different local environments of the phosphorus atom. Table 3.4 lists each V agent's P−S stretching frequency. Four of the five V agents have P surrounded by O, O, C, and S, with the exception being VG, which has P surrounded by three oxygens and a sulfur atom. Not surprisingly, VG's P−S stretching frequency is an outlier compared to the other V agents. The P−S stretching frequency of VM, VR, and VX is in the range 490–506 $\mathrm{cm^{-1}}$, but VE's is at 554 $\mathrm{cm^{-1}}$ even though VE shares the same local environment for its phosphorus atom as VM, VR, and VX. This may be due to VE having C_2H_5 connected to its P atom, whereas VM, VR, and VX have a methyl group in that position.

Table 3.4: P−S stretching frequencies.

Molecule	Frequency, $\mathrm{cm^{-1}}$
VE	554
VG	590
VM	491
VR	506
VX	490

After repeating the calculations for GB and VX with three other basis sets, the overall shape of the spectra remained largely unchanged. The exception was the cc-pVDZ basis set, which resulted in noticeable differences in the P=O region of the VX spectrum. That basis set has the fewest basis functions of the basis sets tested and is therefore expected to give the least accurate results. The other basis sets that were tested, 6-311+G(d,p) and cc-pVTZ, gave frequency shifts of up to ±20 cm^{-1} for the dominant P−O−C and P=O lines. The intensity of the P−O−C line changed by up to 10%; the intensity of the P=O line changed by as much as 30%. This is in agreement with the findings of Sosa and Schlegel (1987), who showed that calculated IR intensities are more sensitive to the choice of basis set than are the vibrational frequencies.

In the calculated IR spectra (Figs. 1–3), the frequencies of the high-frequency lines (∼2800–3200 cm^{-1}) are higher than the experimental frequencies. For example, in Sarin's IR spectrum measured by Durst *et al.* (1998), the highest frequency line is at 2989.2 cm^{-1}, whereas in my calculated spectrum this line is at 3128 cm^{-1} and is identified with three unresolved C−H stretching modes at nearby frequencies. In contrast with the high-frequency lines, the low-frequency lines (∼500–1600 cm^{-1}) in the calculated Sarin spectrum have frequencies that are in much closer agreement with experiment. Durst et al. measured prominent lines at 1015.17 and 1308.89 cm^{-1}, while my calculations place these lines at 1015.40 cm^{-1} (P−O−C vibration) and 1307.64 cm^{-1} (P=O vibration).

The overestimation of vibrational frequencies is a well-known feature of *ab initio* calculations. A major cause of this overestimation is that standard normal mode analysis treats the potential as being purely quadratic ("harmonic") in the displacements of nuclei from their equilibrium positions, whereas higher order ("anharmonic") terms may exist in the real potential, causing the calculated harmonic vibrational frequencies to be higher than the true observed fundamental vibrational frequencies. Other causes of too-high calculated frequencies include incomplete treatment of electron correlation and finite basis sets (Rauhut and Pulay, 1995; Scott and Radom, 1996; Halls *et al.*, 2001; Merrick *et al.*, 2007).

Various schemes have been proposed to correct the calculated vibrational frequencies for better agreement with experiment. These include rescaling all of the calculated normal mode frequencies with a single scale factor (Scott and Radom, 1996; Halls *et al.*, 2001; Merrick *et al.*, 2007; Alecu *et al.*, 2010), rescaling the high and low frequencies separately with different scale factors (Halls *et al.*, 2001), and rescaling the relevant force constants in the Hessian matrix (Rauhut and Pulay, 1995). Although they improve agreement in the high frequency region, they lead to larger differences between calculation and experiment at the more significant lower frequencies. For this reason it was decided to report the uncorrected harmonic frequencies from the *ab initio* calculations.

In conclusion, quantum chemistry methods were used to calculate IR spectra for the nerve agents and related simulant molecules. The dominant peaks arise from P−O−C and P=O vibrations, and their frequencies and relative intensities are in good agreement with experiment. The V agents have a strong line due to P−S stretching that distinguishes them from the G agents and simulants. Although it should be possible to distinguish whether a given agent belongs to the G or V family, it is unlikely that IR spectroscopy could be used to identify a particular agent. These conclusions are not affected by the choice of basis set used for the calculations.

Chapter 4

IR SPECTRA FROM QUANTUM CHEMISTRY AND FORCE FIELDS COMPARED TO EXPERIMENT

This chapter is an attempt to bridge the gap between two methods of calculating IR spectra: quantum chemistry versus force field. In both cases, the approach used is normal mode analysis. The difference is how the potential energy of the molecule is calculated. For large molecules such as proteins, which contain thousands of atoms, calculating the potential energy using a force field is the only feasible option since the computational cost of quantum chemistry calculations increases rapidly with the size of the molecule.

To test the validity of the force field approach, it is necessary to use small molecules (~ 10 atoms) so that both quantum chemistry and force field calculations can be done. Furthermore, both calculation methods need to be compared to experiment. The NIST/EPA Gas-Phase Infrared Database ("NIST database" hereafter) (Stein, 1992) is useful for this purpose. The 5228 molecules in the database range in size from 2 to 74 atoms. This database contains experimental IR spectra covering the frequency range from ≈ 500 to $\approx 3900\,\mathrm{cm}^{-1}$ in steps of $4\,\mathrm{cm}^{-1}$. The documentation for the database stipulates that these spectra should not be used for quantitative purposes because they do not provide molar absorbances. Since the spectra show the relative absorbance at different frequencies, they are useful for compound identification.

There are many programs available for molecular simulation using force fields. Of these, CHARMM (Chemistry at HARvard Molecular Mechanics)

(Brooks *et al.*, 2009, 1983) is particularly useful for calculation of normal modes because of its built-in utility for vibrational analysis called VIBRAN. CHARMM requires the user to specify which force field to use as its potential energy function. But which force field should be used with molecules from the NIST database? Before answering this, it's useful to first review the form of the force fields used by CHARMM.

4.1 Functional form of the CHARMM force field

The potential energy of a single or many molecules is a sum of the interactions between all the atoms of the system. The CHARMM potential energy function ("force field") treats the atoms as points of mass and charge. The approximate effect of the electrons is included in the effective partial charges of the atoms for calculating electrostatic interactions, as well as in the bond lengths, bond angles, and van der Waals forces between atoms.

The CHARMM force field includes both bonded and non-bonded interactions. The total potential energy is the sum of both kinds: $V = V_{\text{bonded}} + V_{\text{non-bonded}}$. Bonded interactions are those between atoms connected by one, two, or three sequential bonds. These interactions include the potential energy due to bond stretching (two atoms, one bond), bond angle bending (three atoms, two bonds), dihedral (torsion) angle rotation (four atoms, three bonds), improper dihedral angles (four atoms, three bonds), and Urey–Bradley terms (three atoms, two bonds):

$$\begin{aligned} V_{\text{bonded}} = \sum_{\text{bonds}} K_b \left(b - b_0\right)^2 + \sum_{\text{angles}} K_\theta \left(\theta - \theta_0\right)^2 + \sum_{\text{dihedrals}} K_\varphi \left[1 + \cos\left(n\varphi - \delta\right)\right] \\ + \sum_{\text{impropers}} K_\omega \left(\omega - \omega_0\right)^2 \sum_{\text{Urey–Bradley}} K_{\text{UB}} \left(S - S_0\right)^2 \ . \end{aligned} \tag{4.1}$$

While the terms for bonds, angles, and dihedrals are fairly easy to understand, the improper dihedral and Urey–Bradley terms are less obvious. Brooks *et al.* explain: “For three bonded atoms A−B−C, the Urey–Bradley term is a quadratic function of the distance, S, between atoms A and C. The improper dihedral angle term is used at branchpoints; that is, for atoms A, B, and D bonded to a central atom, C, the term is a quadratic function of the (pseudo)-dihedral angle defined by A−B−C−D. Both the Urey–Bradley and improper dihedral terms are used to optimize the fit to vibrational spectra and out-of-plane motions . . . Although the improper dihedral term is used very generally in the CHARMM force fields, the Urey–Bradley term tends to be used only in special cases” (Brooks *et al.*, 2009).

To describe interactions between atoms separated by more than three sequential bonds in the molecule, or between atoms from two separate molecules, non-bonded interactions are included in the potential. These are the electrostatic Coulomb interaction between charged atoms and the Lennard–Jones interaction:

$$V_{\text{non-bonded}} = \sum_{\text{non-bonded}} \left\{ \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon\, r_{ij}} + \varepsilon_{ij}^{\text{min}} \left[\left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}^{\text{min}}}{r_{ij}} \right)^{6} \right] \right\} . \qquad (4.2)$$

The Lennard–Jones potential models the van der Waals interaction using a strong ($\sim r^{-12}$) short-range repulsion of atoms so that atoms may not overlap, and a weak ($\sim r^{-6}$) long-range attraction between atoms.

By specifying the force field used by CHARMM, the user is selecting the set of constants in eqs. 4.1 and 4.2: all the force constants K_b and equilibrium bond lengths b_0; the K_θ and θ_0 for bond angles; the K_φ, n, and δ for dihedrals; the K_ω and ω_0 for impropers; the K_{UB} and S_0 for Urey–Bradley terms; the effective partial charges q_i for Coulomb forces; and the $\varepsilon_{ij}^{\text{min}}$ and R_{ij}^{min} for Lennard–Jones interactions. All of these parameters are specific to the atoms in question and the atoms' context within the molecule—that is, what kinds of atoms are involved in each interaction (carbons, nitrogens, oxygens, hydrogens, *etc.*) and how they are bonded to each other.

4.2 CHARMM General Force Field

Returning to the question of which force field to use with molecules from the NIST experimental IR database, since the CHARMM22 force field (MacKerell *et al.*, 1998, 2004) was constructed specifically for proteins, it is not possible to use it with other molecules whose connectivity is not part of the range of possibilities built into the force field. Therefore, I decided to use the CHARMM General Force Field ("CGenFF"— Vanommeslaeghe *et al.*, 2010) with molecules from the NIST database. This force field was created specifically for use with drug-like molecules and serves as an extension to previous CHARMM all-atom force fields for proteins, nucleic acids, lipids, and carbohydrates. The CGenFF is an ongoing project and a new version of the force field is published online about twice per year. At the time of writing, the most recent version is 2b7, which is the version used here.

4.3 Method

Twenty-one molecules were selected for study because they had already been parameterized in the CGenFF and also had an experimental IR spectrum available in the NIST database. Table 4.1 lists their names, chemical formulae, and two numerical identifiers: the Chemical Abstract Service (CAS) registry number and the PubChem Compound ID (CID). Additionally, the column labeled "RESI" gives each molecule's residue code (*e.g.*, NMA for *N*-methylacetamide) that is used by CHARMM to identify the molecule in the topology file for the force field. In the table, the molecules are divided into seven groups based on similarities in structure.

As was done previously for the nerve agents and simulants of Chapter 3, initial structures were obtained from the PubChem Compound Database (Bolton *et al.*, 2008). The calculation of normal modes and IR intensities was done in two ways: first, *ab initio* using Gaussian 03; second, with CHARMM using force fields. The quantum chemistry calculations were performed following the same procedure as in Chapter 3, again using density functional theory with the B3LYP hybrid functional and the 6-31+G(d,p) basis set. In the force field calculation, the atomic coordinates were optimized to obtain a minimum-energy structure through 1000 steps of steepest descent minimization followed by 1000 steps of conjugate-gradient minimization. After minimization, the normal modes were calculated using CHARMM's vibrational analysis utility (VIBRAN). The IR intensity associated with each normal mode was calculated from the normal mode vectors and partial charges according to eq. 2.52.

After obtaining the normal mode frequencies and associated IR intensities for both calculation methods, the IR spectra were simulated by giving the IR lines finite width through convolution with a Lorentzian peak profile. By inspection of several experimental spectra from the NIST database, it was found that the narrowest peaks typically had a full width at half maximum (FWHM) of $\approx 24\,\mathrm{cm}^{-1}$. Therefore, a Lorentzian (eq. 2.45) with this value for the FWHM was used as the line shape for the IR peaks associated with the normal modes.

Table 4.1: Twenty-one molecules that have experimental IR spectra in the NIST/EPA Gas-Phase Infrared Database and have parameters in the CHARMM General Force Field. They are divided into seven groups based on similar structure.

Name	Formula	RESI	CAS #	CID #
N-methylacetamide	C3H7NO	NMA	79-16-3	6582
2-pyrrolidinone	C_4H_7NO	2PDO	616-45-5	12025
nicotinamide	$C_6H_6N_2O$	3NAP	98-92-0	936
toluene	C_7H_8	TOLU	108-88-3	1140
p-xylene	C_8H_{10}	PXYL	106-42-3	7809
3-methylpyridine	C_6H_7N	3MEP	108-99-6	7970
cyclohexene	C_6H_{10}	CHXE	110-83-8	8079
thiane	$C_5H_{10}S$	THPS	1613-51-0	15367
2,3-dihydrofuran	C_4H_6O	2DHF	1191-99-7	70934
thiazole	C_3H_3NS	THAZ	288-47-1	9256
azulene	$C_{10}H_8$	AZUL	275-51-4	9231
benzimidazole	$C_7H_6N_2$	ZIMI	51-17-2	5798
benzothiazole	C_7H_5NS	ZTHZ	95-16-9	7222
quinoline	C_9H_7N	QINL	91-22-5	7047
isoquinoline	C_9H_7N	IQIN	119-65-3	8405
anthracene	$C_{14}H_{10}$	ANTR	120-12-7	8418
acridine	$C_{13}H_9N$	ACRD	260-94-6	9215
phenazine	$C_{12}H_8N_2$	FENZ	92-82-0	4757
phenoxazine	$C_{12}H_9NO$	FEOZ	135-67-1	67278
phenothiazine	$C_{12}H_9NS$	FETZ	92-84-2	7108
carbazole	$C_{12}H_9N$	CRBZ	86-74-8	6854

4.4 Amides: a connection to proteins

The first group in Table 4.1 consists of N-methylacetamide, 2-pyrrolidinone, and nicotinamide. These molecules are classified as amides because they contain O=C−N−H (usually written as CONH) as part of their structures. This same amide subunit occurs along the backbone of proteins; it is the peptide bond that links one amino acid residue to the next in the polypeptide chain (see Fig. 5.1). As will be discussed in Chapter 5, the peptide bonds of a protein result in an IR signature that is characteristic of the protein's secondary-structure content—the fraction of the protein sequence that is in α-helices or β-sheets. In particular, the small, 12-atom molecule N-methylacetamide has been used as a model compound to study the molecular vibrations of the amide groups of the protein backbone (Miyazawa *et al.*, 1958; Miyazawa, 1960; Gaigeot and Sprik, 2003; Gaigeot *et al.*, 2005; Schultheis *et al.*, 2008; Schropp *et al.*, 2010; Kaminský *et al.*, 2011). Keeping in mind the end goal of calculating the IR spectra of proteins using force fields, it is worth paying special attention to these three amide molecules, whose structures are shown in Fig. 4.1. Of the three, N-methylacetamide is the most protein-like because it is a chain molecule with the amide subunit CONH located between two carbon atoms (methyl groups) that are analogous to the C_α atoms of the protein backbone.

Figs. 4.2, 4.3, and 4.4 show that the *ab initio* calculations closely resemble the experimental spectra of N-methylacetamide, 2-pyrrolidinone, and nicotinamide. From observing the overall shape—the frequencies and relative heights of peaks—it's easy to see the one-to-one correspondence between all major experimental peaks and their *ab initio* counterparts. Unsurprisingly, as

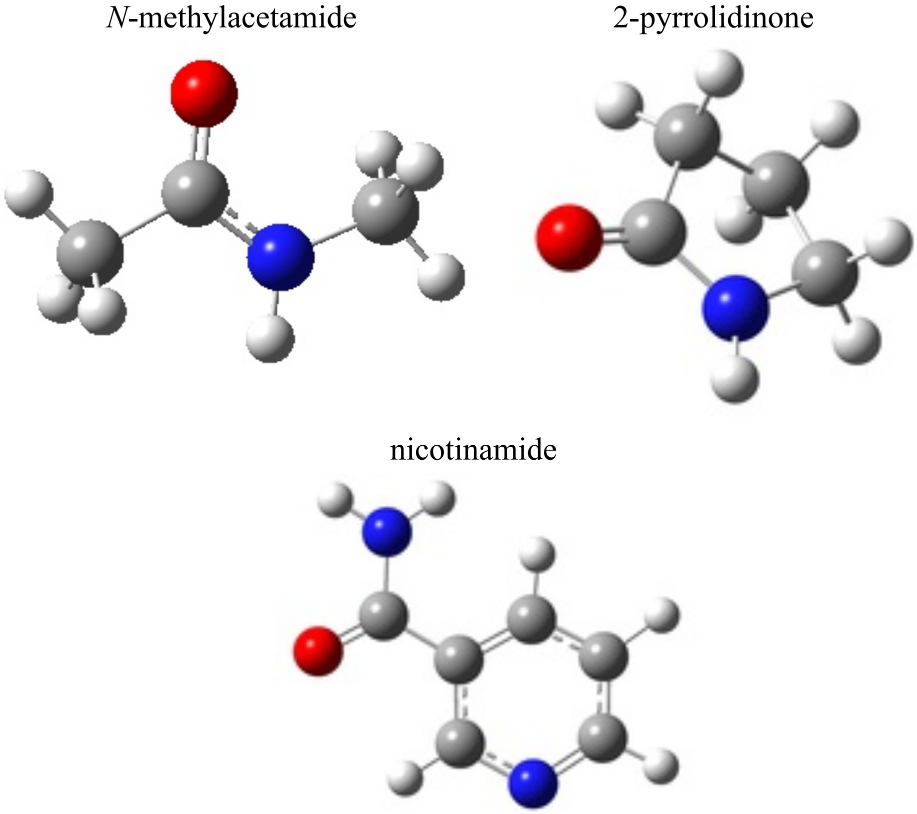

Figure 4.1: Structures of three amides: N-methylacetamide, 2-pyrrolidinone, and nicotinamide.

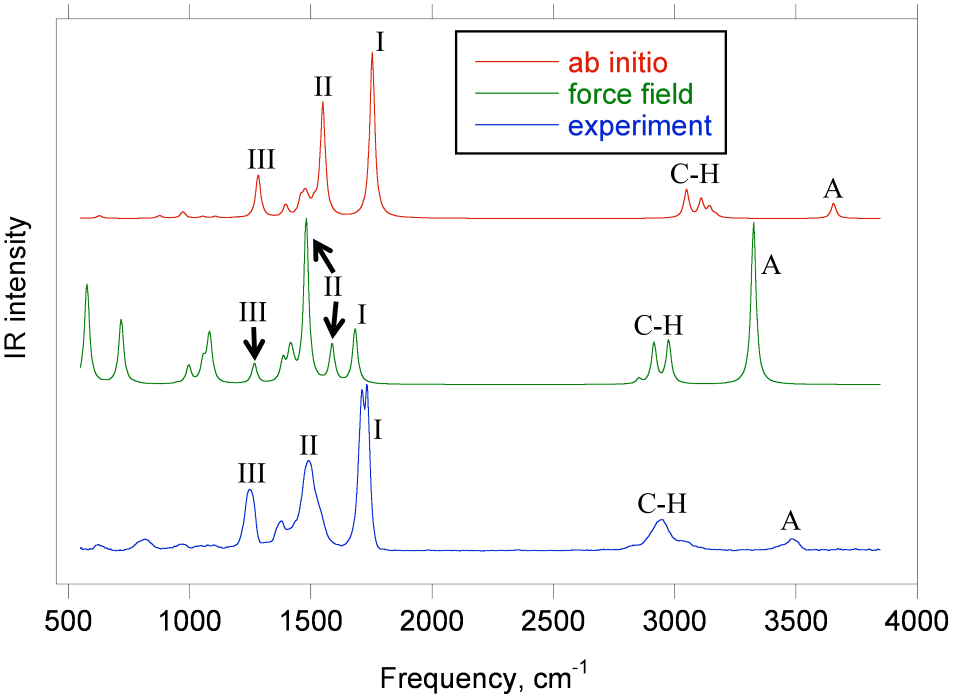

Figure 4.2: IR spectrum of *N*-methylacetamide: two calculated spectra compared to experimental spectrum.

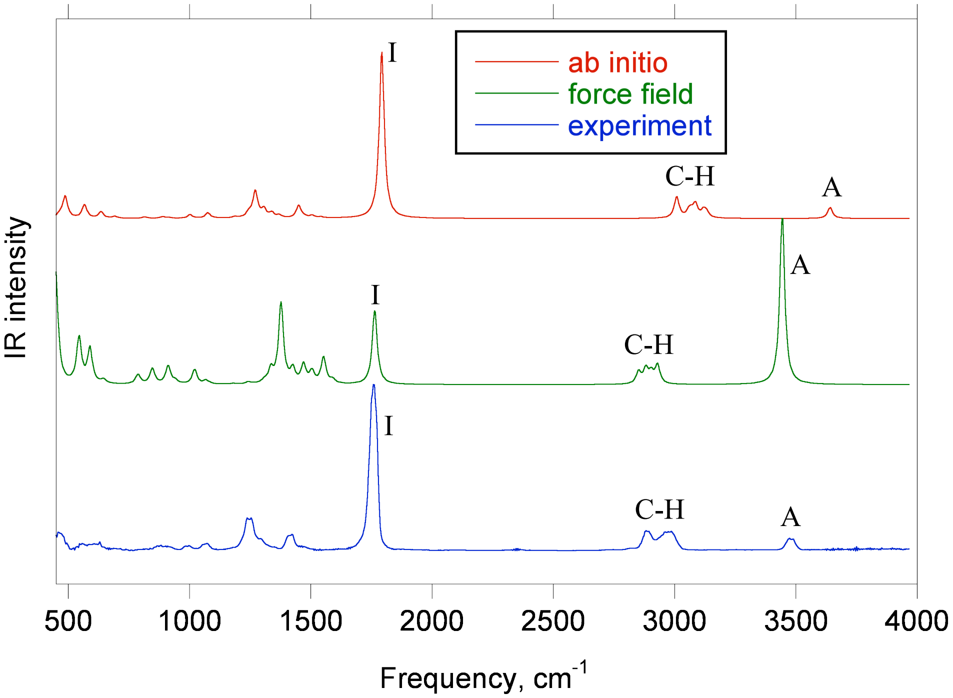

Figure 4.3: IR spectrum of 2-pyrrolidinone: two calculated spectra compared to experimental spectrum.

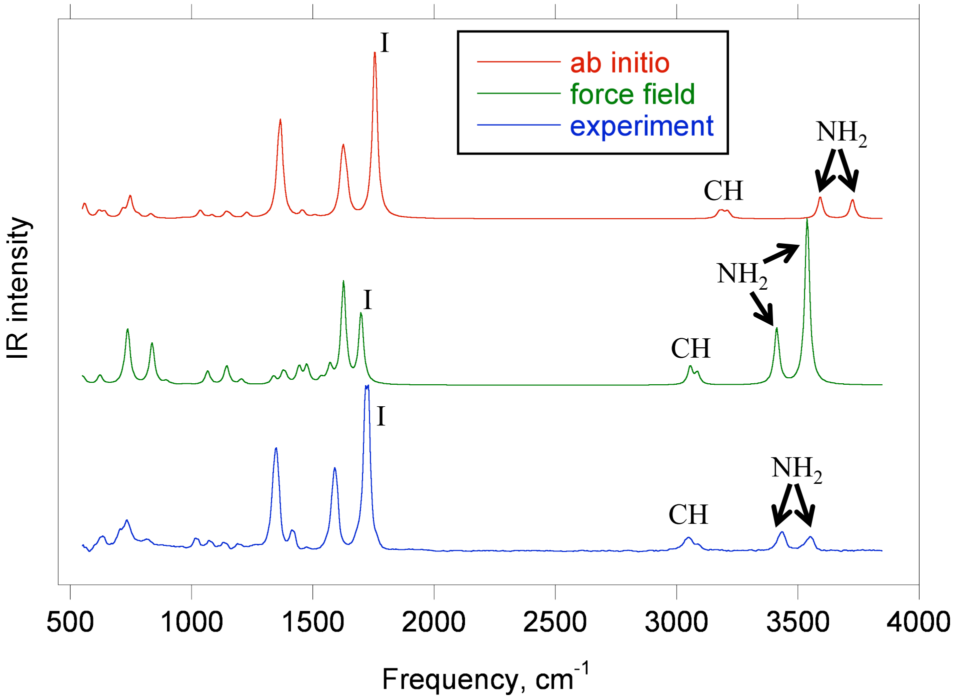

Figure 4.4: IR spectrum of nicotinamide: two calculated spectra compared to experimental spectrum.

with the nerve agents and simulants of Chapter 3, the *ab intio* calculations place the high frequency carbon-hydrogen and nitrogen-hydrogen stretches at too high frequencies. In contrast, for peaks with frequencies less than $\sim 2000\,\text{cm}^{-1}$, the agreement of the *ab initio* frequency with experiment is much better.

Compared to the *ab initio* spectra, the force field calculation more accurately predicts the frequencies of the high-frequency C—H and N—H stretching modes, but exaggerates their intensities. This may be due to the force field assigning to the hydrogen atoms effective partial charges that are too high. Below $2000\,\text{cm}^{-1}$, the force field spectra differ significantly from the experimental spectra. Of the three molecules, 2-pyrrolidinone had the best experimental agreement of its force field based spectrum, with the calculated peaks at 1377, 1553, and 1764 cm^{-1} corresponding to the experimental peaks at 1254, 1422, and 1758 cm^{-1}. For the other two molecules, it is difficult to make the correspondence between the peaks predicted by the force field and those in the experimental spectra. Keep in mind that in the force field calculation, the positions (frequencies) of the IR peaks are determined by the force field parameters, whereas the intensities are primarily determined by the partial charges.

One spectral feature that is observed in the experimental spectra and predicted by both calculation methods is the strong absorption line due to the C=O bond-stretching vibration, whose frequencies are given in Table 4.2. This line is among the strongest observed in the IR spectra of proteins. In the standard nomenclature for the various IR lines associated with vibrations of the amide group (Susi, 1972), this is called the amide I line. Simultaneous to the C=O stretch, the amide I mode involves C−C−N angle bending and N−H bending. The fact that the frequency of the amide I line is sensitive to the local environment of the H−N−C=O group involved in the vibration is a major reason why IR spectroscopy of the amide I band can provide quantitative information about a protein's secondary-structure content.

Table 4.2: Experimental and calculated frequencies of the C=O stretching mode (the amide I line).

Molecule	Frequency, cm^{-1}		
	ab initio	force field	experiment
N-methylacetamide	1753	1683	1721
2-pyrrolidinone	1793	1764	1758
nicotinamide	1756	1699	1730

Another IR line that is common between all three molecules is the amide A line, which is due to N−H stretching. The calculated and experimental amide A frequencies are given in Table 4.3. While *N*-methylacetamide and 2-pyrrolidinone each have a single amide A line corresponding to their single N−H bond, this line gets split into two in the case of nicotinamide, which has an NH_2 group. The two N−H bonds may either stretch in phase or out of phase, resulting in nicotinamide's two N−H stretching modes with different frequencies (Table 4.3).

Table 4.3: Experimental and calculated frequencies of the N−H stretching mode (the amide A line).

Molecule	Frequency, cm^{-1}		
	ab initio	force field	experiment
N-methylacetamide	3655	3326	3482
2-pyrrolidinone	3641	3444	3474
nicotinamide	3592, 3725	3413, 3538	3434, 3550

Two other lines that are observed in the experimental and calculated spectra of *N*-methylacetamide are amide II and III (frequencies in Table 4.4). The amide II mode is a combination of N−H in-plane bending with C−N bond stretching and lesser contributions from C=O bending and C−C stretching (Barth and Zscherp, 2002). The amide III mode is a similar motion to amide II, except with the N−H bending motion having the opposite phase. These lines are also observed in IR spectra of proteins. It is difficult to identify pure amide II and III lines for 2-pyrrolidinone and nicotinamide because the normal modes that involve motions of the relevant atoms are complicated by simultaneous ring distortions.

Table 4.4: Experimental and calculated frequencies of the amide II and III lines of N-methylacetamide. The force field calculation gave two modes with atomic motions consistent with the amide II mode.

Molecule	Frequency, cm^{-1}		
	ab initio	force field	experiment
amide II	1550	1481, 1588	1490
amide III	1283	1268	1246

Since the force field necessarily only gives an approximation to the true energy of the molecule in a given conformation, the normal modes resulting from a force field calculation will not be identical, either in frequency or detailed atomic motions, to those obtained from a more precise quantum mechanics based calculation. Of course, not even a quantum *ab initio* calculation is exact since there are approximations inherent to whichever computational framework is used (*e.g.*, Hartree–Fock or density functional theory), and to whichever finite basis set is used, that are necessary to make the many-particle quantum problem computationally tractable. Thus it is a nontrivial task to unambiguously match up the normal modes obtained from the force field calculation to those obtained from the *ab initio* calculation. This was acknowledged by Vanommeslaeghe *et al.* (2010): "the assignment of a selected QM [quantum mechanics, *ab initio*] normal mode to an MM [molecular mechanics, force field] normal mode is often qualitative in nature, requiring an empirical decision by the user." For a few modes, as in the case of the amide I and amide A modes of the three molecules previously mentioned, the correspondence between the force field and *ab initio* mode will be obvious enough to allow for a unique match. In other cases, there may be no clear match either because no similar mode is found in the other set, or because there is more than one suitable match. Such is the case with the amide II mode of *N*-methylacetamide, for which two modes were found from the force field calculation that reasonably matched the amide II atomic motions seen in the *ab initio* calculation (Table 4.4).

4.4.1 How accurate are force fields for calculating IR?

There are two major approximations involved in using a force field to calculate an IR spectrum. First, the force field approach approximates the molecule's energy for a given geometry using an analytic function that parameterizes the potential energy in terms of bond lengths, bond angles, electrostatic interactions between effective atomic point charges, and van der Waals interactions. Typically the force field parameters (*e.g.*, equilibrium bond lengths, bond "spring" constants K_b) are chosen to agree with experimentally known bond lengths, bond angles, and vibration frequencies. The force field parameters are also informed by *ab initio* calculations of molecular subunits (*e.g.*, to calculate the energy of a small molecule over a full rotation of a dihedral angle so that these data can be fit to a simple analytic function that approximates the dependence of the energy on the rotation of that dihedral angle). As discussed in Chapter 2, the normal modes of vibration are derived from the second derivatives of the potential energy with respect to the atoms' displacements from their equilibrium positions. Thus approximating the molecule's potential energy using a force field results in normal mode frequencies and eigenvectors (detailed atomic motions) that are different from those resulting from the more accurate *ab initio* methods. The extent to which the normal modes derived from a force field agree with those derived from quantum mechanics based calculations is a measure of how accurately the force field approximates the potential energy of the molecule as a function of the atoms' positions.

The second approximation is in the calculation of the IR intensities of the normal modes. In the force field approach, one makes the approximation of calculating the molecule's electric dipole moment using effective partial charges of the atoms instead of a continuous distribution of electron charge. (These partial charges were also part of the first approximation—using a force field to approximate the potential energy.) Recall from eq. 2.46 that the IR intensity of a normal mode Q_i is proportional to $|\partial\mathbf{p}/\partial Q_i|^2$, where $\mathbf{p}$ is the electric dipole moment of the molecule and $\partial\mathbf{p}/\partial Q_i$ is its derivative with respect to displacements of the atomic nuclei from their equilibrium positions along normal coordinate Q_i. An *ab initio* calculation gives the electrons' distribution (probability density) in terms of molecular orbitals. This enables a direct calculation of the dipole (and hence also the dipole derivative) using the charge density

$$\rho(\mathbf{r}) = \rho_e(\mathbf{r}) + \sum_{i=1}^{N} Z_i e\, \delta^3(\mathbf{r} - \mathbf{r}_i)\ , \tag{4.3}$$

where $\rho_e(\mathbf{r})$ is the charge distribution of the electrons only, and the nucleus of atom i has position $\mathbf{r}_i$, atomic number Z_i and nuclear charge $+Z_i e$, considered to be a point charge. Then the dipole moment is

$$\mathbf{p} = \int d^3r\, \rho(\mathbf{r})\, \mathbf{r} = \sum_{i=1}^{N} Z_i e\, \mathbf{r}_i + \int d^3r\, \rho_e(\mathbf{r})\, \mathbf{r}\ . \tag{4.4}$$

That is, $\sum_i Z_i e\, \mathbf{r}_i$ is the contribution of the nuclei to the dipole moment, and $\int d^3r\, \rho_e(\mathbf{r})\, \mathbf{r}$ is the contribution of the electrons. In principle, using the molecular orbitals the dipole derivative $\partial\mathbf{p}/\partial Q_i$ can be calculated numerically through finite difference by first calculating $\mathbf{p}$ in the equilibrium geometry and then recalculating $\mathbf{p}$ with the new electron charge distribution after small displacements of all nuclei along normal coordinate Q_i, although there are more efficient ways of performing this calculation (see eq. 2.54).

The use of partial charges reduces the calculation of the dipole to only include the nuclei, with the charge of the nucleus $+Z_i e$ replaced by an effective partial charge q_i such that $\sum_i q_i = 0$ for a neutral molecule. That is, the charge of the electrons is lumped in with the nuclei. There are many schemes for assigning effective partial charges to the atoms; examples include the Mulliken charges (Mulliken, 1955) and APT charges (Cioslowski, 1989). King (1982) described how effective atomic charges could be derived from experimental IR spectra. Assigning a fixed partial charge to each atom assumes that the charge due to nearby electrons is a point charge located at the atomic nucleus. This neglects the extended, continuous nature of the electrons' charge distribution within the molecule. It also neglects that the electron charge distribution rapidly adjusts in response to displacements of the atomic nuclei. That is, rather than being a fixed value for each atom, the effective atomic charge should adjust dynamically in response to changes in molecular geometry (bond lengths and bond angles).

That the approximation of fixed partial charges results in inaccurately calculated IR intensities of vibrational modes was recognized by Torii and Tasumi (1993). Those authors demonstrated the effect of dynamic charges, which they called "charge flux", in terms of the difference in dipole derivative $(\partial \mathbf{p}/\partial S)_{\mathrm{MO}} - (\partial \mathbf{p}/\partial S)_{\mathrm{FPC}}$, where S is some internal coordinate (a bond length or bond angle). Here the subscripts refer to the two calculation methods: "MO" uses *ab initio* molecular orbitals and "FPC" uses fixed partial charges. Torii and Tasumi studied a small α-helical polypeptide to find out whether or not the fixed partial charges that are used to calculate electrostatic interactions in molecular dynamics simulations of proteins are suitable for calculations of the IR intensities of vibrations of the peptide group. They concluded that the fixed partial charges were insufficient for this purpose and that charge flux (dynamic charge) needs to be included if one is to obtain accurate IR intensities. Kubelka *et al.* (2009, section 4.2) provides some references on the use of partial charges and comments: “These methods are not very accurate since electron charge responds virtually instantaneously to nuclear motion, the basis of the BO [Born–Oppenheimer] approximation. However, they may be still useful in simplified QM/MM [quantum mechanics/molecular mechanics] models to obtain a first approximation of the spectral intensities”.

The use of fixed partial charges gives only a rough approximation of IR intensities. For greater accuracy, *ab initio* methods are necessary. One way of including the effect of dynamic charge is the equilibrium charge/charge flux (ECCF) model used by Torii and Tasumi (1993). But this is just another way of saying that *ab initio* calculations are necessary, as Torii and Tasumi derived the charge fluxes for their α-helical polypeptide from *ab initio* calculations of a smaller molecule (N-methylacetamide) that was used as a model for the peptide bond. For larger molecules such as proteins, for which *ab initio* calculations are computationally impractical, refinement of the partial charges may be necessary if one is to obtain calculated IR spectra that are in better agreement with experiment. That is, the effective partial charges that are appropriate for calculating electrostatic forces in molecular dynamics simulations may be different from the charges that are appropriate for calculating the changes in the dipole moment with nuclear displacements.

4.4.2 Effect of using different partial charges

One check that can be done is to use different sets of partial charges with the same set of normal mode eigenvectors to calculate IR intensities of the modes. That way, one can be sure that differences in the calculated spectra are due solely to the partial charges since in each case the normal mode frequencies and atomic motions are identical. It has already been seen (Figs. 4.2–4.4) that the *ab initio* IR spectra are in excellent agreement with experimental spectra. Thus the *ab initio* normal modes are a good choice for testing different sets of partial charges. This allows one to compare IR intensities calculated using partial charges to IR intensities that were calculated using molecular orbitals.

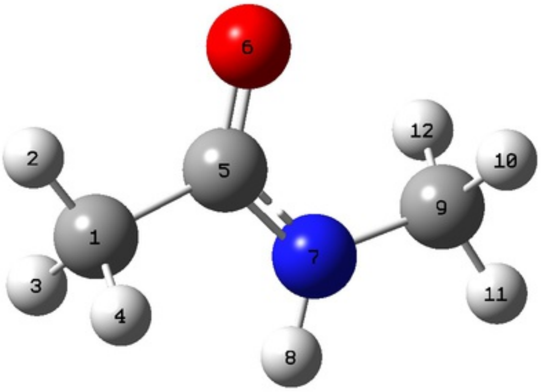

Figure 4.5: Key to numbering of atoms of *N*-methylacetamide in Table 4.5.

Three sets of partial charges for *N*-methylacetamide are shown in Table 4.5 with the atoms numbered as in Fig. 4.5. These are the Mulliken charges (Mulliken, 1955), the APT charges that are defined for each atom as $^1/_3$ of the trace of its atomic polarizability tensor (Cioslowski, 1989), and the charges assigned by the CHARMM General Force Field (CGenFF). These three sets of partial charges were used with the same set of *ab initio* normal modes to

Table 4.5: Three sets of partial charges (in units of e) for N-methylacetamide: Mulliken charges, charges derived from the atomic polar tensor (APT), and charges from the CHARMM General Force Field (CGenFF). For each set, the net charge is zero as expected for this neutral molecule. Key to atom numbering is in Fig. 4.5.

Index	Atom	Mulliken	APT	CGenFF
1	C	−0.562	−0.070	−0.27
2	H	0.188	0.028	0.09
3	H	0.167	0.004	0.09
4	H	0.167	0.004	0.09
5	C	0.489	1.066	0.51
6	O	−0.536	−0.780	−0.51
7	N	−0.386	−0.734	−0.47
8	H	0.299	0.173	0.31
9	C	−0.296	0.355	−0.11
10	H	0.168	−0.019	0.09
11	H	0.133	−0.008	0.09
12	H	0.168	−0.019	0.09
	Sum:	0.000	0.000	0.00

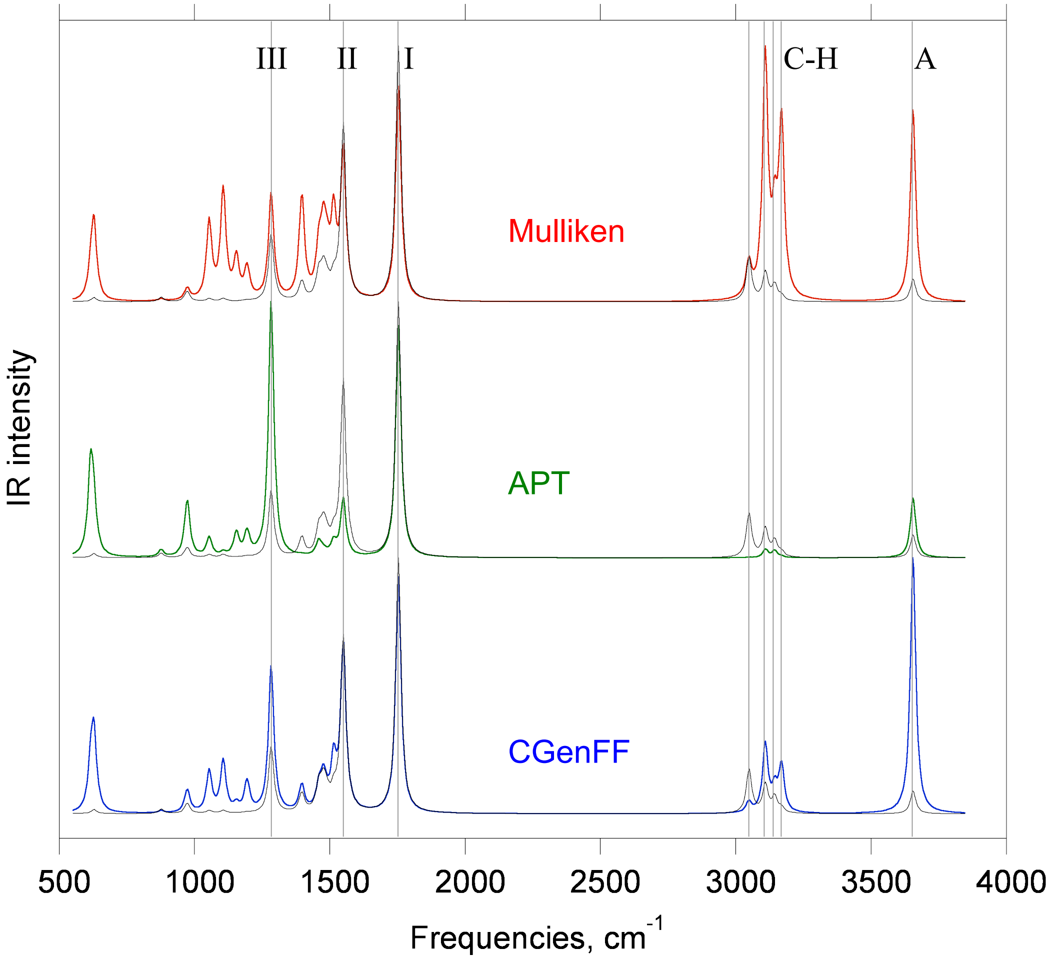

Figure 4.6: IR spectra of *N*-methylacetamide calculated using the *ab initio* normal modes with three different sets of partial charges: Mulliken, APT, and CGenFF. Each is shown in comparison to the same reference spectrum (thin black curve) in which the intensities of the modes were calculated using the *ab initio* molecular orbitals.

calculate IR intensities. The resulting spectra are shown in Fig. 4.6. They are each compared to the *ab initio* spectrum from Fig. 4.2, which was derived from the same set of normal modes but with the mode intensities calculated more accurately using molecular orbitals instead of partial charges.

Of the three sets of charges, the Mulliken charges gave the worst agreement with the *ab initio* IR intensities. Aside from the gross exaggeration of the high-frequency peaks due to C−H and N−H stretching modes, many erroneously high peaks appear in the 1000–1200 cm^{-1} range that are not evident in the *ab initio* spectrum. The APT charges result in a higher quality spectrum than the Mullikens. In particular the heights of the X−H peaks are in good proportion to the dominant C=O peak; this is due to the lowering of the hydrogen charges (see Table 4.5). However, the amide III line at 1283 cm^{-1} has been exaggerated in strength by a factor of ≈ 3 and the amide II line at 1550 cm^{-1} has been diminished by about the same factor. Of the three sets, the CGenFF charges give the best results in the low-frequency region below 2000 cm^{-1}, although the 1000–1200 cm^{-1} region was exaggerated as in the case of the Mullikens, but not as badly. All three sets of charges exaggerated the intensity of the peak at 628 cm^{-1}, which is due to an in-plane bending motion of the entire molecule, combined with a C−C stretch. Note that even though the Mulliken and APT charges were both derived from population analysis of the *ab initio* molecular orbitals, the IR intensities obtained from these charges were actually *worse* than the CGenFF charges in replicating the *ab initio* IR intensities. This demonstrates that the use of partial charges introduces significant errors in the calculation of IR intensities, even if those partial charges were derived from the same molecular orbitals that were used to calculate the *ab initio* intensities. Information has been lost in substituting an effective atomic charge in place of the 3×3 atomic polar tensor (the derivatives of the x, y, and z components of the molecule's dipole moment with respect to an atom's displacement in the x, y, and z directions), which contains all the information necessary to accurately calculate the dipole derivative $d\mathbf{p}/dQ_i$ for a normal mode Q_i.

In comparing the middle spectrum in Fig. 4.2 with the bottom spectrum in Fig. 4.6, one sees the different results obtained from using the same set of partial charges (CGenFF) with two different sets of normal modes—those calculated using the force field and those calculated *ab initio*. Together these figures show that, at least in the case of *N*-methylacetamide, the poor experimental agreement of the force field based spectrum (Fig. 4.2) is less due to the CGenFF partial charges than it is due to the normal modes that were calculated from the force field. When the CGenFF charges were used with the *ab initio* normal modes to calculate the IR spectrum (Fig. 4.6, bottom), the relative heights of the major peaks (amide I, II, and III) agreed reasonably well with experiment. But when these same charges were used with the force field based normal modes (Fig. 4.2, middle), the relative heights of the major peaks were completely wrong.

4.5 Reducing hydrogen charges in intensity calculation

From inspecting many normal modes, one observes that typically the hydrogen atoms have the largest displacements in most molecular vibrations. This means that the calculated spectrum is particularly sensitive to the partial charges of the hydrogen atoms since the atomic displacements are multiplied by the partial charges to calculate the IR line strength (intensity) of each mode (eq. 2.52). If the hydrogen charges are too high, then the intensities of many normal modes will be exaggerated. This explains why in Fig. 4.6 the spectrum calculated using the Mulliken charges has many peaks with intensities that are much higher than those of the experimental peaks: Table 4.5 shows that the Mulliken charges of hydrogen atoms are typically much larger than the APT or CGenFF charges.

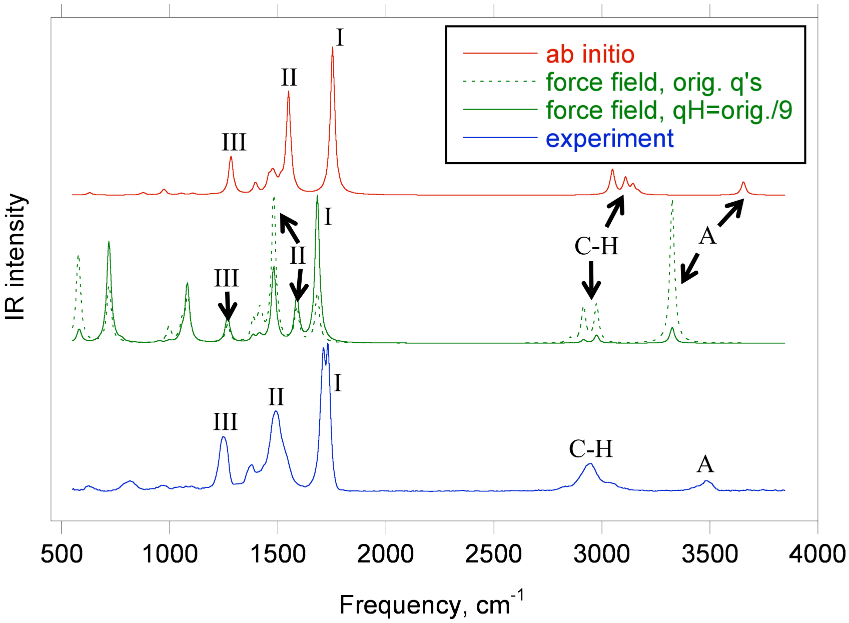

Figure 4.7: Repeat of Fig. 4.2 showing the effect on the force field based spectrum of reducing the partial charges of hydrogen atoms by a factor of 9 when calculating the IR intensities of the normal modes. For comparison, the dotted curve is the force field based spectrum from Fig. 4.2 using the original CGenFF charges (Table 4.5).

As shown in Table 4.5, the force field (CGenFF) assigns a charge of $+0.09e$ to the hydrogen atoms in N-methylacetamide's two CH_3 groups. To test the effect of reducing hydrogen charges, I reduced the methyl H charges by a factor of 9 to $+0.01e$ (with the H connected to N also reduced in charge by this same factor). This new set of charges was then used with the force field based normal modes to recalculate the IR intensities of the modes. To clarify, normal modes were *not* recalculated using the reduced H charges for Coulomb forces between atoms; the same set of normal modes obtained using the original

CGenFF charges was used with the new set of charges for the calculation of the IR intensities of the modes. The resulting IR spectrum obtained from the force field based normal modes and the reduced hydrogen charges is shown in Fig. 4.7. In switching from the original CGenFF charges (dotted curve) to reduced H charges (solid curve), the intensities of several peaks have been reduced in the force field based spectrum, resulting in better agreement with experiment and with the *ab initio* calculation. Note that the peaks that have been decreased in intensity are not just the high frequency C−H and N−H peaks; the peaks at $1481\,\mathrm{cm}^{-1}$ and $581\,\mathrm{cm}^{-1}$ have also been attenuated.

4.6 Other molecules

The remaining 18 molecules in Table 4.1 include planar six-membered rings (toluene, p-xylene, 3-methylpyridine); non-planar six-membered rings (cyclohexene, thiane); five-membered rings (2,3-dihydrofuran, thiazole); five-membered rings joined to seven- or six-membered rings (azulene, benzimidazole, benzothiazole); two six-membered rings joined together (quinoline, isoquinoline); and three rings joined together (anthracene, acridine, phenazine, phenoxazine, phenothiazine, carbazole). Five of these molecules contain non-planar rings (cyclohexene, thiane, 2,3-dihydrofuran, phenoxazine, phenothiazine), whereas the other 13 molecules contain only planar rings.

For the sake of brevity, rather than discussing all 18 of these molecules in detail, I will make some general comments and give a couple of examples. There was considerable variation among these molecules in how well the force field calculation could reproduce the experimental spectra. For example, toluene's force field based spectrum agreed quite well with experiment, whereas for 2,3-dihydrofuran the agreement was poor. Force field calculations of molecules containing planar rings tended to yield better agreement with experiment than those containing non-planar rings.

4.6.1 2,3-dihydrofuran

For example, 2,3-dihydrofuran is a non-planar five-membered ring molecule (Fig. 4.8). Its experimental and calculated (*ab initio* and force field) IR spectra are shown in Fig. 4.9.

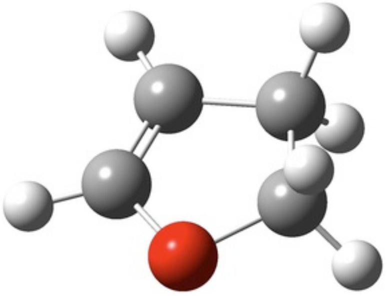

Figure 4.8: Structure of 2,3-dihydrofuran.

At first glance the force field spectrum in Fig. 4.9 does not match the *ab initio* spectrum (or experiment) either in terms of the frequencies of the peaks or their relative intensities. However, upon closer inspection of the normal modes, I was able to find correspondences between the spectra. From viewing animations of the normal modes obtained from the two calculation methods, I was able to match several of the force field based normal modes (IR peaks) to their counterparts in the *ab initio* spectrum; these matches are marked by dotted lines. The frequencies in cm^{-1} of the most prominent peaks are annotated in Fig. 4.9. Peaks marked “RD” are due to normal modes that cause ring deformations (always coupled with movements of the hydrogen atoms). The strong peak annotated with “C=C” is due to a ring deformation that is dominated by C=C stretching. The high-frequency peaks marked “CH_2” are

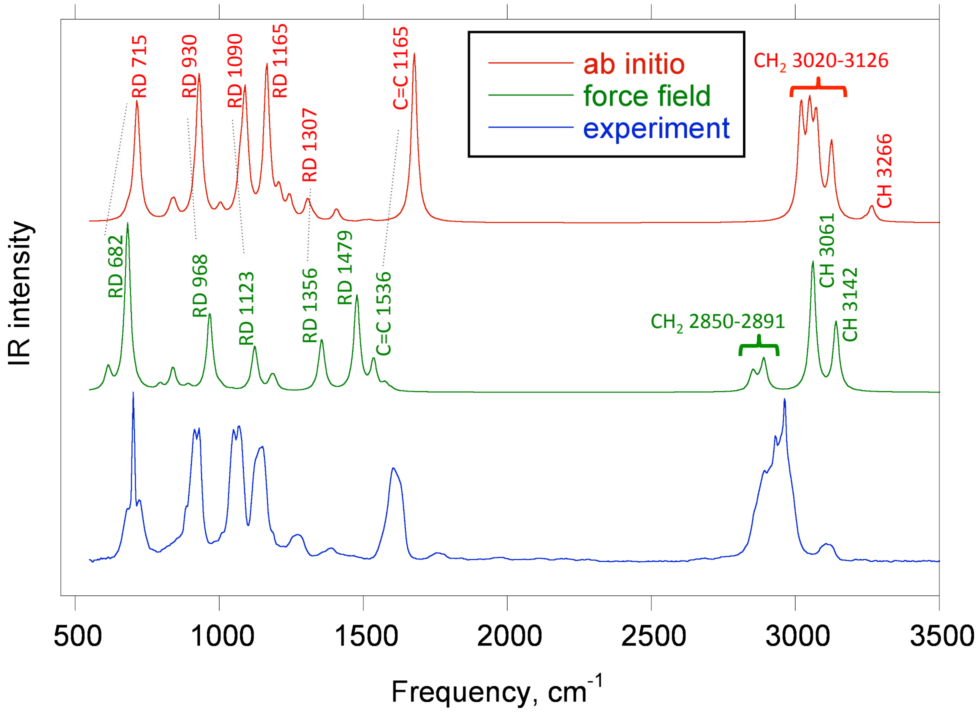

Figure 4.9: IR spectrum of 2,3-dihydrofuran: two calculated spectra compared to experimental spectrum.

due to symmetric and antisymmetric stretching of C−H bonds in the two CH_2 groups; the peaks marked "CH" are due to C−H stretching in the two CH groups.

4.6.2 Toluene

Toluene is an example of a simple molecule that contains a planar ring (Fig. 4.10). Its experimental and calculated (*ab initio* and force field) IR spectra are shown in Fig. 4.11. Among the molecules in Table 4.1, toluene had the best agreement between the force field based spectrum, the *ab initio* spectrum, and experiment.

The simplicity of toluene's structure is likely responsible for the success of the force field in this case. In Fig. 4.11, matches between IR peaks in the

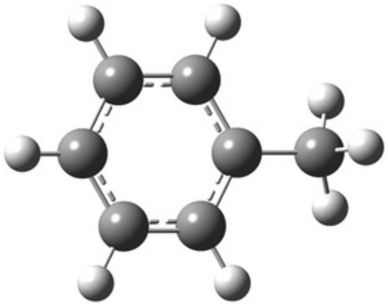

Figure 4.10: Structure of toluene.

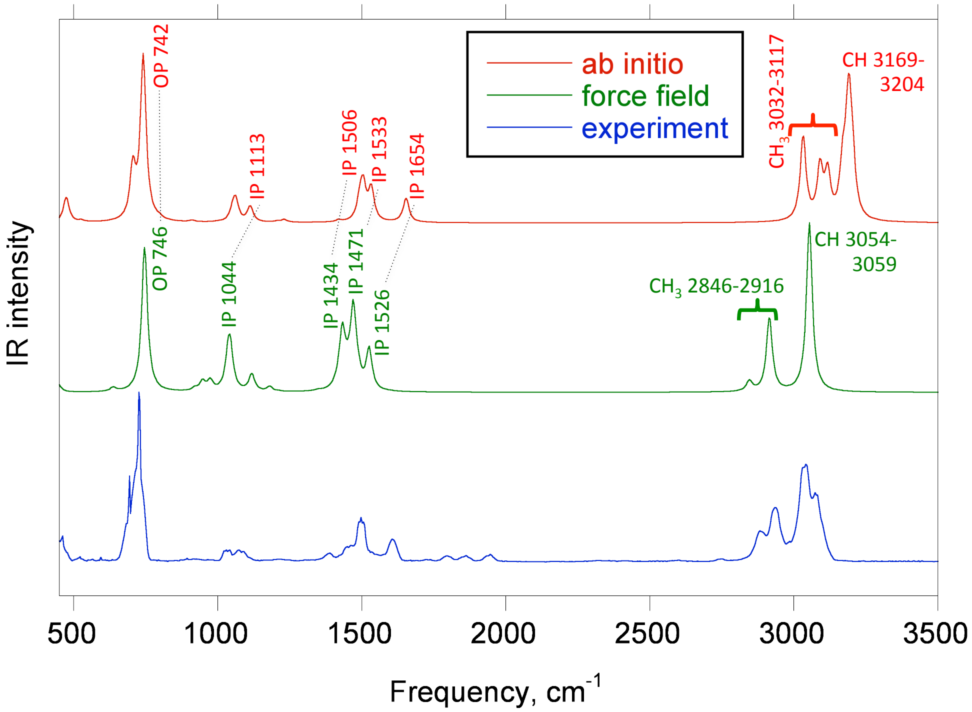

Figure 4.11: IR spectrum of toluene: two calculated spectra compared to experimental spectrum.

two calculated spectra are marked with dotted lines, with the frequencies in cm^{-1} also labeled. As with 2,3-dihydrofuran, these matches were made after comparing animations of the normal modes. The IR peak annotated with “OP” is due to an out-of-plane ring deformation; the peaks marked with “IP” are due to in-plane ring deformations. The ring-deformation modes always involve rocking or deformation of the CH_3 group as well as displacements of the hydrogens in the ring. The high-frequency peaks are due to C−H bond stretching in the CH_3 group and in the CH groups in the ring. The results for toluene and 2,3-dihydrofuran demonstrate that the experimental agreement of the force field based spectra can differ significantly from one molecule to another.

4.7 Future directions

For a given molecule, it should be possible to optimize the partial charges to obtain the best agreement between the IR spectrum calculated from partial charges and a reference spectrum. The reference or target spectrum could be either from experiment or from a quantum chemistry calculation. The root-mean-square (RMS) difference σ between the calculated and reference spectrum is

$$\sigma^2 \equiv \frac{1}{\nu_2 - \nu_1} \int_{\nu_1}^{\nu_2} \left[I_{\mathrm{FPC}}(\nu) - I_{\mathrm{REF}}(\nu) \right]^2 d\nu \ , \tag{4.5}$$

where $I_{\mathrm{FPC}}(\nu)$ is the IR spectrum calculated using fixed partial charges, $I_{\mathrm{REF}}(\nu)$ is the reference spectrum, and the frequency range of both spectra is from ν_1 to ν_2. Alternatively, the discrete set of normal mode intensities (eq. 2.52) could be compared:

$$\sigma^2 \equiv \frac{1}{3N-6} \sum_{n=7}^{3N} \left[I_{\mathrm{FPC},n} - I_{\mathrm{REF},n} \right]^2 \ , \tag{4.6}$$

where N is the number of atoms in the molecule, and the $3N-6$ vibrational modes are indexed from 7 to $3N$. Either way, starting from guess values for the partial charges, these charges could be optimized to minimize the RMS difference σ between the calculated and reference spectra. It would be interesting to see how much the force field based IR spectra could be improved by optimizing the partial charges in this systematic manner. Since the mode frequencies are determined from the force field, this charge-fitting procedure obviously could not improve the experimental agreement of the positions of the calculated IR peaks, but it could improve the experimental agreement of the calculated intensities. To improve the experimental agreement of the frequencies that are calculated from the force field would require fine tuning of the force field parameters for each molecule.

Chapter 5

IR SPECTRA OF PROTEINS FROM ALL-ATOM NORMAL MODE ANALYSIS

While there has been considerable past experimental and theoretical work done on the terahertz and infrared spectra of proteins, often researchers have focused their efforts solely on certain spectral regions. In contrast, the focus of this chapter will be on the complete spectrum of protein vibrations obtained through normal mode analysis. An attempt will be made to ascertain the spectral region where proteins' signatures differ the most from one another, allowing for the possibility of identifying unknown proteins based on comparisons to reference spectra. Before proceeding, it is useful to briefly review some of the extensive literature on the response of proteins to THz and IR radiation.

In the last decade, there has been considerable interest in the THz (far-infrared) spectra of proteins due to the development of experimental techniques that have made such measurements possible. The review by Markelz (2008) defines this frequency range to be from 1 to $100\,\mathrm{cm}^{-1}$ (from 0.03 to 3 THz), although another review by Plusquellic *et al.* (2007) extends the range up to 10 THz ($333\,\mathrm{cm}^{-1}$). Protein motions at these low frequencies tend to be global in nature rather than being localized to certain atoms or groups, as is the case with higher frequency vibrations. As such, THz spectroscopy provides a probe into the large-scale, collective dynamics of proteins, which may be important for conformational changes and function. Theoretical interpretation of the experimental spectra has been aided by normal mode analysis and molecular dynamics simulations.

At higher than THz frequencies, IR spectroscopy has been used extensively as an approximate measure of protein structure. For the large number of proteins whose detailed three dimensional structures have not yet been determined from X-ray crystallography, either because these proteins do not readily form crystals or because their structures are intrinsically disordered, IR spectroscopy is one of the analytical methods that can detect the presence of secondary structures such as α-helices and β-sheets and even provide quantitative estimates of the abundance of these structures within the protein. This line of research has benefited greatly from the development of Fourier transform infrared (FTIR) spectrometers and accompanying data-analysis techniques (Arrondo *et al.*, 1993). As introduced in Chapter 4 and shown in Fig. 5.1, the protein backbone consists of a repeating structural unit; the rest of a protein's structure is determined from its sequence of amino acid residues (the R's in the figure) bonded to the C_α atoms. The peptide bond (the amide CONH grouping) links one segment of the chain to the next and is thus a ubiquitous feature along the backbone. In IR spectroscopy, the peptide bond is associated with nine characteristic absorption bands called the amide bands. Susi (1972) gives a table of the generally accepted names for these bands along with their frequencies and the types of molecular motions that give rise to them. Similar tables are given by Bandekar (1992); Arrondo *et al.* (1993); Tamm and Tatulian (1997); Kong and Yu (2007).

Among these amide bands, the amide I band in the range 1600–1700 cm^{-1} has received by far the most attention for its sensitivity to proteins' secondary structure. Since water also has strong absorption in the amide I frequency range, often IR spectra are measured with the protein in a solution of deuterium oxide (written as D_2O or ^2H_2O) instead of ^1H_2O, in which case the

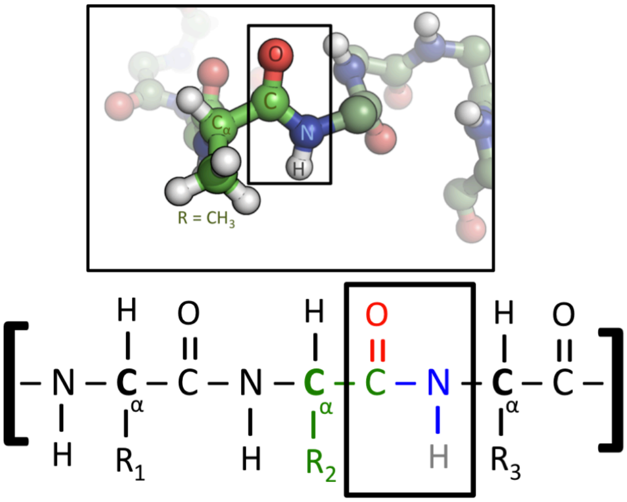

Figure 5.1: Schematic diagram of the backbone of a protein. The peptide bond (CONH group) is enclosed in a box. (This public-domain graphic was obtained from Wikipedia.)

modified IR bands of the deuterated protein are denoted by amide I′, II′, and so on. As discussed in Chapter 4, the amide I absorption band is due to C=O stretching vibrations in the peptide bonds of the protein backbone. Rather than occurring at a single frequency, the various C=O vibrations have a distribution of frequencies that is characteristic of the protein's geometry. Thus the amide I absorption band is broad and smooth, consisting of many overlapping, unresolved absorption lines. Nevertheless, the shape of this band contains information about the secondary structures present in the protein. Despite the difficulties posed by the amide I band's lack of fine structure, there has been considerable success in interpreting this band to derive quantitative estimates of the fraction of a protein's structure that is in α-helices and β-sheets. As reviewed by Hering and Haris (2009) and Barth and Zscherp (2002), there have been two main approaches to estimating secondary-structure content based on the amide I band: curve fitting and pattern recognition.

The first approach is best illustrated in the classic paper by Byler and Susi (1986). They demonstrated that after Fourier self-deconvolution (Kauppinen *et al.*, 1981), the amide I$'$ band could be decomposed into six to nine Gaussian components at 11 well-defined frequencies. Informed by empirical and theoretical knowledge of the amide I frequencies associated with different secondary structures, the Gaussian components were assigned to β-strands, α-helices, unordered segments, or "turns and bends". The main result was that the "β-content" of the protein—the fraction of the protein's sequence that is folded into β-strands—could be estimated with surprising accuracy based on the fraction of the total integrated amide I$'$ band area comprised by the Gaussian components that were associated with β-strands. The α-helix content of the protein was estimated in the same way, and these estimates were found to be accurate to within 4% of the actual β-strand or α-helix fractions for 11 proteins whose structures were already known from X-ray crystallography.

The second approach to determine fractions of secondary structure from IR spectra is pattern recognition: a category that includes both multivariate data analysis and artificial neural network methods (Hering and Haris, 2009). The basic idea is that by using a calibration set or "training set" of IR spectra of proteins whose structures are known, the α-helix and β-sheet fractions can be related to the spectra in a systematic way, which then can be used to predict the secondary-structure content for a protein outside of the calibration set. For example, Dousseau and Pezolet (1990) applied two multivariate data analysis methods, classical least-squares and partial least-squares, to a calibration set of 13 proteins in H_2O solution, making use of both the amide I and II regions of the spectra and correcting for water absorption. This resulted in structure fraction predictions of comparable accuracy to those obtained by Byler and

Susi (1986) using curve fitting. The pattern-recognition approach involves less subjectivity than the curve-fitting approach: in the former case, the spectra do not require deconvolution or second derivatives prior to analysis, and no band assignments need to be made by the researcher (Hering and Haris, 2009).

In addition to the previously mentioned reviews by Hering and Haris (2009) and Barth and Zscherp (2002), there are several other excellent reviews with an experimental emphasis (Kong and Yu, 2007; Jackson and Mantsch, 1995; Tamm and Tatulian, 1997; Bandekar, 1992; Arrondo *et al.*, 1993). In contrast, the reviews by Schweitzer-Stenner (2006), Barth and Zscherp (2002), and Krimm and Bandekar (1986) focus more on the theoretical aspects of predicting the IR spectra, particularly the amide I band. The theoretical overview by Barth and Zscherp is an excellent starting point. The theory that has been most often used to explain the amide band shapes is transition dipole coupling: "It is a resonance interaction between the oscillating dipoles of neighbouring amide groups and the coupling depends upon the relative orientations of, and the distance between, the dipoles" (Barth and Zscherp, 2002). Early efforts to apply transition dipole coupling to the amide vibrations of proteins considered idealized, infinite β-sheets or α-helices, taking advantage of symmetry to calculate the frequency shifts of the amide I and II vibrations of these structures relative to the frequencies of isolated oscillators (Miyazawa, 1960). While these calculations helped characterize the amide I bands due to different secondary structures, more realistic calculations were carried out by Torii and Tasumi (1992, 1996), who used the known crystal structures of eight proteins to locate and orient the "transition dipole" oscillator representing each peptide group in a protein and treated the interaction between the oscillators using transition dipole coupling. Their approach was quite successful in reproducing the observed amide I band for proteins, and many subsequent studies have made use of their method. In recent years more sophisticated approaches have been developed that make use of quantum chemistry calculations of molecular subunits to derive force fields appropriate for the amide vibrations along the protein backbone (Kubelka and Keiderling, 2001; Choi *et al.*, 2007; Choi and Cho, 2009; Kubelka *et al.*, 2009; Grahnen *et al.*, 2010).

While these theoretical methods have been shown to deliver high-quality predictions of the amide bands, especially amide I, the goal of this chapter is the prediction of a protein's complete vibrational spectrum—not just the vibrations of the peptide bonds of the protein backbone that give rise to the amide bands. As in previous chapters, the approach used here is normal mode analysis, although molecular dynamics can also be used to predict IR spectra. Since proteins have a large number of atoms it is only practical to calculate the Hessian matrix from force fields. The first detailed normal mode analysis of a protein was done by Brooks and Karplus (1983) using the CHARMM force field. The protein studied was bovine pancreatic trypsin inhibitor, and all heavy atoms and polar hydrogens (580 atoms in total) were considered in the model (other hydrogens were considered to be part of the heavy atom to which they were bonded). The software developed for normal mode analysis was later incorporated into the CHARMM package (Brooks *et al.*, 2009) as the VIBRAN set of commands. For this reason CHARMM is an ideal program for performing normal mode analysis of proteins.

5.1 Method

I used CHARMM (version c35b1r1) to perform normal mode analysis for the 13 proteins in Table 5.1. All calculations were performed using the Saguaro computer cluster at Arizona State University, a Linux-based cluster of over 5000 Intel Xeon processors. For the steps of minimization and calculation of IR intensities, I was able to speed up the calculations by doing them in parallel on several processors. However, I was not successful in finding a way to diagonalize the Hessian matrix using the parallel capabilities of the computer cluster and was restricted to always using a single processor for this step.

The proteins' crystal structures were obtained from the Protein Data Bank (Berman *et al.*, 2000) as PDB files and their unique four-character identifiers are given in the table. I will often refer to these PDB IDs in place of the protein's name, *e.g.*, 1BTI for bovine pancreatic trypsin inhibitor. Since X-ray crystallography cannot resolve the hydrogen atoms of proteins, the positions of the hydrogen atoms are not given in the PDB files; these positions are assigned by CHARMM based on the known positions of the other atoms and the parameter file for the force field. Several of the protein structures contained more than one chain; in all but one case (1T6B) the sequence of amino acids was identical from one chain to the next. These protein structures were dimers (2TRX, 1YPY, 1UZG, and 1OAN); trimers (2EBO and 1K4R); and one hexamer (2I39). For simplicity, when a protein had more than one chain, I only considered a single chain in my calculation, the "A" chain in the PDB file (or X chain in the case of 1T6B). The exception was 2EBO for which I did the calculations in two ways: first with the A chain by itself, and then with all three chains of the trimer. The number of atoms, including hydrogens, in the protein chain(s) considered are listed in Table 5.1. In a few cases, the tabulated number of residues in the chain may be less than the length of the sequence because several residues were missing coordinates in the PDB files.

However, this posed no problem for inputing these structures into CHARMM because the residues with missing coordinates were either at the beginning or end of the sequence. Thus there were no problematic, disconnected gaps in the structure, and the normal mode analysis and spectrum calculation could be done for the continuous portion of the sequence whose coordinates were known.

The largest protein considered was protective antigen from anthrax toxin (1T6B), and this was a special case. The two chains are very different with the X chain being 735 residues in length and the Y chain being 189. Neither one of the chains had its complete structure in the PDB file: 59 residues of the X chain were missing coordinates, 19 residues of Y. Some of the missing coordinates were in the middle of the sequence, causing a disconnected gap in the structure. To obtain a more complete structure, I submitted the protein's FASTA sequence to the ModWeb comparative modeling web server (Pieper *et al.*, 2011; Eswar *et al.*, 2003). This gave a structure with coordinates for 722 residues of the X chain (missing coordinates for only the first 13 residues), which I then used for normal mode analysis and the spectrum calculation. Since the structure obtained from this homology modeling may differ significantly from true structure, the spectrum calculated based on this structure may not be accurate. The goal here was to test the spectrum calculation method on a fairly large bio-threat-related molecule for which the computing resources required—run time, memory, disk space—were substantial.

Two other important steps to get CHARMM to work with the protein are the specification of the protonation state of the histidine residues and the specification of disulfide bonds. First, all the histidine residues (residues named HIS in the PDB file) must be renamed to one of the three types of histidine residues parametrized by the CHARMM force field. CHARMM uses the residue names HSD, HSE, and HSP to differentiate between the three protonation states. These are defined by which of the two nitrogen atoms in the imidazole ring is bonded to a hydrogen atom: one, the other, or both. I chose the HSD protonation state for all histidine residues. Second, disulfide bonds in the protein must be specified in the CHARMM input script. These bonds between the sulfur atoms of nearby cysteine residues are usually specified in the PDB file itself in lines beginning with "SSBOND". For example, residues number 6 and 127 of lysozyme are cysteine residues that are close to each other in the three-dimensional, folded structure and they are connected by a disulfide bond. This bond is specified in the CHARMM input script with the command, "patch disu A 6 A 127". While the specification of the disulfide bonds and the protonation state of the histidine residues can be done manually, these steps are more easily accomplished automatically by using the CHARMM-GUI web-based application to generate the necessary inputs (Jo *et al.*, 2008).

Table 5.1: Thirteen proteins whose IR spectra were calculated.

PDB ID	Name	Atoms	Residues	Chain	Out of
1L2Y	Trp-cage miniprotein construct TC5b	304	20	A	1 of 1 chain
1BTI	Bovine pancreatic trypsin inhibitor	882	58	A	1 of 1 chain
2EBO	Envelope glycoprotein GP2 from Ebola virus	1203	74	A	1 of 3 chains
		1203 × 3	74 × 3	A, B, & C	3 of 3 chains
2TRX	Thioredoxin from E. coli	1653	108	A	1 of 2 chains
2I39	N1L protein from vaccinia virus	1943	117	A	1 of 6 chains
6LYZ	Hen egg-white lysozyme	1961	129	A	1 of 1 chain
1YMB	Horse heart metmyoglobin	2411	153	A	1 of 1 chain
1YPY	L1 protein from vaccinia virus	2698	182	A	1 of 2 chains
3KGQ	Carboxypeptidase A	4729	303	A	1 of 1 chain
1K4R	Envelope glycoprotein E from dengue virus	6012	395	A	1 of 3 chains
1UZG	Envelope glycoprotein E from dengue virus type 3	6050	392	A	1 of 2 chains
1OAN	Envelope glycoprotein E from dengue virus type 2	6129	394	A	1 of 2 chains
1T6B	Protective antigen from anthrax toxin	11 352	722	X	1 of 2 chains

The potential energy of the proteins was calculated using the CHARMM22 force field (MacKerell *et al.*, 1998, 2004). Besides the terms in this potential for bonded and non-bonded interactions shown in eqs. 4.1 and 4.2, there is an additional set of terms for the energy-correction map (CMAP). These terms, introduced by MacKerell *et al.* (2004), are a function of the dihedral angles and their use "corrects certain small systematic errors in the description of the protein backbone by the all-atom CHARMM force field" and "significantly improves the structural and dynamic results obtained with MD simulations of proteins in crystalline and solution environments" (Brooks *et al.*, 2009).

5.1.1 Minimization

Prior to calculation of the Hessian matrix and normal modes, the structure needs to be optimized to minimize the potential energy, as described in Chapter 2. In order to prevent too much distortion of the structure during minimization, harmonic constraints of the form $K_i\,(x_i - x_{i,\mathrm{orig}})^2$ were temporarily added to the potential energy for each atom i except hydrogens, which had the effect of attracting each non-hydrogen atom to its original position. The strengths of the harmonic constraints K_i were taken to be proportional to each atom's mass. At first, the minimization was done with stiff constraints, $K_i/m_i = 10^4\,\mathrm{kcal\ mol^{-1}\ \AA^{-2}}$ (with m_i in atomic masses), which hardly allowed for any movement of the atoms except for the hydrogens. In successive iterations, the strength of the constraints was lowered by a factor of ten to $10^3, 10^2, \ldots, 10^{-4}$ kcal mol^{-1} Å^{-2}. For each of the powers of 10, minimization was done with 2000 steps of the steepest descent method followed by 5000 steps of the adopted basis Newton-Raphson method. After this, the harmonic restraints were completely removed, and an additional 5000 steepest descent steps and 10 000 adopted basis Newton-Raphson steps were taken to arrive at the final minimized structure. This rigorous minimization procedure became quite time consuming as the size of the protein increased. The minimization can be sped up by running CHARMM in parallel on many processors (I used up to 32 processors for the larger proteins). However, the vibrational analysis commands of CHARMM have not been implemented to work in parallel (at least not in the version of CHARMM I used), so there is no speedup from using more than one processor to diagonalize the mass-weighted Hessian matrix using CHARMM.

5.1.2 Diagonalization

CHARMM has the ability to calculate the second derivatives in the Hessian matrix (eq. 2.7) analytically, or numerically based on finite differences. I always used the default option to calculate the second derivatives analytically. The structure of the matrix is illustrated in Fig. 5.2. It is a sparse, symmetric matrix with most of its nonzero elements located near the diagonal; these values are mostly due to bonded interactions between atoms. The nonzero elements that are far from the diagonal are due to non-bonded interactions: electrostatic forces and van der Waals forces. The pattern of non-bonded interactions away from the diagonal is a signature of the three-dimensional, folded structure of the protein, which places residues close together in space that are far apart in the sequence. The same pattern would likely emerge if instead of plotting the Hessian matrix in Fig. 5.2, one plotted a “contact map” with dark pixels indicating that the atom pair corresponding to that row and column are separated by a distance less than a threshold of a few Å. The extent of the non-bonded interactions depend on the cutoff distances. I used the default cutoffs specified in the parameter file for the CHARMM22 force field. These default cutoffs ignored any electrostatic or van der Waals interactions between atoms separated by more than 12 Å.

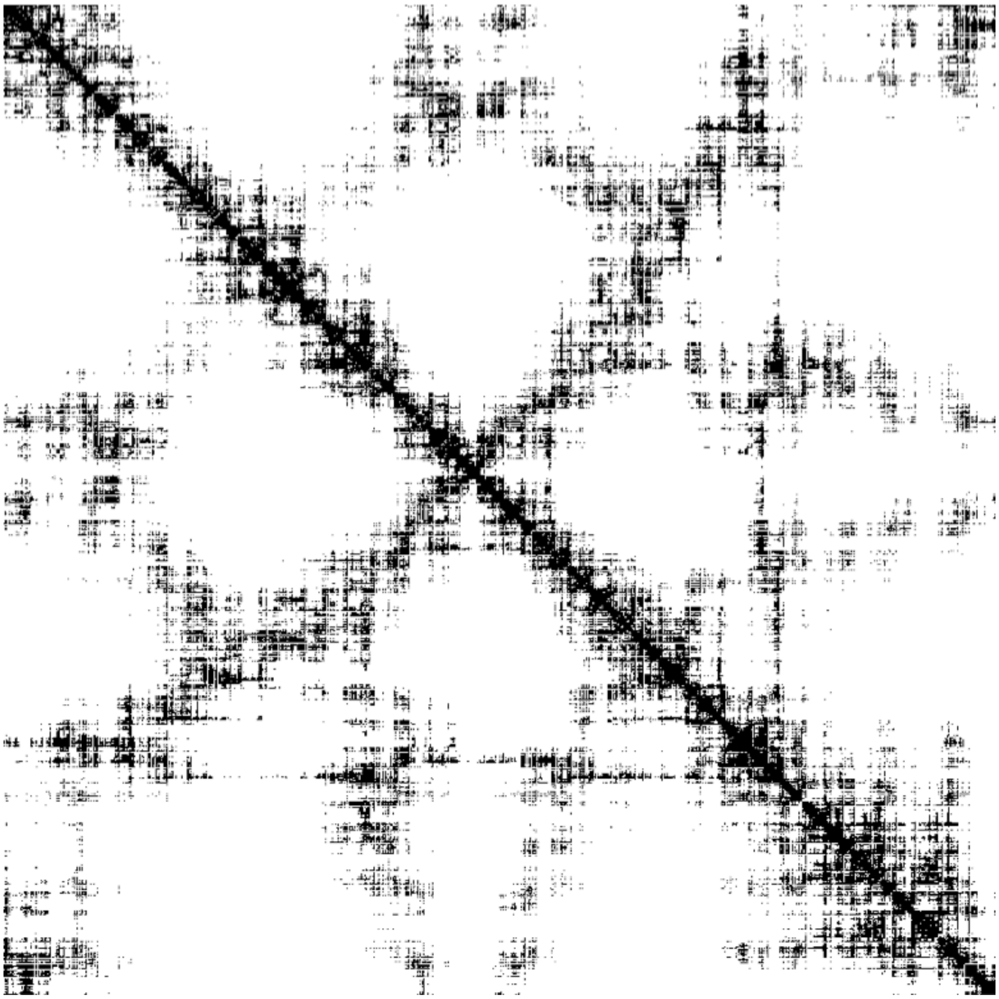

Figure 5.2: Illustration of the structure of a typical Hessian matrix, with matrix elements above a certain threshold in absolute value represented as black pixels. Row 1, column 1 of the matrix is in the upper-left corner of the image, and the rows and columns are enumerated as in eq. 2.7.

I explored two options for diagonalizing the mass-weighted Hessian matrix of eqs. 2.6 and 2.7: either the diagonalization was performed using the DIAG command in CHARMM, or the matrix was saved to disk for diagonalization using LAPACK (Anderson *et al.*, 1999), which is available in Intel's Math

Kernel Library for use with Intel's Fortran compiler. Care was taken to save the Hessian to disk in binary format to retain machine precision rather rather than saving the data to text files with reduced precision; the binary data files also took up less disk space than text files. The LAPACK subroutine for diagonalizing a symmetric matrix of double precision floating point values is called DSYEV; for single precision, the subroutine to use is SSYEV.

I found that the time needed to diagonalize the Hessians of proteins smaller than 6000 atoms was comparable for all three matrix-diagonalization methods (CHARMM's DIAG and LAPACK's DSYEV and SSYEV), but for the proteins larger than 6000 atoms CHARMM's DIAG routine typically took twice as long to finish the calculation compared to the time spent by either of the two LAPACK routines, whose times were still comparable to each other. For example, 1UZG's 6050 atoms results in a Hessian matrix of size 18 150 × 18 150; to diagonalize this matrix, CHARMM's DIAG spent 7.3 hours, whereas DSYEV took 2.8 hours and SSYEV took 3.4 hours.

These times include disk access times, which can be significant. Running the program to diagonalize the largest matrix, that of the X chain of 1T6B with 11 352 atoms, using DSYEV took 21.2 hours, of which time 2.1 hours was spent reading the Hessian from disk and 3.0 hours was spent saving the resulting eigenvectors to disk. The 34 056 × 34 056 double-precision values (8 bytes each) of the 34 056 eigenvectors require 9.3 Gbytes of memory. The full Hessian matrix requires this same amount of memory, although in saving it to disk one can halve the size by storing only half of the symmetric matrix; further reductions in file size can be achieved by only storing the nonzero matrix elements. For example, 89% of the Hessian matrix elements for bovine pancreatic trypsin inhibitor were exactly zero, and the distribution of the nonzero elements is shown in Fig. 5.3. The memory requirements can be halved again if single-precision values (4 bytes each) are used instead of doubles (8 bytes each). There was some noticeable loss of precision if the diagonalization was done in single precision (SSYEV) instead of double (DSYEV): in the former case the first six modes, which are expected to have zero frequency, strayed a few cm^{-1} from zero. However, this didn't make much difference in the final calculated spectrum.

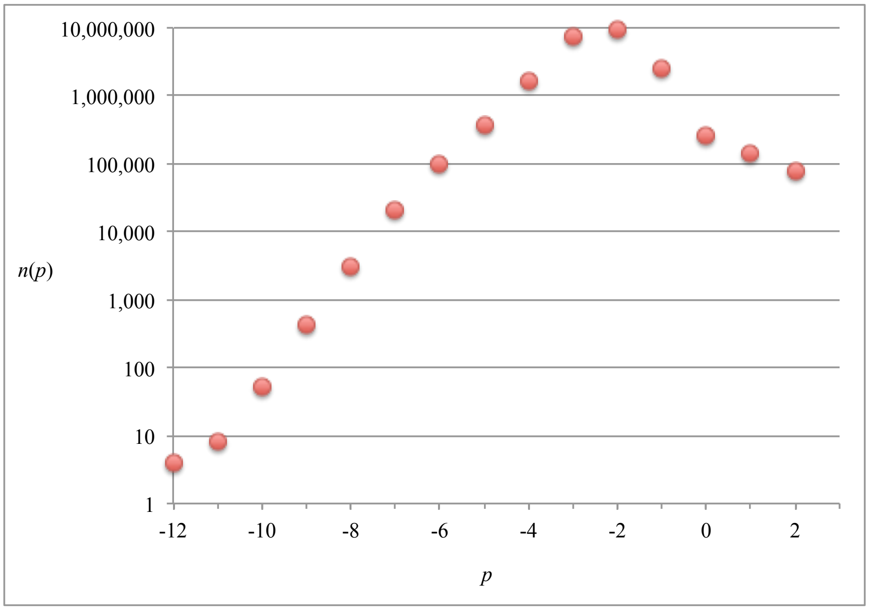

Figure 5.3: Distribution of the magnitude of nonzero elements of the non-mass-weighted Hessian matrix of bovine pancreatic trypsin inhibitor (1BTI). Each matrix element's magnitude is calculated as $p = \text{floor}\,(\log_{10}|H_{i,j}|)$ with $H_{i,j}$ in kcal Å^{-2}. Then $n(p)$ is the number of matrix elements having a given value of p.

For the larger proteins, one could avoid some of the burdensome disk access times by never saving the eigenvectors to disk, only working with them in memory and saving only the final mode frequencies and IR intensities at the end. To really speed things up, what is needed is a diagonalization routine that can be run in parallel on many processors. ScaLAPACK (Choi *et al.*,

1996) implements parallel versions of the DSYEV and SSYEV subroutines, named PDSYEV and PSSYEV; however, the file input and output is made much more complex by the need to distribute the matrix data across many computers and then gather the results after diagonalization. To extend the IR spectrum calculation using all-atom normal mode analysis to proteins larger than those in Table 5.1, parallelization of the diagonalization step would be required.

5.1.3 Spectrum Calculation

After diagonalization to obtain the normal mode eigenvalues and eigenvectors, the IR intensity I_n associated with each normal mode n was calculated according to eq. 2.52. The masses m_i and effective partial charges q_i of the atoms were obtained from the PSF (protein structure file) produced by CHARMM. As discussed in Chapter 2, the normal mode frequencies ω_n are found from the square roots of the eigenvalues, but conversion of units is required. The eigenvalues have the same units as the elements of the mass-weighted Hessian matrix (eq. 2.7), which are kcal g^{-1} Å^{-2}. After converting to SI units (kcal to J, g to kg, and Å to m), the eigenvalue is an angular frequency squared in rad s^{-2}. After taking its square root and dividing by 2π rad/cycle, it is a frequency in Hz. Divide by 10^{12} to get it in THz, and use eq. 1.2 to get the frequency in cm^{-1}. From start to finish, if one simply takes the square root of an eigenvalue of the mass-weighted Hessian matrix in its original units of kcal g^{-1} Å^{-2}, then by multiplying the answer by the conversion factor

$$\frac{10^3\sqrt{4.184}}{2\pi(2.997\,924\,58)} \tag{5.1}$$

one obtains that mode's frequency in cm^{-1}.

The calculation of the IR intensities is easily parallelized, as each I_n in eq. 2.52 is independent of the others. I used the OpenMP parallelization scheme in my Fortran program to split the loop over mode number n among eight processors on Saguaro sharing the same memory. That is, rather than a single processor sequentially calculating the intensities of all $3N-6$ vibrational modes, each of the eight processors was assigned 1/8 of the modes to work on, which sped up the calculation considerably. This leaves diagonalization as the only step in my spectrum calculation process that I haven't yet been able to parallelize, and this limits the size of the protein that can be dealt with.

After calculating the normal mode frequencies ω_n and IR intensities I_n, the IR absorption spectrum was simulated by broadening with a line shape function as described by eq. 2.37. For this purpose, I wrote a Fortran program called CONVOLVE, which takes as its input the list of normal mode frequencies and associated intensities, and gives as its output the convolution described in eq. 2.37. The CONVOLVE program allows the user to select from one of three line shapes: the $S_n(\omega)$ of eq. 2.39, the Lorentzian function $L_n(\omega)$ of eq. 2.45, or a Gaussian function. To calculate the IR spectra shown in this chapter, I used a Lorentzian function. The same peak width (FWHM) of $\Gamma = 10\,\mathrm{cm}^{-1}$ was used for all normal modes. As a final step, each spectrum was normalized to have a maximum value of unity. Fig. 5.4 illustrates the effect of using four different peak widths for the calculation of a protein's IR spectrum. With $\Gamma = 20\,\mathrm{cm}^{-1}$, the peak just above $3300\,\mathrm{cm}^{-1}$ appears as a single peak; with a narrow peak width ($\Gamma = 5\,\mathrm{cm}^{-1}$) it is resolved into two peaks.

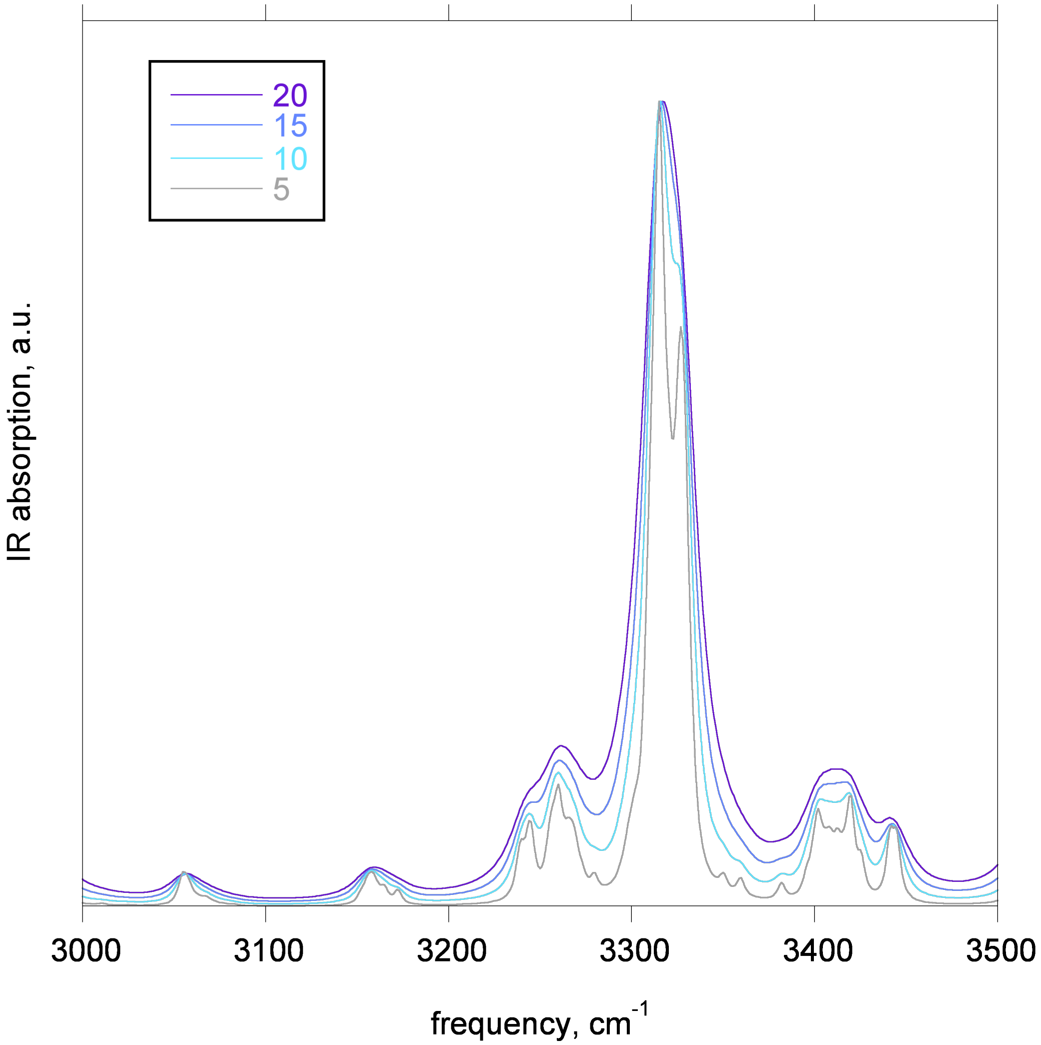

Figure 5.4: Illustration of the broadening of absorption lines in the IR spectrum due to using different values for the FWHM of the assumed Lorentzian line shape: $\Gamma = 20, 15, 10,$ and $5\,\mathrm{cm}^{-1}$.

5.1.4 Calculation Times

Comparisons of the computer run times required for the matrix diagonalizations of the 13 proteins of Table 5.1 are shown in Fig. 5.5. For these time comparisons the diagonalization was done using the DSYEV routine on a single processor of the Saguaro cluster. The times shown do not include the time spent reading or writing data files. These times can be thought of as the time required to calculate an IR spectrum from normal mode analysis since the most time-consuming step in this process is the diagonalization of the mass-weighted Hessian matrix to get the normal mode frequencies and eigenvectors. As shown in Fig. 5.5, the time required to diagonalize the matrix scales as N^3 (with N being the number of atoms), whereas in Chapter 2 it was discussed that the time required for the calculation of the IR intensities scales as only N^2.

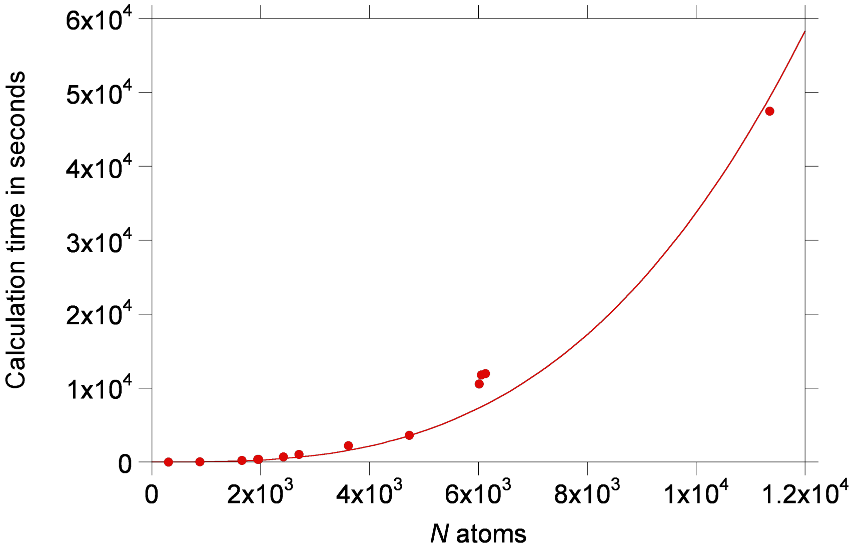

Figure 5.5: Computer run times needed to diagonalize the Hessian matrix for the proteins in Table 5.1. The diagonalization used the DSYEV routine of LAPACK. These times were measured on a single processor of ASU's Saguaro computer cluster; disk access times were excluded. The fitting function is $y = (3.37 \pm 0.14) \times 10^{-8} \times x^3$.

Fig. 5.7 shows calculated IR spectra for three of the proteins of Table 5.1. The entire frequency range of vibrational modes is plotted, from the terahertz region ($\lesssim 100\,\mathrm{cm}^{-1}$) to the highest frequency vibrations at $\approx 3700\,\mathrm{cm}^{-1}$. The protein spectra show a gap from 1800 to $2800\,\mathrm{cm}^{-1}$ in which there are no vibrational modes.

5.2.1 Low-frequency region including THz range

The lowest frequency modes are global in nature; *i.e.*, rather than the motions being localized in specific parts of the protein, these modes involve the flexing of the complete structure and collective motions of entire secondary structures relative to each other.

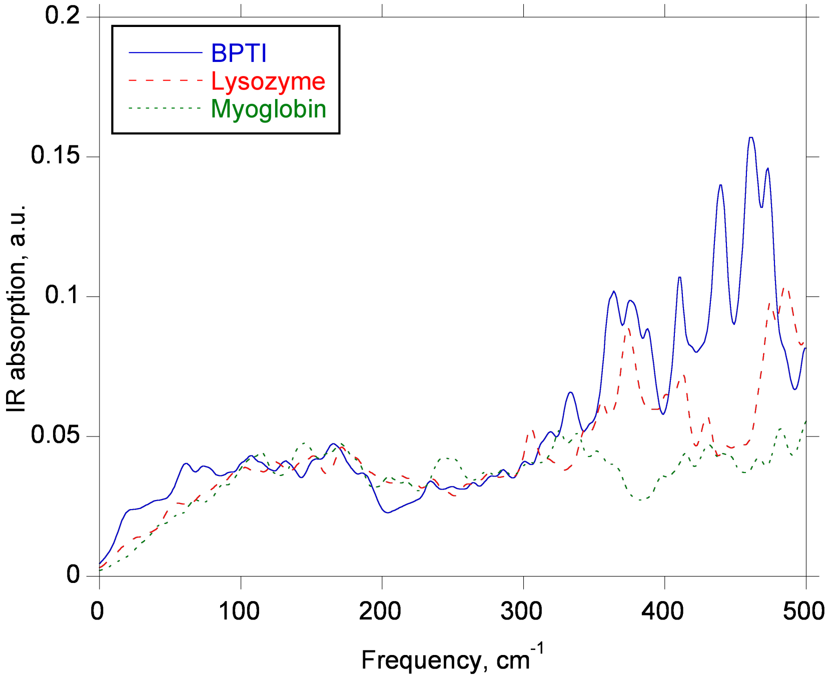

Figure 5.6: Low-frequency region (including THz frequencies $\lesssim 100\,\mathrm{cm}^{-1}$) of the calculated IR spectra of three proteins: bovine pancreatic trypsin inhibitor (1BTI), lysozyme (6LYZ), and myoglobin (1YMB).

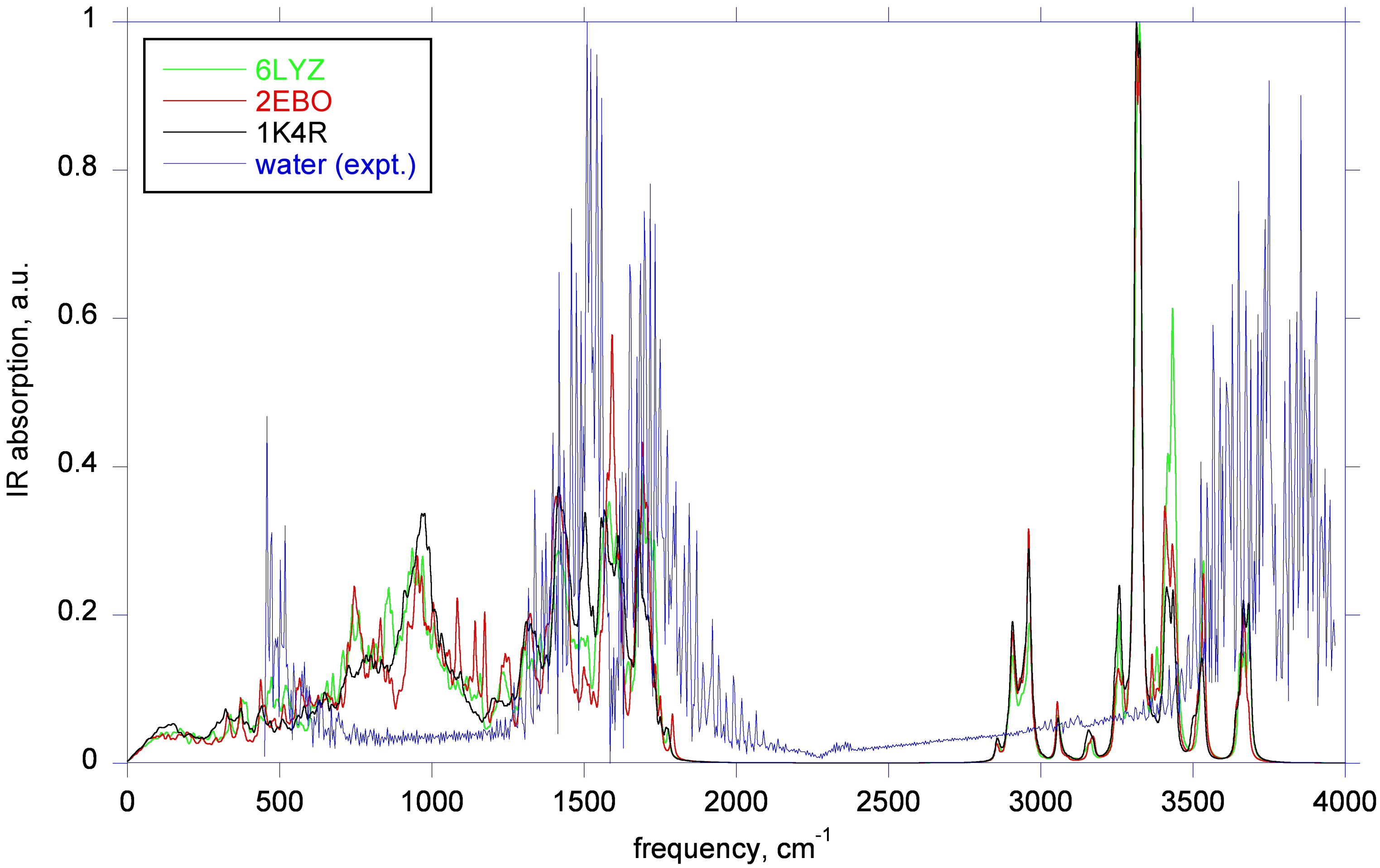

Figure 5.7: Calculated IR absorption spectra of three of the proteins from Table 5.1 using Lorentzian line shapes with FWHM = 10 cm^{-1}. Also shown is an experimental spectrum of water from the NIST/EPA Gas-Phase Infrared Database (Stein, 1992) to illustrate the spectral regions with strong absorption by water.

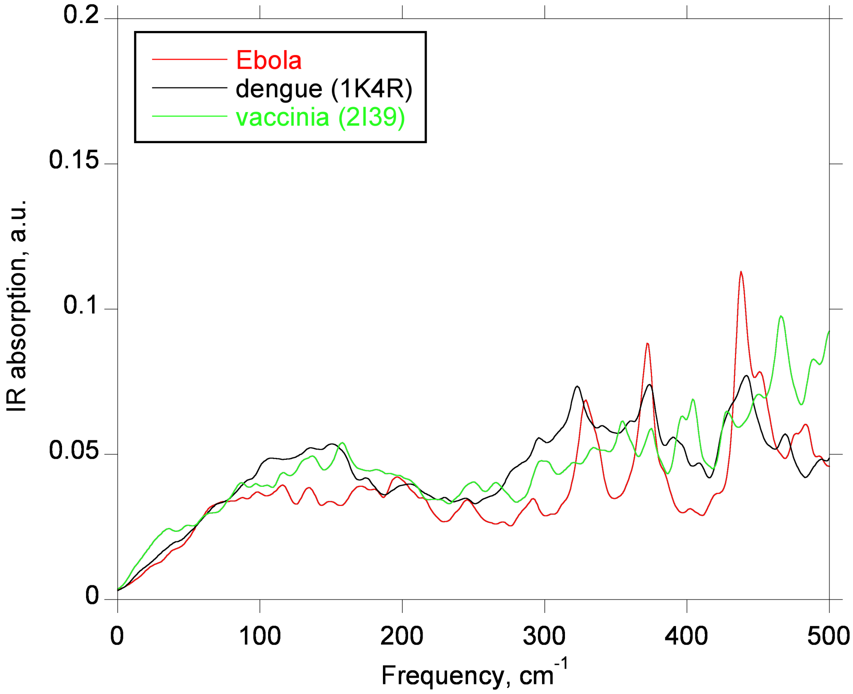

Figure 5.8: Low-frequency region (including THz frequencies $\lesssim 100\,\mathrm{cm}^{-1}$) of the calculated IR spectra of proteins associated with Ebola (2EBO), dengue (1K4R), and vaccinia (2I39) virus.

As shown in Figs. 5.6 and 5.8, the spectra of various proteins in the THz region ($\lesssim 100\,\mathrm{cm}^{-1}$) are very similar to one another, which makes it impossible to identify proteins based on their THz spectra. Above $300\,\mathrm{cm}^{-1}$ the different proteins' spectra begin to differ from one another.

5.2.2 Intermediate region—window of low absorption by water

Fig. 5.7 suggests that the spectral region from about 600 to $1300\,\mathrm{cm}^{-1}$ may be promising for discriminating between proteins in water solution since in this window there is comparatively little absorption by water. Fig. 5.9 shows this spectral region for three of the proteins in Table 5.1—bovine pancreatic trypsin inhibitor, lysozyme, and myoglobin. These three proteins have somewhat different spectral signatures in this region, so differentiating between them based on their spectra might be possible. This same spectral region is shown in Fig. 5.10 for three proteins associated with viruses—Ebola (2EBO), dengue (1K4R), and vaccinia (2I39).

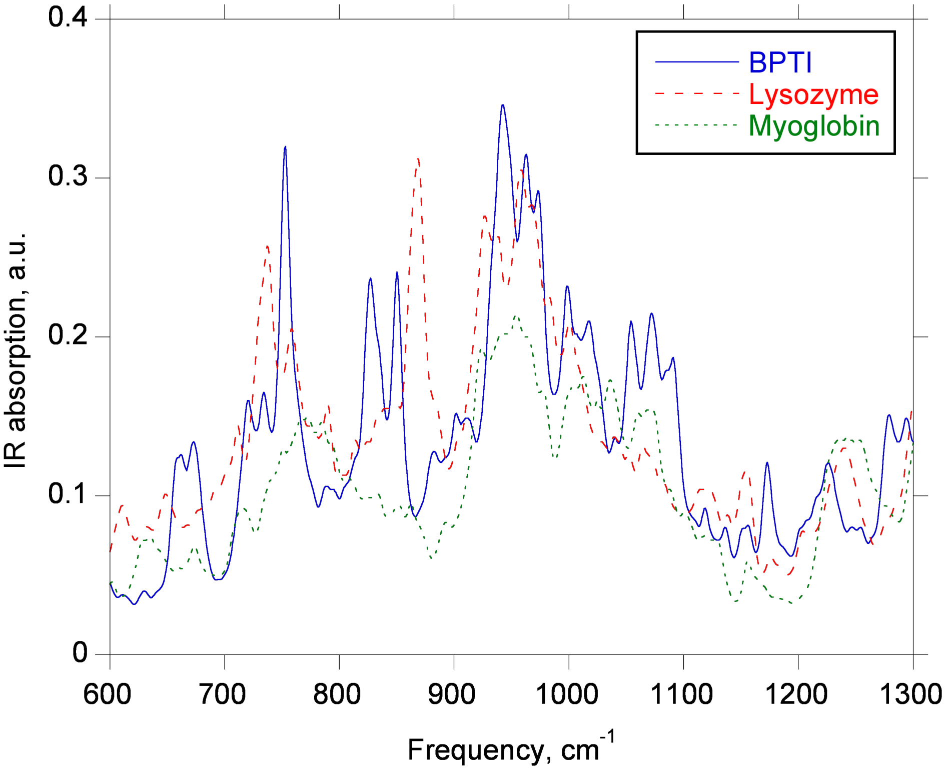

Figure 5.9: Calculated IR spectra for three proteins—bovine pancreatic trypsin inhibitor (1BTI), lysozyme (6LYZ), and myoglobin (1YMB)—in the intermediate spectral region where IR absorption by water is expected to be low.

IR spectra were calculated for three different versions of envelope glycoprotein E from dengue virus: PDB IDs 1K4R, 1UZG, and 1OAN. Not surprisingly, these three very similar structures result in similar calculated IR spectra, as shown in Fig. 5.11. However, the two proteins studied that were associated with vaccinia virus—N1L protein (2I39) and L1 protein (1YPY)—were quite different in size and structure; hence their calculated IR spectra are quite different, as shown in Fig. 5.12.

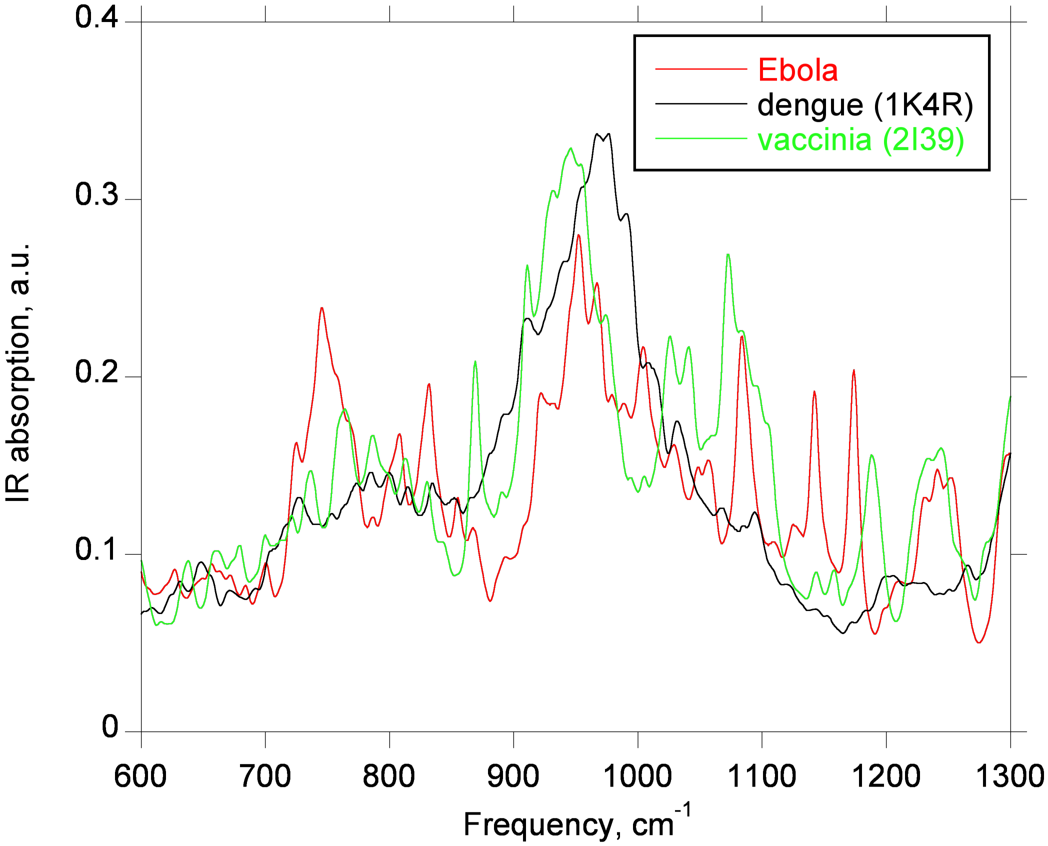

Figure 5.10: Calculated IR spectra for three proteins associated with viruses—Ebola (2EBO), dengue (1K4R), and vaccinia (2I39)—in the intermediate spectral region.

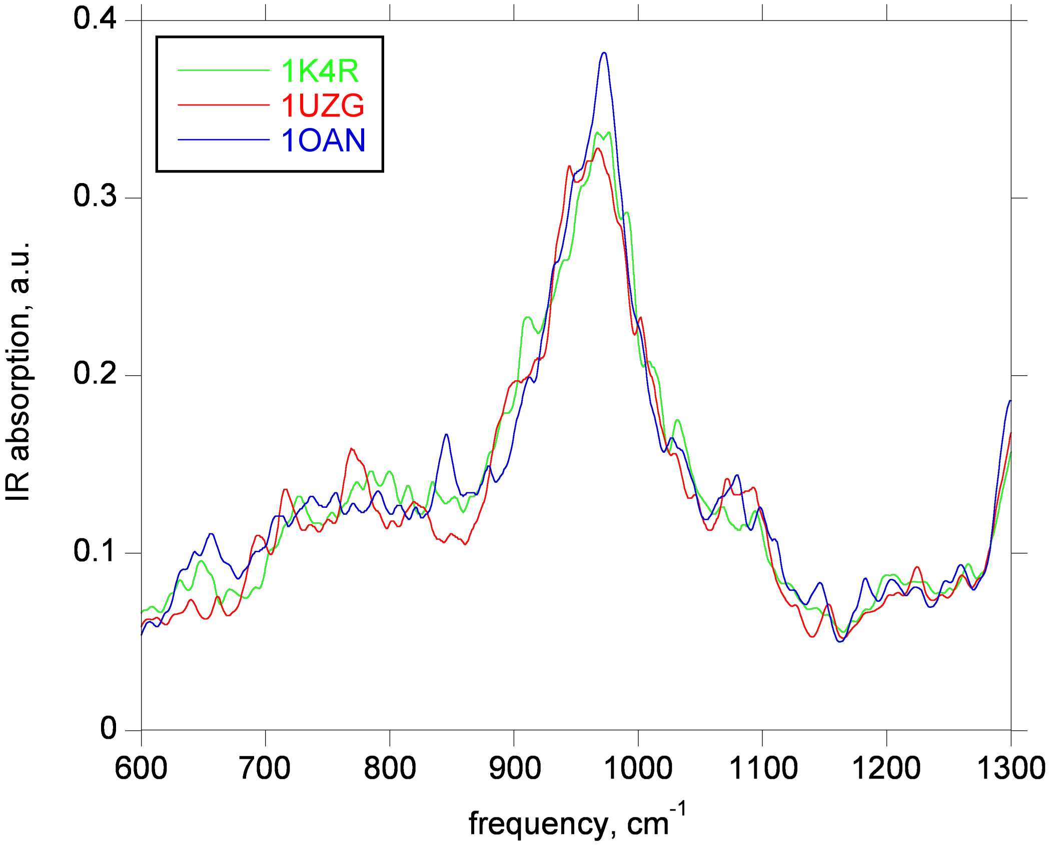

Figure 5.11: Calculated IR spectra for three versions of envelope glycoprotein E from dengue virus in the intermediate spectral region.

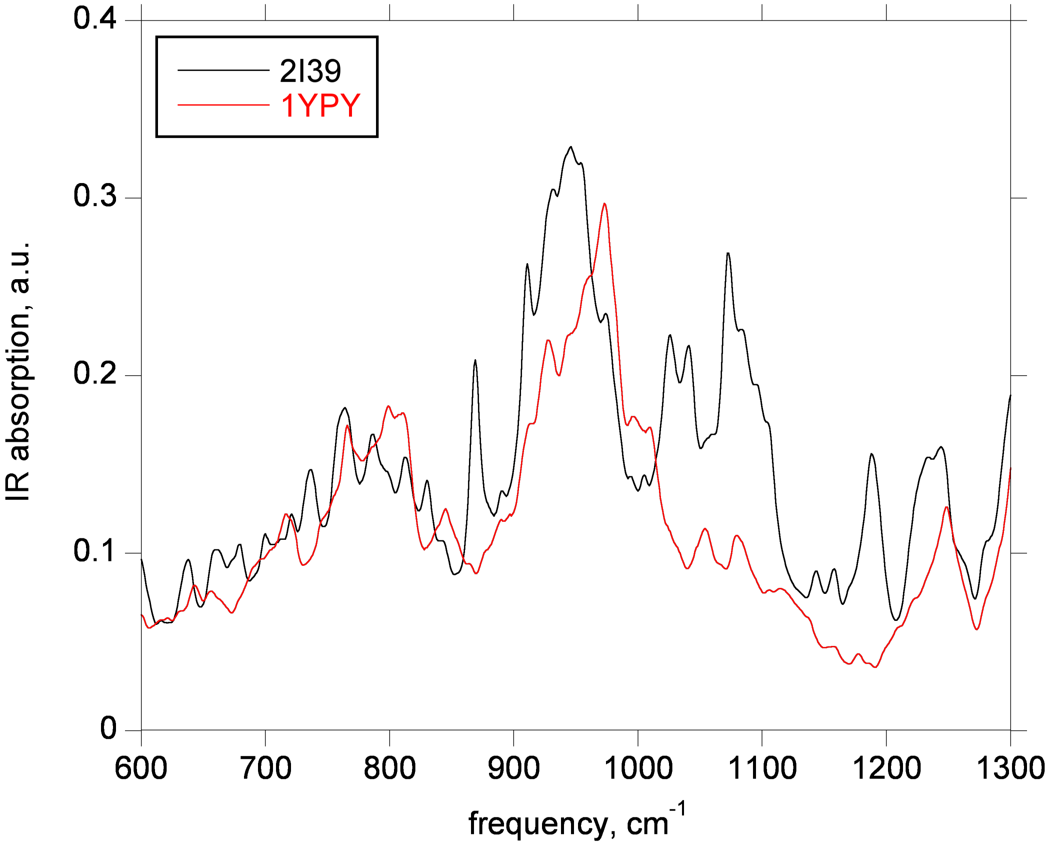

Figure 5.12: Calculated IR spectra for two proteins associated with vaccinia virus—N1L protein (2I39) and L1 protein (1YPY)—in the intermediate spectral region.

5.2.3 High-frequency region

The strong absorption bands on the high-frequency side of the gap are due to bond-stretching vibrations of hydrogen atoms bonded to heavier atoms (carbon, nitrogen, or oxygen). The high-frequency region (2800–3800 cm^{-1}) of the calculated IR spectrum of bovine pancreatic trypsin inhibitor is shown in Fig. 5.13. This is compared to the density of vibrational modes: a histogram of the normal mode frequencies in bins of width 10 cm^{-1}. The type of atomic displacements responsible for each IR peak is labeled in the figure. These peak assignments were made after examining the few most IR-active normal modes near the frequency of each peak. To examine the modes, individual modes were animated by generating trajectories of the atomic positions over one cycle of oscillation of a specific mode. I wrote a Fortran program that allows the user to select a specific normal mode and generate a coordinate trajectory in the DCD file format, which can then be viewed as a movie using the program VMD (Humphrey *et al.*, 1996). CHARMM already offers this feature, but the advantage of my program is that it allows the user to select the maximum atomic displacement (*e.g.*, 1 Å) in the mode animation, whereas CHARMM uses a temperature argument to scale the atomic displacements.

As labeled in Fig. 5.13, the peaks in the high-frequency region are due to bond stretching in various C−H bonds (2858–3056 cm^{-1}), N−H bonds (3154–3537 cm^{-1}), and O−H bonds (3683 cm^{-1}). In principle, there should also be a peak due to stretching of the S−H bonds in cystine (Cys) residues (this protein has six Cys residues), but these do not appear in the calculated spectrum because each Cys residue is paired with another Cys residue by a disulfide bond. When CHARMM patches two Cys residues together, it modifies the

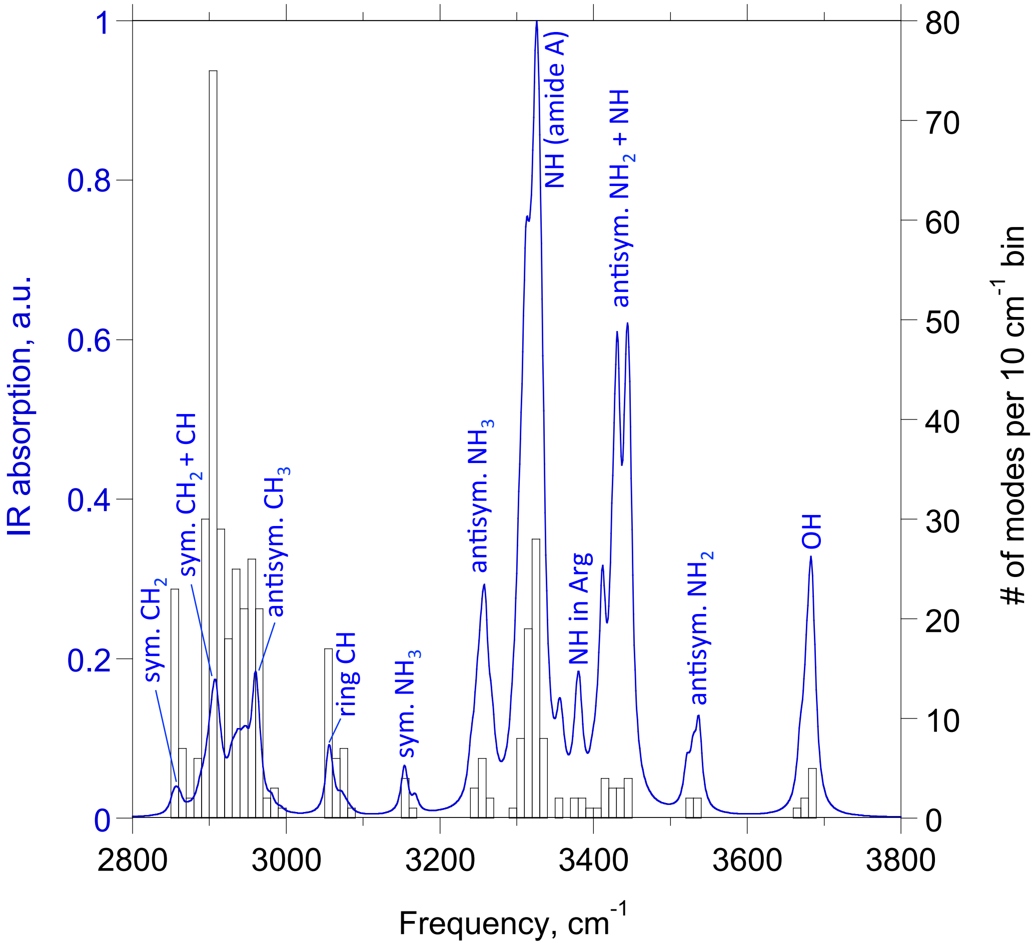

Figure 5.13: High-frequency region (2800–3800 cm^{-1}) of the calculated IR spectrum (blue curve) of bovine pancreatic trypsin inhibitor compared to the density of vibrational modes (histogram) in bins of width 10 cm^{-1}. The IR peaks in this region are due to X−H stretching modes with X = C, N, or O. The labels give the type of X−H stretching mode that is responsible for each peak.

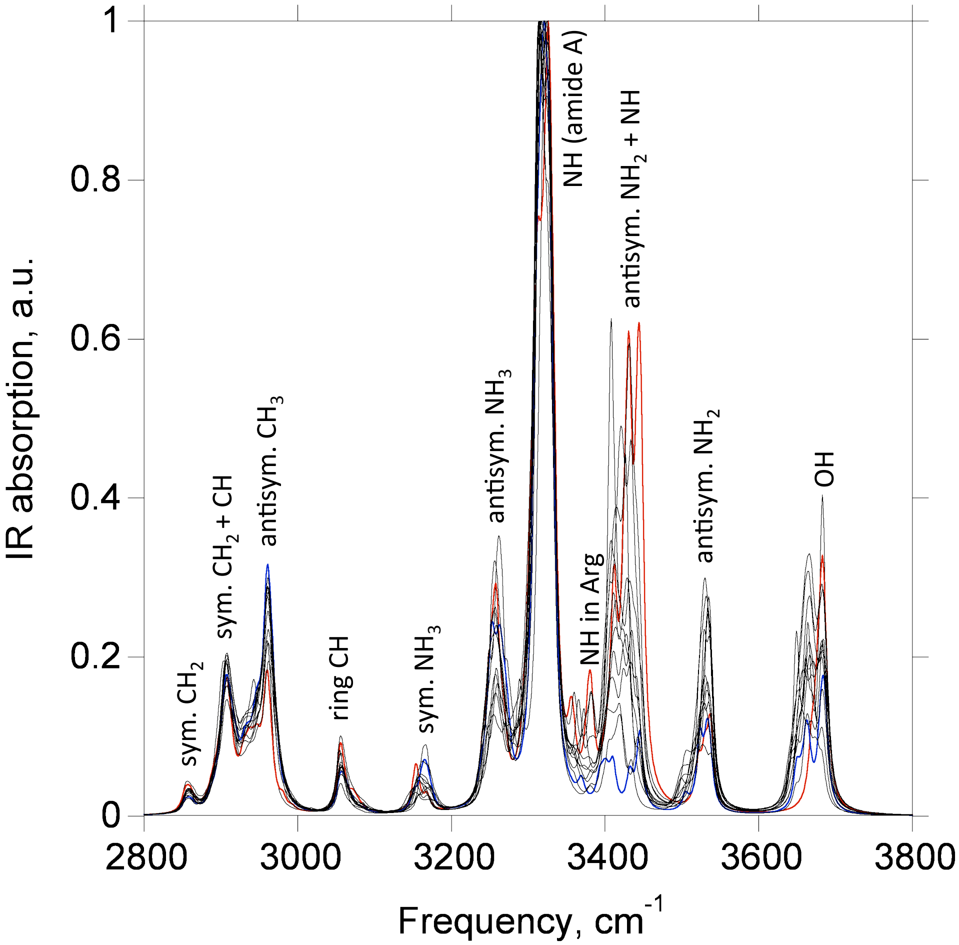

Figure 5.14: High-frequency region (2800–3800 cm^{-1}) of the calculated IR spectra of the 13 different proteins in Table 5.1. Peak labels transferred from the assignments made for bovine pancreatic trypsin inhibitor (1BTI) in Fig. 5.13. The protein with the lowest abundance of arginine residues (thioredoxin, 2TRX) is in blue, while the one with the highest Arg abundance (1BTI) is in red.

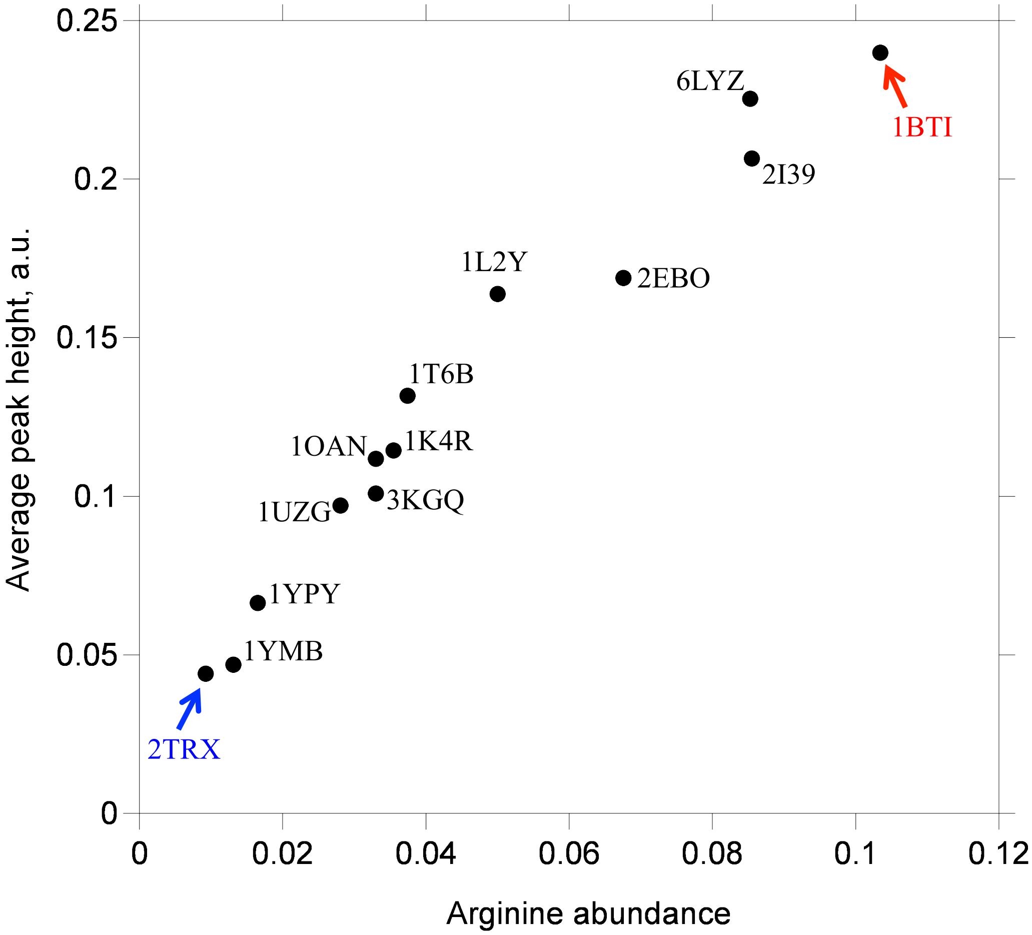

Figure 5.15: For each of the 13 proteins in Table 5.1, the average height of the peak at $\approx$ 3440 cm^{-1} in the calculated IR spectrum (Fig. 5.14) is based on the integrated area over 3390–3487 cm^{-1}. This is plotted against the proteins' arginine abundance: the ratio of the number of Arg residues to the total number of residues.

model structure by adding an S$-$S bond between the two Cys residues and eliminating the hydrogen atom that was originally bonded to each sulfur atom.

In Fig. 5.13 the peak with the largest number of normal modes is at 2908 cm^{-1} and is due to symmetric stretching of CH_2 groups combined with stretching of an adjacent C$-$H bond; these modes were localized in individual lysine, arginine, and methionine residues. The peak with the strongest IR intensity is at 3326 cm^{-1} and is due to N$-$H bond stretching in the peptide bond of the protein backbone. As discussed in Chapter 4, this is called the amide A band. At 3431–3444 cm^{-1}, the next strongest peak (appearing as a double peak for this particular protein) is due to atomic displacements localized in arginine (Arg) residues: asymmetric stretching in arginine's two NH_2 groups coupled with stretching of the adjacent N$-$H bond (not the backbone, amide N$-$H). Typically the motion was localized in a single Arg residue for a given mode, with different modes vibrating in different Arg residues (this protein has six Arg residues). The various N$-$H stretches are more strongly weighted than C$-$H stretches in the calculated spectrum because of the different partial charges assigned to the hydrogen atoms. The hydrogen charge is $+0.31e$ in the backbone NH, $+0.46e$ in NH_2 groups, and $+0.33e$ in NH_3 groups, whereas the hydrogen charge is only $+0.09e$ in CH_2 and CH_3. This explains why the N$-$H peak at 3326 cm^{-1} has much stronger intensity, even though it is a sum of fewer normal modes, than the C$-$H peak at 2908 cm^{-1}.

When the calculated IR spectra of all 13 proteins from Table 5.1 are compared in the high-frequency region (Fig. 5.14), they are found to be quite similar. The relative intensities of the various peaks are comparable from one protein to the next, which is an indication that the relative numbers of the different chemical groups (CH, CH_2, CH_3, NH, NH_2, NH_3, and OH) giving rise to these peaks are similar from one protein to the next. The exception to this rule is the peak at $\approx 3430\,\mathrm{cm}^{-1}$ whose intensity varies drastically among the proteins. As previously discussed, in bovine pancreatic trypsin inhibitor the normal modes responsible for this IR peak were found to be localized to arginine residues. Thus I expected the height of this peak to depend on the relative abundance of Arg in the protein sequence. As demonstrated by Fig. 5.15, this was indeed the case. The average intensity of the peak at $\approx 3430\,\mathrm{cm}^{-1}$ was obtained from numerical integration (using Simpson's rule) over over 3390–3487 cm^{-1} to get the area of the peak. The protein with the highest Arg abundance, bovine pancreatic trypsin inhibitor (6 out of 58 residues are Arg), also had the highest average intensity of the peak at $\approx 3430\,\mathrm{cm}^{-1}$. The protein with the lowest Arg abundance, thioredoxin (1 out of 108 residues are Arg), had the lowest average peak intensity.

5.2.4 Spectrum of monomer compared to trimer

As discussed earlier, calculations were done in two ways for envelope glycoprotein GP2 from Ebola virus: first, in its monomer form (chain A only, 1203 atoms), and then in its trimer form (chains A, B, and C; 3609 atoms). The calculated IR spectra for the monomer and trimer are compared in Fig. 5.16. There are some noticeable differences in the spectra below about $1200\,\mathrm{cm}^{-1}$, although it seems unlikely that such differences could be measured experimentally given that experimental spectra would likely be significantly smoother (broader intrinsic line widths) than these calculated spectra.

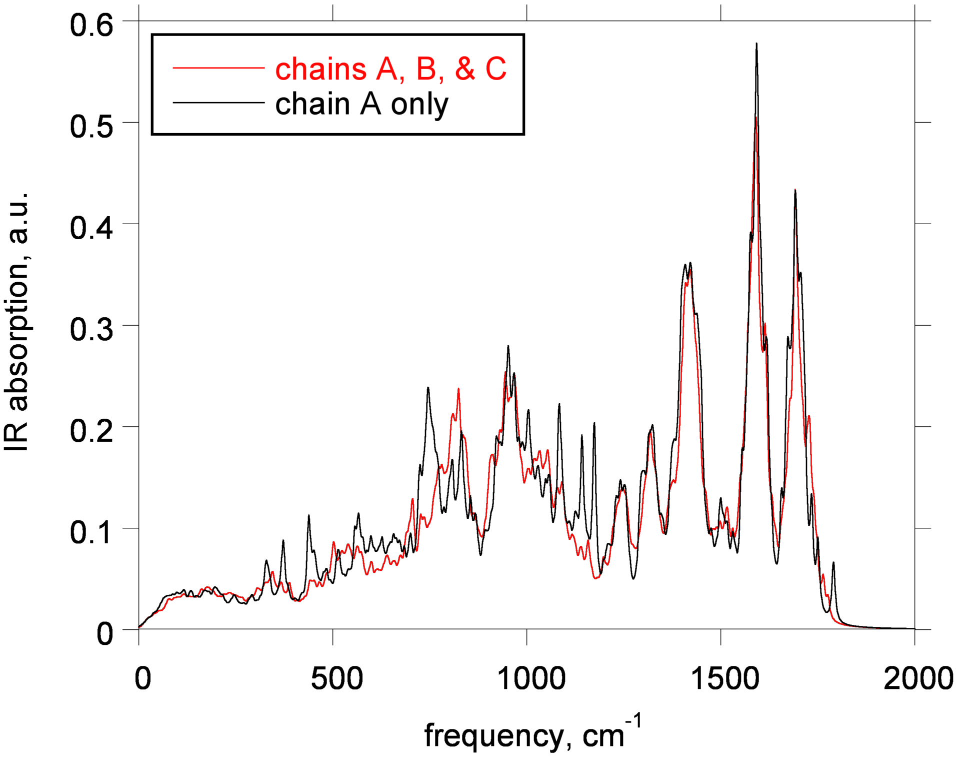

Figure 5.16: Comparison of the calculated IR spectrum of a single chain, the A chain (1203 atoms), of envelope glycoprotein GP2 from Ebola virus to that of the complete trimer consisting of chains A, B, and C (3609 atoms). Both spectra were calculated using the same Lorentzian line shape with FWHM = $10\,\mathrm{cm}^{-1}$.

5.2.5 Amide I and II bands—comparison with experiment

An experimental IR spectrum of hen egg-white lysozyme from 1200 to 2000 cm^{-1} was downloaded from the website of the Protein Infrared Database (Dong and Caughey, 1994) and compared with my calculated spectrum in Fig. 5.17. The experimental spectrum shows both the amide I (1600–1700 cm^{-1}) and amide II (1500–1600 cm^{-1}) bands; this spectrum was measured with the protein in aqueous solution (not D_2O) at a pH of 7.3. Lysozyme spectra have also been measured in aqueous solution by Pérez and Griebenow (2000). Spectra of dried lysozyme were measured by Liltorp and Maréchal (2005) and Belton and Gil (1994). In Fig. 5.17, the calculated amide I and II bands are shifted by $\approx 40\,\mathrm{cm}^{-1}$ to higher frequency compared to the experimentally observed bands. The amide I peak was measured at 1655 cm^{-1}, but the calculation placed it at 1694 cm^{-1}. Similarly, the amide II peak was measured at 1543 cm^{-1}, but the calculation placed it at 1583 cm^{-1}.

The curve-fitting method of Byler and Susi (1986) assumes that the amide I (or I′) band can be decomposed as a sum of six to nine Gaussians whose center frequencies are the characteristic vibration frequencies of the C=O modes associated with different secondary structures. While this simple picture has been successfully used for prediction of secondary-structure content, there are far more than nine normal mode frequencies within the amide I band. The distribution (histogram) of normal mode frequencies of lysozyme in bins of width 10 cm^{-1} is shown in the bottom panel of Fig. 5.17. If one considers the calculated amide I band to go from 1670 to 1740 cm^{-1} (approximately the interval corresponding to the band's full width at half maximum), then there are 174 normal modes in this band. Similarly, if one counts modes in the

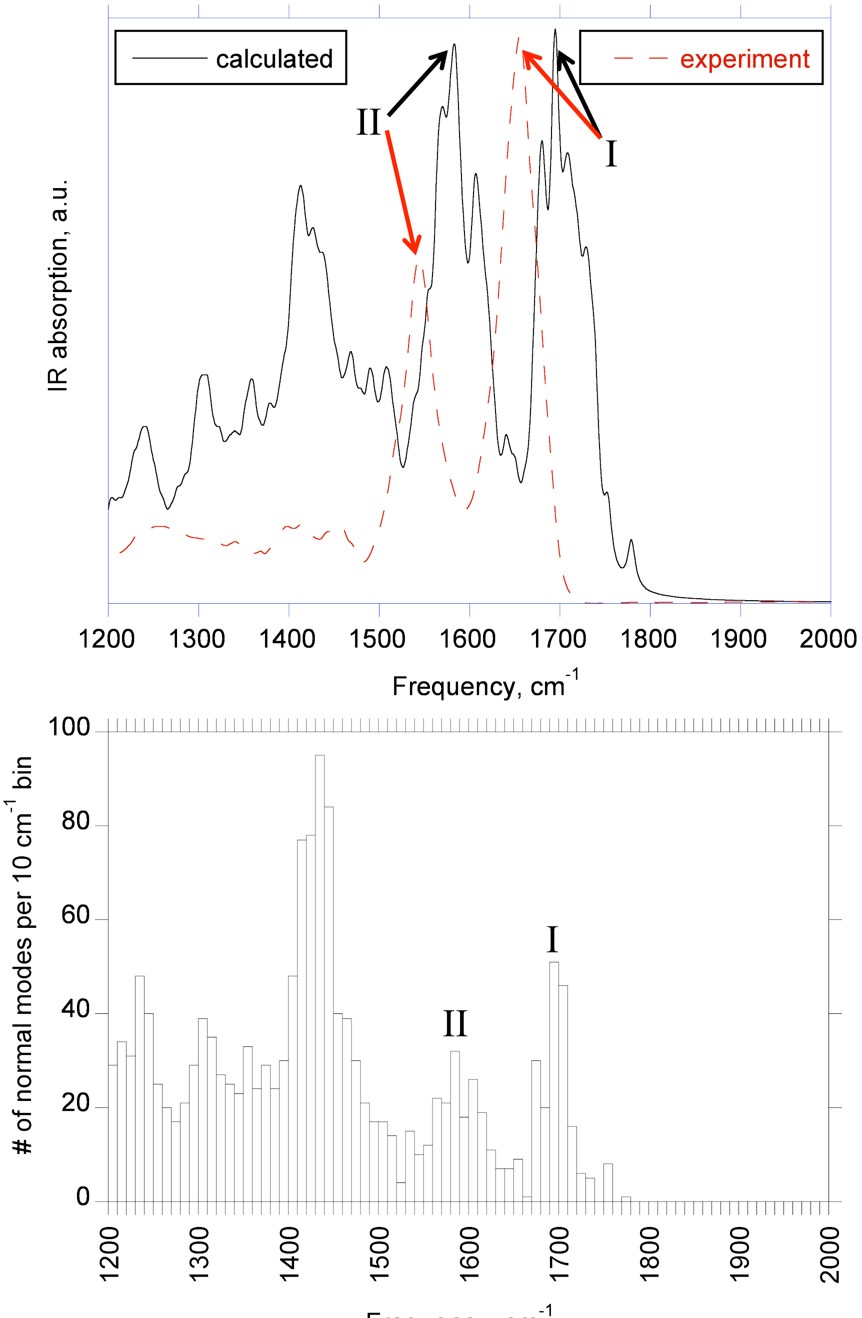

Figure 5.17: Top: Comparison of calculated IR spectrum of hen egg-white lysozyme with experimental spectrum from 1200 to 2000 cm^{-1}, showing the amide I and II bands. Bottom: histogram of normal mode frequencies in bins of width 10 cm^{-1}.

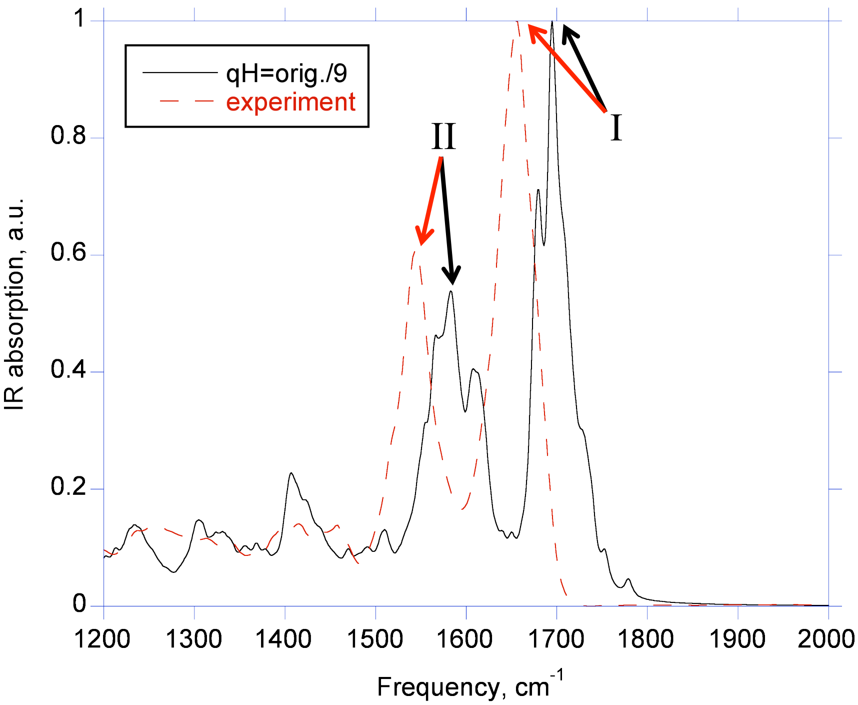

Figure 5.18: Repeat of Fig. 5.17 with recalculated IR spectrum using hydrogen charges that were reduced by a factor of 9.

amide II band in the range 1540–1630 cm^{-1}, there are 171 normal modes.

Furthermore, when one views animations of the most IR-active normal modes near the centers of the amide I and II bands, one observes that these normal modes often involve vibrations of side chains in addition to the expected amide I and II vibrations in the protein backbone. For example, the animation of a mode near the center of the amide I band of lysozyme showed the expected C=O stretching vibration in a leucine residue, but also included in this mode was a vibration involving stretching and bending of several C−N bonds in an arginine residue.

In Chapter 4 it was seen that hydrogen atoms, being the lightest atoms, tend to have the largest displacements in molecular vibrations. These large displacements, combined with the fixed partial charges assigned to the hydrogen atoms by the force field, exaggerate the contribution of hydrogens' motions to changes in the molecule's dipole moment. For many normal modes this results in calculated IR intensities that are too high compared to experimental spectra. The H atoms' contribution to the calculated spectrum can be reduced by simply reducing the effective charges of the H atoms when calculating the IR intensities in eq. 2.52. The charges assigned to the H atoms of lysozyme range from $+0.05e$ to $+0.46e$, with the most common charge being $+0.09e$. In reducing the H charges, it is preferable to scale them all down by the same factor rather than assigning an equal charge to all H atoms since the latter would ignore the different charges given to different H atoms based on their location in the molecule. As a test of whether reducing H charges can improve the agreement of the calculated spectrum with experiment, I recalculated the IR intensities for lysozyme using H charges that were ⅑ times their original values. The result is shown in Fig. 5.18. Notice that now the relative heights of the amide I and II bands in the calculated spectrum are in much better agreement with experiment, as is the portion of the spectrum below $\approx 1525\,\mathrm{cm}^{-1}$. This demonstrates that reducing the effective charges of hydrogen atoms when calculating the IR intensities of the normal modes can significantly improve the agreement of the calculated spectrum with experiment.

In addition to reducing all the H charges by the same factor, I also tried setting all H charges to $+0.01e$ or even setting them all to zero. In each case the calculated spectrum below $2000\,\mathrm{cm}^{-1}$ remained essentially unchanged from the result obtained from dividing all the H charges by 9. Of course, reducing H charges had the effect of attenuating the high-frequency X$-$H stretch region (2800–$3800\,\mathrm{cm}^{-1}$). When the H charges were divided by 9, the intensities of the peaks in this region were uniformly attenuated by a factor of ≈ 9. When instead all H charges were set to $0.01e$, there was a nonuniform attenuation of the high-frequency peaks. With all H charges set to zero, these peaks were attenuated even more, but were still present since even in X−H stretching vibrations the heavy atoms undergo small displacements.

5.3 Summary

I have demonstrated that calculating IR spectra based on all-atom normal mode analysis is practical for proteins having up to $\sim 11\,000$ atoms using the current level of computing power available to users of the Saguaro computer cluster at Arizona State University. As computer power improves, spectrum calculations for ever larger proteins will become practical. The most time consuming step is the diagonalization of the mass-weighted Hessian matrix to obtain the normal modes. As of yet, I have not been able to run a matrix diagonalization routine in parallel, so I have not been able to take advantage of the parallel processing capabilities of the computer cluster for the diagonalization step. Parallel processing of the matrix diagonalization will be necessary if one is to extend this method to larger proteins than presented here.

This study could be improved by including the effects of water in the normal mode analysis. This could be done using either explicit or implicit solvation methods. It would be interesting to see how solvation affects the resulting spectra. Modeling the proteins in solution is desirable for comparisons with experimental IR spectra.

Identifying proteins based solely on their IR spectra would be challenging if not practically impossible. The calculated spectra shown here for isolated proteins have narrower intrinsic line widths than are likely to be experimentally observed. Even with these narrow line widths, the calculated spectra are generally quite similar, and to highlight the differences one needs to “zoom in” on certain spectral regions. Certainly, proteins’ signatures in the THz region ($\lesssim 100\,\mathrm{cm}^{-1}$) are too generic to allow for much, if any, differentiation. The same is true for the high-frequency region, 2800–3700 cm^{-1}, which is due to bond stretching vibrations of hydrogens bonded to heavier atoms. The calculated spectra suggest that the intensity of one of the high-frequency peaks is quite sensitive to the abundance of arginine in the protein; this prediction needs to be tested experimentally. The intermediate spectral region from 600 to 1300 cm^{-1} is somewhat promising for differentiating between IR signatures of proteins, as IR absorption by water is low in this window and the calculated spectra show some differences here.

The prominent amide I band at 1600–1700 cm^{-1} is known to be sensitive to secondary-structure content. Analysis of this band has been used to measure the fraction of a protein that is folded in α-helices or β-sheets to within a few-percent accuracy, but this at best will narrow the number of possible matches between the spectrum of an unknown protein and a library of reference spectra, not allow for unique identification.

In comparing the calculated IR spectrum of lysozyme to an experimental spectrum in the region containing the amide I and II bands, it was seen that the agreement with experiment was significantly improved when the effective partial charges of hydrogen atoms were reduced in the calculation of the IR intensities of the normal modes. This suggests that with an adequate choice of partial charges, normal mode analysis of proteins can yield calculated IR spectra that are in reasonable agreement with experimental spectra even for the amide I and II bands.

Chapter 6

CONCLUSIONS

As the number of atoms in a molecule increases, so too does the number of vibrational modes; the infrared absorption lines of the various modes begin to overlap, leading to bands of unresolved absorption lines. For small molecules such as the organophosphorus nerve agents and simulants of Chapter 3, the IR spectrum contains many discrete, non-overlapping lines, which makes IR spectroscopy a good tool for identifying these molecules. However, for larger molecules the IR spectrum will tend to contain fewer distinguishing features as the lines blend together. This is especially true for proteins because they contain many copies of the same structural units—the peptide bonds of the protein backbone and the amino acid residues. Each structural unit has its own characteristic modes of vibration, but its frequencies may be perturbed due to effects of its local environment, such as its location in an α-helix or β-sheet, or the amount of water present. The shape of the protein spectrum reflects to some degree the protein's secondary structure and population of amino acids in the sequence, but my calculations suggest that overall, the IR spectra of proteins are quite similar from one protein to another. Considering that in a detection scenario, when one is presented not with a pure, isolated substance but with a mixture of unknown substances—various chemical, viral, or bacterial components—the prospect of identifying the substances from IR spectroscopy becomes yet more difficult. Hence IR spectroscopy is likely to remain an important tool for analyzing substances, but its utility for unique identification of complex molecules is limited.

REFERENCES

Alecu, I. M., J. J. Zheng, Y. Zhao and D. G. Truhlar, "Computational thermochemistry: Scale factor databases and scale factors for vibrational frequencies obtained from electronic model chemistries", Journal of Chemical Theory and Computation **6**, 9, 2872–2887 (2010).

Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney and D. Sorensen, *LAPACK Users' Guide* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999), third edn.

Arrondo, J. L. R., A. Muga, J. Castresana and F. M. Goni, "Quantitative studies of the structure of proteins in solution by Fourier-transform infrared spectroscopy", Progress in Biophysics & Molecular Biology **59**, 1, 23–56 (1993).

Ault, B. S., A. Balboa, D. Tevault and M. Hurley, "Matrix isolation infrared spectroscopic and theoretical study of the interaction of water with dimethyl methylphosphonate", Journal of Physical Chemistry A **108**, 46, 10094–10098 (2004).

Bahar, I., T. R. Lezon, A. Bakan and I. H. Shrivastava, "Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins", Chemical Reviews **110**, 3, 1463–1497 (2010).

Bandekar, J., "Amide modes and protein conformation", Biochimica Et Biophysica Acta **1120**, 2, 123–143 (1992).

Barth, A. and C. Zscherp, "What vibrations tell us about proteins", Quarterly Reviews of Biophysics **35**, 4, 369–430 (2002).

Belton, P. S. and A. M. Gil, "IR and Raman spectroscopic studies of the interaction of trehalose with hen egg-white lysozyme", Biopolymers **34**, 7, 957–961 (1994).

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, "The protein data bank", Nucleic Acids Res. **28**, 1, 235–242 (2000).

Bermudez, V. M., "Computational study of the adsorption of trichlorophosphate, dimethyl methylphosphonate, and Sarin on amorphous SiO_2", Journal of Physical Chemistry C **111**, 26, 9314–9323 (2007a).

Bermudez, V. M., "Quantum-chemical study of the adsorption of DMMP and Sarin on γ-Al_2O_3", Journal of Physical Chemistry C **111**, 9, 3719–3728 (2007b).

Bermudez, V. M., “Ab initio study of the interaction of dimethyl methylphosphonate with rutile (110) and anatase (101) TiO_2 surfaces”, Journal of Physical Chemistry C **114**, 7, 3063–3074 (2010).

Bolton, E. E., Y. Wang, P. A. Thiessen and S. H. Bryant, “PubChem: Integrated platform of small molecules and biological activities”, in “Annual Reports in Computational Chemistry”, vol. 4, chap. 12 (American Chemical Society, Washington, DC, 2008).

Brooks, B. and M. Karplus, “Harmonic dynamics of proteins: Normal modes and fluctuations in bovine pancreatic trypsin inhibitor”, Proceedings of the National Academy of Sciences of the United States of America **80**, 21, 6571–6575 (1983).

Brooks, B. R., C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, “CHARMM: The biomolecular simulation program”, Journal of Computational Chemistry **30**, 10, 1545–1614 (2009).

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, “CHARMM: A program for macromolecular energy, minimization, and dynamics calculations”, Journal of Computational Chemistry **4**, 2, 187–217 (1983).

Byler, D. M. and H. Susi, “Examination of the secondary structure of proteins by deconvolved FTIR spectra”, Biopolymers **25**, 3, 469–487 (1986).

Choi, J., J. Demmel, I. Dhillon, J. Dongarra, S. Ostrouchov, A. Petitet, K. Stanley, D. Walker and R. C. Whaley, “ScaLAPACK: A portable linear algebra library for distributed memory computers—design issues and performance”, Computer Physics Communications **97**, 1-2, 1–15 (1996).

Choi, J.-H. and M. Cho, “Computational linear and nonlinear IR spectroscopy of amide I vibrations in proteins”, in “Biological and Biomedical Infrared Spectroscopy”, edited by A. Barth and P. I. Haris, vol. 2 of *Advances in Biomedical Spectroscopy*, pp. 224–260 (IOS Press, Amsterdam, 2009).

Choi, J. H., H. Lee, K. K. Lee, S. Hahn and M. Cho, “Computational spectroscopy of ubiquitin: Comparison between theory and experiments”, Journal of Chemical Physics **126**, 4, 14 (2007).

Cioslowski, J., “A new population analysis based on atomic polar tensors”, Journal of the American Chemical Society **111**, 22, 8333–8336 (1989).

Creasy, W. R., J. R. Stuff, B. Williams, K. Morrissey, J. Mays, R. Duevel and H. D. Durst, "Identification of chemical-weapons-related compounds in decontamination solutions and other matrices by multiple chromatographic techniques", Journal of Chromatography A **774**, 1-2, 253–263 (1997).

Dong, A. C. and W. S. Caughey, "Infrared methods for study of hemoglobin reactions and structures", Hemoglobins, Pt C **232**, 139–175 (1994).

Dousseau, F. and M. Pezolet, "Determination of the secondary structure content of proteins in aqueous solutions from their amide I and amide II infrared bands: Comparison between classical and partial least-squares methods", Biochemistry **29**, 37, 8771–8779 (1990).

Durst, H. D., J. R. Mays, J. L. Ruth, B. R. Williams and R. V. Duevel, "Microscale synthesis and in-situ spectroscopic characterization of some chemical weapons related organophosphonate compounds", Analytical Letters **31**, 8, 1429–1444 (1998).

Eswar, N., B. John, N. Mirkovic, A. Fiser, V. A. Ilyin, U. Pieper, A. C. Stuart, M. A. Marti-Renom, M. S. Madhusudhan, B. Yerkovich and A. Sali, "Tools for comparative protein structure modeling and analysis", Nucleic Acids Res **31**, 13, 3375–80 (2003).

Ewig, C. S. and J. R. Van Wazer, "The ab initio structure of *O*-methyl methylphosphonofluoridate", Journal of Molecular Structure: THEOCHEM **122**, 3-4, 179–187 (1985).

Finnish Institute for Verification of the Chemical Weapons Convention (VERIFIN), *Chemical and Instrumental Verification of Organophosphorus Warfare Agents* (The Ministry of Foreign Affairs of Finland, Helsinki, 1977).

Flanigan, D. F., "Hazardous cloud imaging: a new way of using passive infrared", Applied Optics **36**, 27, 7027–7036 (1997).

Frisch, M., G. Tucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, J. Montgomerey, T. Vreven, K. Kudin, J. Burant, J. Millam, S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Hoda, O. Kitao, H. Nakai, M. Klene, X. Li, J. Knox, H. Hratchian, J. Cross, V. Nakken, C. Adamo, J. Jaramillo, R. Gomperts, R. Stratmann, O. Yazyev, A. Austin, R. Cammi, C. Pomelli, J. Ochterski, P. Ayala, K. Morokuma, G. Voth, P. Salvador, J. Dannenberg, V. Zakrzewski, S. Dapprich, A. Daniels, M. Strain, O. Farkas, D. Malick, A. Rabuck, K. Raghavachari, J. Foresman, J. Ortiz, Q. Cui, A. Baboul, S. Clifford, J. Cioslowski, B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Martin, D. Fox, T. Keith, M. Al-Laham, C. Peng, A. Nanayakkara, M. Challacombe, P. Gill, B. Johnson, W. Chen, M. Wong, C. Gonzalez and J. Pople, *Gaussian 03, revision E.01*, Gaussian, Inc., Wallingford, CT (2003).

Gaigeot, M. P. and M. Sprik, “Ab initio molecular dynamics computation of the infrared spectrum of aqueous uracil”, Journal of Physical Chemistry B **107**, 38, 10344–10358 (2003).

Gaigeot, M. P., R. Vuilleumier, M. Sprik and D. Borgis, “Infrared spectroscopy of *N*-methylacetamide revisited by ab initio molecular dynamics simulations”, Journal of Chemical Theory and Computation **1**, 5, 772–789 (2005).

Galabov, B. S. and T. Dudev, *Vibrational Intensities*, vol. 22 of *Vibrational Spectra and Structure* (Elsevier, Amsterdam, 1996).

Goldstein, H., C. P. Poole and J. L. Safko, *Classical Mechanics* (Addison Wesley, San Francisco, 2002), 3rd edn.

Gordon, M. S. and M. W. Schmidt, “Advances in electronic structure theory: GAMESS a decade later”, in “Theory and Applications of Computational Chemistry: the first forty years”, edited by C. E. Dykstra, G. Frenking, K. S. Kim and G. E. Scuseria, pp. 1167–1189 (Elsevier, Amsterdam, 2005).

Grahnen, J. A., K. E. Amunson and J. Kubelka, “DFT-based simulations of IR amide I$'$ spectra for a small protein in solution: Comparison of explicit and empirical solvent models”, Journal of Physical Chemistry B **114**, 40, 13011–13020 (2010).

Griffiths, D. J., *Introduction to Electrodynamics* (Prentice Hall, Upper Saddle River, N.J., 1999), 3rd edn.

Gurton, K. P., M. Felton, R. Dahmani and D. Ligon, “In situ infrared aerosol spectroscopy for a variety of nerve agent simulants using flow-through photoacoustics”, Applied Optics **46**, 25, 6323–6329 (2007).

Halls, M. D., J. Velkovski and H. B. Schlegel, “Harmonic frequency scaling factors for Hartree–Fock, S-VWN, B-LYP, B3-LYP, B3-PW91 and MP2 with the Sadlej pVTZ electric property basis set”, Theoretical Chemistry Accounts **105**, 6, 413–421 (2001).

Hering, J. A. and P. I. Haris, “FTIR spectroscopy for analysis of protein secondary structure”, in “Biological and Biomedical Infrared Spectroscopy”, edited by A. Barth and P. I. Haris, vol. 2 of *Advances in Biomedical Spectroscopy*, pp. 129–167 (IOS Press, Amsterdam, 2009).

Humphrey, W., A. Dalke and K. Schulten, “VMD: Visual molecular dynamics”, Journal of Molecular Graphics & Modelling **14**, 1, 33–38 (1996).

Jackson, J. D., *Classical Electrodynamics* (Wiley, New York, 1999), 3rd edn.

Jackson, M. and H. H. Mantsch, “The use and misuse of FTIR spectroscopy in the determination of protein structure”, Critical Reviews in Biochemistry and Molecular Biology **30**, 2, 95–120 (1995).

Jo, S., T. Kim, V. G. Iyer and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM", J Comput Chem **29**, 11, 1859–65 (2008).

Kaminský, J., P. Bouř and J. Kubelka, "Simulations of the temperature dependence of amide I vibration", The Journal of Physical Chemistry A **115**, 1, 30–34 (2011).

Kauppinen, J. K., D. J. Moffatt, H. H. Mantsch and D. G. Cameron, "Fourier self-deconvolution: A method for resolving intrinsically overlapped bands", Applied spectroscopy **35**, 3, 271–276 (1981).

King, W. T., "Effective atomic charges", in "Vibrational Intensities in Infrared and Raman Spectroscopy", edited by W. B. Person and G. Zerbi, vol. 20 of *Studies in Physical and Theoretical Chemistry*, chap. 6, pp. 122–142 (Elsevier, Amsterdam, 1982).

Komornicki, A. and R. L. Jaffe, "An ab initio investigation of the structure, vibrational frequencies, and intensities of HO_2 and HOCl", Journal of Chemical Physics **71**, 5, 2150–2155 (1979).

Kong, J. and S. Yu, "Fourier transform infrared spectroscopic analysis of protein secondary structures", Acta Biochimica Et Biophysica Sinica **39**, 8, 549–559 (2007).

Krimm, S. and J. Bandekar, "Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins", Advances in Protein Chemistry **38**, 181–364 (1986).

Kubelka, J., P. Bouř and T. A. Keiderling, "Quantum mechanical calculations of peptide vibrational force fields and spectral intensities", in "Biological and Biomedical Infrared Spectroscopy", edited by A. Barth and P. I. Haris, vol. 2 of *Advances in Biomedical Spectroscopy*, pp. 178–223 (IOS Press, Amsterdam, 2009).

Kubelka, J. and T. A. Keiderling, "Differentiation of β-sheet-forming structures: Ab initio-based simulations of IR absorption and vibrational CD for model peptide and protein β-sheets", Journal of the American Chemical Society **123**, 48, 12048–12058 (2001).

Liltorp, K. and Y. Maréchal, "Hydration of lysozyme as observed by infrared spectrometry", Biopolymers **79**, 4, 185–196 (2005).

MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", Journal of Physical Chemistry B **102**, 18, 3586–3616 (1998).

MacKerell, A. D., M. Feig and C. L. Brooks, “Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations”, Journal of Computational Chemistry **25**, 11, 1400–1415 (2004).

Markelz, A. G., “Terahertz dielectric sensitivity to biomolecular structure and function”, IEEE Journal of Selected Topics in Quantum Electronics **14**, 1, 180–190 (2008).

Merrick, J. P., D. Moran and L. Radom, “An evaluation of harmonic vibrational frequency scale factors”, Journal of Physical Chemistry A **111**, 45, 11683–11700 (2007).

Miyazawa, T., “Perturbation treatment of the characteristic vibrations of polypeptide chains in various configurations”, Journal of Chemical Physics **32**, 6, 1647–1652 (1960).

Miyazawa, T., T. Shimanouchi and S. I. Mizushima, “Normal vibrations of *N*-methylacetamide”, Journal of Chemical Physics **29**, 3, 611–616 (1958).

Mulliken, R. S., “Electronic population analysis on LCAO-MO molecular wave functions. I”, Journal of Chemical Physics **23**, 10, 1833–1840 (1955).

Pérez, C. and K. Griebenow, “Fourier-transform infrared spectroscopic investigation of the thermal denaturation of hen egg-white lysozyme dissolved in aqueous buffer and glycerol”, Biotechnology Letters **22**, 23, 1899–1905 (2000).

Pieper, U., B. M. Webb, D. T. Barkan, D. Schneidman-Duhovny, A. Schlessinger, H. Braberg, Z. Yang, E. C. Meng, E. F. Pettersen, C. C. Huang, R. S. Datta, P. Sampathkumar, M. S. Madhusudhan, K. Sjolander, T. E. Ferrin, S. K. Burley and A. Sali, “ModBase, a database of annotated comparative protein structure models, and associated resources”, Nucleic Acids Res **39**, Database issue, D465–74 (2011).

Plusquellic, D. F., K. Siegrist, E. J. Heilweil and O. Esenturk, “Applications of terahertz spectroscopy in biosystems”, ChemPhysChem **8**, 17, 2412–2431 (2007).

Rauhut, G. and P. Pulay, “Transferable scaling factors for density-functional derived vibrational force-fields”, Journal of Physical Chemistry **99**, 10, 3093–3100 (1995).

Rivin, D., G. Meermeier, N. S. Schneider, A. Vishnyakov and A. V. Neimark, “Simultaneous transport of water and organic molecules through polyelectrolyte membranes”, Journal of Physical Chemistry B **108**, 26, 8900–8909 (2004).

Schmidt, M. W., K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, "General atomic and molecular electronic-structure system", Journal of Computational Chemistry **14**, 11, 1347–1363 (1993).

Schropp, B., C. Wichmann and P. Tavan, "Spectroscopic polarizable force field for amide groups in polypeptides", Journal of Physical Chemistry B **114**, 19, 6740–6750 (2010).

Schultheis, V., R. Reichold, B. Schropp and P. Tavan, "A polarizable force field for computing the infrared spectra of the polypeptide backbone", Journal of Physical Chemistry B **112**, 39, 12217–12230 (2008).

Schweitzer-Stenner, R., "Advances in vibrational spectroscopy as a sensitive probe of peptide and protein structure: A critical review", Vibrational Spectroscopy **42**, 1, 98–117 (2006).

Scott, A. P. and L. Radom, "Harmonic vibrational frequencies: An evaluation of Hartree–Fock, Møller–Plesset, quadratic configuration interaction, density functional theory, and semiempirical scale factors", Journal of Physical Chemistry **100**, 41, 16502–16513 (1996).

Shagidullin, R. R., A. V. Chernova, V. S. Vinogradova and F. S. Mukhametov, *Atlas of IR Spectra of Organophosphorus Compounds (Interpreted Spectrograms)* (Kluwer, Dordrecht, 1990).

Söderström, M. T., "Identification of VX type nerve agents using cryodeposition GC-FTIR", AIP Conference Proceedings **430**, 1, 457–462 (1998).

Söderström, M. T., "Fourier transform infrared spectroscopy in analysis of chemicals related to the chemical weapons convention", in "Chemical Weapons Convention Chemicals Analysis: Sample Collection, Preparation and Analytical Methods", edited by M. Mesilaakso, chap. 14, pp. 353–385 (John Wiley & Sons, Ltd, 2005).

Sosa, C. and H. B. Schlegel, "A theoretical study of the infrared vibrational intensities of CH_3F", Journal of Chemical Physics **86**, 12, 6937–6945 (1987).

Stein, S. E., "NIST standard reference database 35. NIST/EPA gas-phase infrared database—JCAMP format", (1992).

Suenram, R. D., F. J. Lovas, D. F. Plusquellic, A. Lesarri, Y. Kawashima, J. O. Jensen and A. C. Samuels, "Fourier transform microwave spectrum and ab initio study of dimethyl methylphosphonate", Journal of Molecular Spectroscopy **211**, 1, 110–118 (2002).

Susi, H., "Infrared spectroscopy—conformation", in "Enzyme Structure, Part C", edited by C. H. W. Hirs and S. N. Timasheff, vol. 26 of *Methods in Enzymology*, pp. 455–472 (Academic Press, 1972).

Tamm, L. K. and S. A. Tatulian, "Infrared spectroscopy of proteins and peptides in lipid bilayers", Quarterly Reviews of Biophysics **30**, 4, 365–429 (1997).

Torii, H. and M. Tasumi, "Model calculations on the amide-I infrared bands of globular proteins", Journal of Chemical Physics **96**, 5, 3379–3387 (1992).

Torii, H. and M. Tasumi, "Infrared intensities of vibrational modes of an α-helical polypeptide: Calculations based on the equilibrium charge/charge flux (ECCF) model", Journal of Molecular Structure **300**, 171–179 (1993).

Torii, H. and M. Tasumi, "Theoretical analyses of the amide I infrared bands of globular proteins", in "Infrared Spectroscopy of Biomolecules", edited by H. H. Mantsch and D. Chapman, chap. 1, pp. 1–18 (Wiley-Liss, New York, 1996).

Vanommeslaeghe, K., E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. MacKerell, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields", Journal of Computational Chemistry **31**, 4, 671–690 (2010).

Vishnyakov, A., G. Y. Gor, M. T. Lee and A. V. Neimark, "Molecular modeling of organophosphorous agents and their aqueous solutions", Journal of Physical Chemistry A **115**, 20, 5201–5209 (2011).

Vishnyakov, A. and A. V. Neimark, "Molecular model of dimethylmethylphosphonate and its interactions with water", Journal of Physical Chemistry A **108**, 8, 1435–1439 (2004).

Walker, A. R. H., R. D. Suenram, A. Samuels, J. Jensen, M. W. Ellzy, J. M. Lochner and D. Zeroka, "Rotational spectrum of Sarin", Journal of Molecular Spectroscopy **207**, 1, 77–82 (2001).

Wilson, E. B., J. C. Decius and P. C. Cross, *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra* (McGraw-Hill, New York, 1955).

Zerbi, G., "Introduction to the theory of vibrational frequencies and vibrational intensities", in "Vibrational Intensities in Infrared and Raman Spectroscopy", edited by W. B. Person and G. Zerbi, vol. 20 of *Studies in Physical and Theoretical Chemistry*, chap. 3, pp. 23–64 (Elsevier, Amsterdam, 1982).