MAX-PLANCK-GESELLSCHAFT

vrije Universiteit amsterdam

Master's Thesis

Anharmonic Infrared Spectra from Short QM/MM Simulations

prepared by

Timo Marcel Daniel Graen

from Hannover

at the Vrije Universiteit Amsterdam and the
Max Planck Institute for Biophysical Chemistry Göttingen

Thesis period:	1st August 2010 until 30th July 2011
Supervisor:	Dr. Gerrit Groenhof
First Referee:	Prof. Dr. Marloes Groot
Second Referee:	Dr. Daan P. Geerke

Abstract

Proteins are the workhorses of life. Understanding protein function on the molecular level is a very active field of research. Experimental infrared difference spectra contain valuable information about protein dynamics but the experimental data is difficult to interpret. Calculated infrared spectra from simulated protein dynamics contain the desired all-atom information but rely on many approximations which require experimental validation. Together, experiment and simulation can greatly increase the knowledge gained about the investigated system.

This thesis presents a method which enables the calculation of anharmonic vibrational difference spectra from short quantum mechanics/molecular mechanics simulations including the full protein environment. The developed simulation scheme extends state of the art dipole moment time series analysis methods to active centers of large solvated proteins. The method does not depend on the choice of quantum method and can be applied to ground and excited states. Short trajectory lengths limit the spectral resolution of Fourier spectra. Parameter based alternatives were investigated to overcome the Fourier resolution limit.

The vibrational difference spectrum for the photoactive yellow protein and its locked mutant was calculated. The computational cost of simulating the green fluorescent protein active pocket exceeded the available resources and low spectral resolution was obtained. The parametric Burg method was identified as a working Fourier transform alternative for data analysis. A suitable model order estimation scheme was developed based on the normal mode frequency density. High level anharmonic vibrational spectra make harmonic normal mode spectra obsolete.

Keywords: Infrared spectrum, anharmonic spectrum, difference spectra, dipole moment time series analysis, charged dipole moment, GFP, PYP, QM/MM simulations

Contents

1 Introduction

Atoms are peculiar objects. The quantum mechanical behavior of atoms is counterintuitive to the human perception of the world. In the classical world, objects can store continues amounts of energy and trajectories are deterministic. The properties of atomic systems remain hidden to human sensory organs and can rarely be found in the macroscopic world. Music instruments are a simple example of quantum properties in the macroscopic world. The oscillation of the guitar string is a particularly good example. The guitar string can be extended continuously out of its equilibrium position. However, the boundary conditions of the instrument restrict the solutions of the resulting differential equation to discrete harmonic sine functions. This had important implications for scientists in the eighteenth century [53]. An arbitrary function of the string extension was expressed as an infinite sum of harmonic sine functions. This result is a substantial cornerstone in the formulation of Fourier transforms. In the early twenties century, a series of publications by Erwin Schrödinger [57] beautifully illustrates how the discrete nature of the experimentally measured hydrogen spectrum contributed to the formulation of modern quantum physics. The publications relate the differential equation describing the oscillating string to the spectrum of the hydrogen atom and solve the discrete eigenvalues of the hydrogen atom wave function.

Today, state of the art experiments reach atomic time and length scales. The focus has shifted from single atoms towards molecules and large protein systems. This has vastly increased the complexity of both experimental setups and data analysis. The purpose is to understand life on all its levels ranging from macroscopic understanding of the ecosystem earth all the way down to single molecule mechanisms at atomic resolution. The author of this thesis is interested in the later, namely understanding the dynamics of life at the atomic scale.

All life is sustained by proteins. Proteins are a diverse set of molecules which are predominantly assembled from 20 main amino acids found in nature. Interestingly, certain chain combinations of amino acids form stable reoccurring structures under narrow environmental conditions while others do not. This process is known as self-assembly or protein folding. Predicting how a sequence of amino acids will fold is currently only possible for very small proteins [22]. Nature solved the folding problem in a long natural evolutionary process which is still ongoing today. A disruption in the expected protein fold will almost certainly result in miss folding and death of the cell up to the death of the human host. However, it is important to notice that only part of the information needed to understand life on the atomic scale is contained in the protein structure. The more important part of the information is contained in the protein dynamics [40]. This can be related to knowing the shape of a tool without knowing how to use it or better, how to build a new one.

To the physicist, proteins are molecular machines preassembled by nature with lost construction manuals. Proteins self assemble without eminent presence of symmetry or a deterministic relation between folded structure and function. Proteins consist of bosons and fermions which are subject to electromagnetism. The laws of statistical and quantum physics describe protein motion and do, in principle, solve all problems in biology and chemistry. However, the reach of analytical molecular physics ends somewhere between the hydrogen atom and the hydrogen molecule. From there on, many approximations are necessary to study and predict protein dynamics from theory. Without approximations, the computational cost of predicting the time evolution for thousands of atoms is far out of reach even for the most advanced computers today and at the current rate of chip development also for many years to come. Approximations enormously reduce the cost of calculating protein dynamics but the list of required approximation is long and it is not always clear beforehand whether a theoretical model will hold or not. Therefore, computational studies of protein systems must heavily rely on experimental input.

Studying proteins in experiments is not an easy task either. A major challenge in the experimental quest is information gathering at length scales of nuclear resolution which corresponds to around 10^{-10} m. This is well below the wavelength of visual light and can therefore not be observed using conventional microscopy. X-ray scattering is a powerful tool to overcome this barrier for structure determination. Here, a periodic crystal of the bio-molecule is required. In the experiment, the diffraction pattern of the crystal is measured. This corresponds to the absolute square value of the Fourier coefficients of the structure, the intensity. As only these intensities can be measured, the complex phase information is lost. The phase information has to be recovered using computational models of the structure. An alternative method are NMR experiments. No crystal is required and measurements can be performed on solvated proteins. NMR experiments measure the spin coupling of ^{1}H and ^{13}C atoms. The experimental data is an average of atomic distances r as $\langle r^{-6} \rangle$ which can then be used as constrains in computational structure prediction models. However, direct determination of molecular structure and especially dynamics currently remains an unsolved problem.

The inaccessibility of molecular dynamics at atomic resolution challenged the creativity of the scientific community. The holy grail is the creation of molecular movies with very high time and spatial resolution. Light is a powerful tool to study protein dynamics. The effects of light are diverse and can be roughly separated into three groups.

First, the UV range of the spectrum which carries enough energy to break molecular bonds and can be used to study repair dynamics, i.e. after DNA photo damage. Second, the visible spectrum of light which excites electronic states of molecules to higher energy states. This changes the energy landscape generated by the electrons on the nuclei. The absorbed energy can be passed on as is naturally observed in the dynamics of plant photo systems or used in the design of FRET experiments. The energy can also dissipate into heat causing molecular vibrations to increase in amplitude. Alternatively, the molecular vibrations can directly be excited through infrared light. Infrared light is the third region of light which is highly interesting for studying molecular dynamics.

The interaction of infrared light with molecules probes the vibrational degrees of freedom of the nuclear wave function. In the picture of quantum mechanics, the motion of atomic nuclei in a molecule is coupled and quantized. For N atoms there are $3N$ degrees of freedom which is reduced to $3N - 6$ by removing three translational and three rotational degrees of freedom for the whole molecule. Thus, a water molecule of three atoms has three collective nuclear degrees of freedom or collective vibrational modes. Vibrational modes can be localized to a few atoms or extended over many bonds involving a large number of atoms simultaneously. Modes can be coupled and exchange energy depending on the overlap between their vibrational wave functions. According to the standard model, the photon is the force carrier of electromagnetism. This allows an intuitive description of molecular orbitals in terms of photon energies. The same holds for molecular vibrations. Vibrational modes can be quantified in terms of photon energies. For low excitation numbers, the allowed quantized energy states of each collective vibrational mode are an integer multiple of the corresponding ground state photon energy. For a single water molecule, this results in three distinct ground state photon energies for all three vibrational modes. Each mode acts as a photon bin which can collect photons matching its energy signature. For the first few collected photons this signature is roughly constant but decreased with the number of collected infrared photons. This corresponds to the solutions of the anharmonic oscillator. The physical effect of collecting photons is an increase in amplitude of the collective vibrations. At a certain point the amplitude increases to a point where the molecule falls apart and the atoms are no longer bound.

Vibrational spectra of proteins can be measured experimentally. The protein is exposed to a broad pulse of infrared light which is partially absorbed by the protein. Unfortunately, this experimental pathway towards obtaining vibrational spectra is not accessible in computer simulations. This is mainly due to the fact that the collective modes are unknown beforehand and the vibrational wave function is too expensive to calculate. An elegant way to circumvent the computational complexity of calculating the vibrational wave function is to explore the potential energy landscape using quantum (QM), classical (MM) or combined QM/MM simulations. The time series data generated from multiple simulation trajectories can then be used to reconstruct frequency information about the energy landscape [6].

Vibrational difference spectra can be measured in experiments. The experimental data contains high quality information about differences in protein dynamics between two protein states. This can be the difference between two protein mutants or the difference between ground and excited states. However, the data analysis is often very difficult due to the large number of modes and the lack of all-atom resolution. This is were simulations can greatly help to extend the knowledge gained from the experiment. The aim of this thesis is to calculate protein difference spectra from all-atom molecular simulations which can then be compared to the experimentally measured spectra.

This thesis attempts to extend the state of the art anharmonic infrared simulation methodology [1, 9, 24, 37, 38, 44, 58, 74, 77]. The presented method enables the calculation of anharmonic vibrational spectra from QM regions in QM/MM simulations while the full protein environment is included. The method does not depend on the choice of QM method which is why excited state spectra can also be calculated. It does not make assumptions about the underlying potential and does not require the harmonic approximation. Temperature effects are naturally included and spectral bands contain absolute values for the power.

Applications for this method are chromophores and photoactive protein centers. Two important experimental systems are discussed in this thesis, the green fluorescent protein (GFP) and the photoactive yellow protein (PYP). Both proteins are complex enough to be interesting but also small enough to allow QM/MM simulations. As both systems have charged active pockets, the behavior of the ill-defined dipole moments for charged systems is investigated and a correction scheme is introduced. The influence of short simulation trajectories from high level QM/MM simulations is discussed with respect to the spectral resolution limit of the Fourier transform. It is investigated how the resolution limit can be increased by using parameter based Fourier transform alternatives. The maximum entropy method as well as the autoregressive filter based Burg method are investigated. As parameter based spectra strongly depend on the model order, an objective model order estimation scheme is introduced based on the normal mode frequency distribution.

High level anharmonic vibrational difference spectra of the PYP chromophore and its locked mutant are calculated to illustrate the power of the developed simulation scheme. The quality of the results render normal mode spectra obsolete. QM/MM simulations are far superior for the calculation of protein difference spectra.

2 Theory

2.1 Propagating Proteins in Time

The focus of this study are proteins. The most accurate description of the nuclear and electronic processes within proteins is the framework of quantum mechanics. For most proteins relativistic effects are negligible and the time dependent Schrödinger equation [57] is a very accurate model for propagating quantum mechanical systems in time.

$$i\hbar \frac{\partial \Psi(t)}{\partial t} = H\Psi(t) \tag{2.1}$$

Unfortunately, the Schrödinger equation can only be solved analytically for the smallest of all systems such as the Hydrogen atom, the rigid rotor, the harmonic oscillator or the particle in a box [2]. The following sections will focus on approximations which help to extend the Schrödinger equation to large protein systems.

2.2 Molecular Dynamics Principles

The Molecular Dynamics simulation framework reduces the vast computational complexity of solving the time dependent Schrödinger equation down to solving the time evolution of many classical point charges in a classical, often even partially harmonic, potentials. Many approximations are necessary to propagate thousands of atoms on microsecond timescales. Three main approximations are introduced in the following paragraphs. First, the Born-Oppenheimer approximation is applied to separate the wave function. Second, classical Newton mechanics are introduced to propagate the nuclei in time. At this point Hartree-Fock (HF) theory and the post HF complete active space self consistent field (CASSCF) method are introduced to describe the electronic wave function. The density formulation of the electron wave function is introduced subsequently in terms of Density Functional Theory (DFT). Third, the efficient force field approximation to the electronic wave function is introduced.

2.2.1 Born-Oppenheimer Approximation

The Born-Oppenheimer Approximation [8] decouples the electronic from the nuclear degrees of freedom. In the following, the time independent Schödinger equation is used but all assumptions directly relate to the time dependent version through the time propagation operator. First, the wave function $\Psi(r, R)$ is expanded as a product of the electronic wave function $\phi_e(r, R)$ and the nuclear wave function $\psi_n(R)$ while the expansion is truncated after the first term,

$$\Psi(r, R) \approx \phi_e(r, R)\psi_n(R). \tag{2.2}$$

Here, r are the coordinates of the electrons and R are the coordinates of the nuclei. Next, the nuclear degrees of freedom are assumed to be much slower than the electronic degrees of freedom. One argument to support this approximation is the large difference in mass between nuclei and electrons leading to much faster electronic motion. The consequence of this approximation is a parametric dependence of the nuclear positions in the electronic wave function $\phi_e(r; R)$ instead of a variable dependence $\phi_e(r, R)$.

2.2.2 Classical Equations of Motion

Within the framework of the Born-Oppenheimer approximation, the Schrödinger equation can now be solved in two steps. First, the electronic wave function for a given set of nuclear positions is solved. Second, the electronic potential enters the nuclear Schrödinger equation as

$$(\hat{T}_n + \hat{V}_{nn} + E_e(R))\psi_n(R) = V_n(R)\psi_n(R). \tag{2.3}$$

The energy $V_n(R)$ of the nuclear wave function is then determined by the nuclear kinetic energy operator $\hat{T}_n$, the nuclear-nuclear potential energy operator $\hat{V}_{nn}$ and the contribution from the electronic potential $E_e(R)$. In this approximation, quantum effects of the nuclear motion are neglected and the nuclear positions are propagated using Newton's equations of motion:

$$F = m\frac{d^2R(t)}{dt^2} = -\nabla V_n(R) \tag{2.4}$$

At this point the electronic wave function is required to propagate the system in time. The problem of obtaining this wave function for reasonably large systems is the central problem of computational quantum chemistry. Before continuing with the Molecular Dynamics (MD) force field approximation to the electronic wave function, three methods of quantum chemistry are introduced. First, the fundamental Hartree-Fock theory [section 2.2.3] and its extension the Complete Active Space SCF method [section 2.2.4] are discussed, followed by a short introduction to density functional theory [section 2.2.5].

2.2.3 Hartree-Fock Theory

Describing wave functions of many electron molecules cannot be handled analytically and must currently rely on approximations to reduce the computational cost of the calculation. Hartree-Fock (HF) theory [56, 79] is a fundamentally important approach for solving the electronic wave function. The theory is mathematically elegant and computationally efficient. Even though HF theory neglects correlation effects beyond Coulomb and electron exchange, it provides a suitable basis for higher level quantum chemical methods. Hartree-Fock theory produces relatively good ground state structures but fails to describe almost all relevant chemical properties of molecular systems. More accurate multi Slater determinant and perturbation based methods exist for including the missing correlation effects and describing excited states but only at very high computational cost. The reader is referred to methods such as CCSD, CISD, MP4, CASSCF as described in the literature [35].

Hartree-Fock theory is formulated as an optimization problem [2, 35, 45] where the optimal solution to the Schrödinger equation is obtained by minimizing the energy E as a function of a trial wave function Ψ. The energy estimation is obtained from the expectation value of the Hamiltonian

$$H_e = T_e + V_{ne} + V_{ee} + V_{nn} \tag{2.5}$$

as

$$E_e = \frac{\langle \Psi | H_e | \Psi \rangle}{\langle \Psi | \Psi \rangle}. \tag{2.6}$$

The electron/nuclei interaction is described by $\mathcal{V}_{ne}$ and analogously the electron/electron and nuclei/nuclei interaction as $\mathcal{V}_{ee}$ and $\mathcal{V}_{nn}$. The kinetic energy of the electrons is $\mathcal{T}_e$ and the energy is given in $\langle bra|ket \rangle$ notation. Individual contributions to equation 2.5 can be grouped together according to their number of electron different indices. This leads to the electron operators

$$\begin{aligned} h_i &= -\frac{1}{2}\nabla_i^2 - \sum_a^{N_{nucl}} \frac{Z_a}{|R_a - r_i|}, \\ g_{ij} &= \frac{1}{|r_i - r_j|}. \end{aligned} \tag{2.7}$$

The r_i denote electron coordinates, R_i and Z_i nuclear coordinates and charges respectively. For normalized wave functions, the denominator in equation 2.6 is $\langle\Psi|\Psi\rangle = 1$. And equation 2.6 can be simplified to

$$E_e = \left\langle \Psi| \sum_i^{N_{elec}} h_i + \sum_{j>i}^{N_{elec}} g_{ij} + \mathcal{V}_{nn}|\Psi \right\rangle. \tag{2.8}$$

The molecular orbitals ϕ_i in the trial wave function Ψ_{trial} are products of a spatial and a spin function. The required antisymmetry of the wave function as is required by the Pauli principle is included using the mathematical properties of determinants. Thus, a possible guess Ψ_{trial} for the HF minimization problem is often a so called Slater determinant Φ_{SD},

$$\Psi_{trial} = \Phi_{SD} = \frac{1}{\sqrt{N!}} \begin{pmatrix} \phi_1(1) & \phi_2(1) & \cdots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \cdots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \cdots & \phi_N(N) \end{pmatrix}. \tag{2.9}$$

Equation 2.6 can now be written as a Lagrange optimization problem under the condition that the molecular orbitals remain orthogonal, $\langle\phi_i|\phi_j\rangle = \delta_{ij}$:

$$L = E_e - \sum_{ij}^{N_{elec}} \lambda_{ij}(\langle\phi_i|\phi_j\rangle - \delta_{ij}). \tag{2.10}$$

The variation of the Lagrangian follows as

$$\delta L = \delta E_e - \sum_{ij}^{N_{elec}} \lambda_{ij}(\langle\delta\phi_i|\phi_j\rangle - \langle\phi_i|\delta\phi_j\rangle). \tag{2.11}$$

while δE_e can be reduced to

$$\delta E_e = \sum_i^{N_{elec}} \left(\langle\delta\phi_i|F_i|\phi_i\rangle + \langle\phi_i|F_i|\delta\phi_i\rangle\right). \tag{2.12}$$

Here,

$$F_i = h_i + \sum_{j}^{N_{elec}} (J_i - K_i) \tag{2.13}$$

is the Fock operator and J_i and K_i are the Coulomb and exchange operators, respectively. The Lagrange condition can be reduced further to the final Hartree-Fock equations in matrix notation:

$$F_i \phi_i = \sum_{j}^{N_{elec}} \lambda_{ij} \phi_i. \tag{2.14}$$

Through unitary transformation, the matrix of Lagrange multipliers λ_i can be made orthogonal resulting in a pseudo eigenvalue problem of the canonical molecular orbitals (MO) ϕ_i':

$$F_i \phi_i' = \epsilon_i \phi_i'. \tag{2.15}$$

Specific Fock orbitals cannot be calculated individually. This requires the Hartree-Fock equations to be solved iteratively. Thus, solutions to the HF equations are called self consistent field (SCF) orbitals.

2.2.4 CASSCF

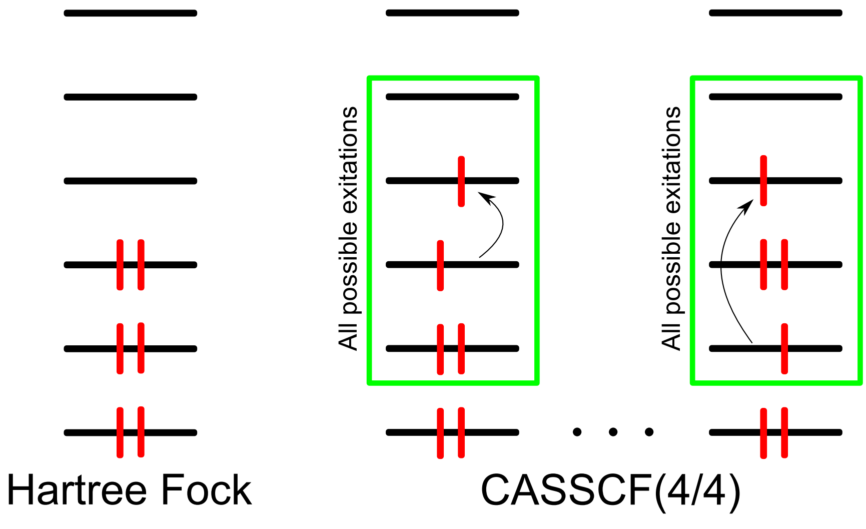

Figure 2.1: Illustration of a 4 electron 4 orbital CASSCF(4/4) active space; left) HF wave function; right) two representations of the full active space (green box).

The Complete Active Space Self Consistent Field (CASSCF) [29] method is a computationally efficient truncated version of the full Configuration Interaction (CI) method. The full CI method is an extension of the Hartree-Fock method which converges towards the exact solution of the Schrödinger Equation within the basis set accuracy. The extremely high accuracy comes at the cost of $N!$ scaling with the number of electrons [35]. This makes it physically impossible to apply the method to anything larger than a few electrons. The underlying idea of the CI method is to include not only the HF ground state Slater determinant but also all possible single electron and higher excitations [35] as a weighted sum

$$\Psi_{CI} = a_0 \Phi_{HF,GS} + \sum_{single} a_s \Phi_s + \sum_{doubble} a_d \Phi_d + \sum_{tripple} a_t \Phi_t + \ldots = \sum_i a_i \Phi_i. \tag{2.16}$$

In this representation of the wave function, each $\Phi_{s,d,t,...}$ is a set of many Slater determinants including all possible combinations for all possible excitation levels. The new set of coefficients a_i is then optimized under the condition of orthogonality $\langle \Psi_{CI} | \Psi_{CI} \rangle = 1$ using Lagrange multipliers λ_i as

$$L = \langle \Psi_{CI} | H | \Psi_{CI} \rangle - \lambda(\langle \Psi_{CI} | \Psi_{CI} \rangle - 1). \tag{2.17}$$

The CASSCF truncation of this methods is illustrated in figure 2.1. Instead of treating all electrons at the Full CI level of theory, only a set of selected orbitals is considered in the CI expansion. However, this approach requires optimizing both the CI coefficients a_i as well as the basis set coefficients from the Slater determinants.

CASSCF recovers most of the static correlation energy which is assumed to be more important for accurate descriptions of excited states. The dynamic correlation energy is only recovered poorly for small active spaces. However, including more of the dynamic correlation energy also significantly increases the computational cost of the calculation which is not acceptable for calculating dynamics. The accuracy of the CASSCF method critically depends on the quality of the chosen active space. This is also the reason why CASSCF is not a black box method. Careful selection of orbitals is required. Dynamics of large molecules can only be performed using reduced active spaces. For most reductions, a large number of orbital combinations exists but only very few will result in high quality wave functions.

2.2.5 DFT

Large effort has been put into the development of electron density and time dependent density based theories which do not require explicit electron coordinates but calculate the wave function based on its spatial electron density [4, 14, 21, 55]. Thus, the integration is reduced from 3N to 3 dimensions with a vast increase in performance compared to multi determinant methods. However, currently there is no sufficiently physical description of the correlation/exchange contribution with respect to the electron density. Therefore, Density Functional Theory (DFT) requires fitting against high level quantum Monte Carlo simulations and/or experimental data which is why many DFT based methods are considered semi empirical.

The underlying idea of Density Functional Theory (DFT) [32, 41] is to separate the density dependent energy

$$E[\rho] = T_s[\rho] + E_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \tag{2.18}$$

into single electron contributions for kinetic energy $T_s[\rho]$, electron/nuclear interaction $E_{ne}[\rho]$ and Coulomb interaction $J[\rho]$. The correlation and exchange part of the energy is excluded from the description and stored in the undefined E_{xc} contribution. The correlation part E_c of this energy is also the motivation for expanding the HF determinants in the full CI expansion. In HF theory, E_x is exact while E_c is missing completely. Unfortunately, the HF exact exchange is not directly compatible with the DFT E_{xc} energy. The first approaches towards describing the correlation exchange energy were based on tabulating exact results from high level homogeneous electron gas simulations. These are homogeneous local density (LDA) and local gradient extended generalized gradient (GGA) functionals [42]. Note, the often mentioned non-local corrections refer only to the inclusion of local gradients. This group of xc functionals fails to describe charge transfer as the needed 1/r behavior is not recovered with respect to the distance r. Instead, they decay exponentially. The reader is referred to a graphical representation of non-locality of the correlation and exchange holes as was calculated by Towler [61] using the variational Monte Carlo method.

Second, a set of hybrid functionals was developed [42], including B3LYP, PBE0, Half/Half and others. For these functionals, some part of the missing long range exchange is mixed into the pure adiabatic approximation (AA) DFT exchange by directly including Hartree-Fock (HF) exchange. The amount of exact HF exchange then determines the long range scaling of the functional, 0.2/r in the case of B3LYP. Note, further increasing the amount of HF exchange reduces the needed cancellation of errors in xc-functionals between the correlation and exchange energies. Thus, simply increasing HF exchange in B3LYP until the 1/r long range behavior is recovered will introduce large errors in the local description.

Third, long range corrected, Coulomb attenuated or range separated xc-functionals [13, 28, 67, 68] have been developed in recent years, referred to as LC-functionals in the following text. They all rely on the separation of the two particle electron-electron interaction into a short and a long range part as described by Leininger *et al.* [46] as

$$\frac{1}{r_{i,j}} = \underbrace{\frac{g(r_{i,j})}{r_{i,j}}}_{\text{LR}} + \underbrace{\frac{1-g(r_{i,j})}{r_{i,j}}}_{\text{SR}} . \tag{2.19}$$

The separating function was conveniently chosen as $g(r_{i,j}) = erf(\mu r_{i,j})$, the error function. From this definition, it becomes clear that LC-functionals introduce an additional parameter μ which determines the transition between the short range and the long range description [54, 55]. The underlying idea is to leave the local description of the xc-functionals, i.e. LDA, GGA unchanged. Thus, retaining the important cancellation of errors while including exact HF exchange for long range interactions. The error function is then used to mediate the two contributions to the energy. Therefore LC-functionals do not fully correct for the one electron SIE [68], as the local SIE still remains and only the long range SIE is compensated by the correct 1/r behavior.

All of the mentioned DFT methods are single configuration methods and especially for the long range corrected functionals the computational cost becomes similar to small multiconfigurational CAS active spaces. Thus, excited state simulations using the CASSCF method are still appealing for QM/MM simulations.

2.2.6 Force Field Approximation

The approximations up to this point allow the accurate time propagation of several dozen atoms. This level of theory is referred to as Born-Oppenheimer dynamics (BO) and still too expensive to model large bio-organic systems in solution. The quantum mechanical potential from equation 2.4 is now replaced by a purely classical spring potential, the force field.

The basis for the molecular dynamics (MD) simulations performed in this thesis is the 2006 AMBER99sb force field [33] which is based on the Wang et al. AMBER99 [69] and Cornell et al. AMBER94 force field [17]. The AMBER force fields originate from the 1984 Weiner force field [72]. The AMBER force field is a pairwise empirical fit to the potential of the electronic wave function. The functional form has not changed over the years and the differences are in the fitting of the functional parameters. The function used is a classical ball and spring model of the form

$$\begin{aligned} E_{total} &= \sum_{bonds} K_r(r-r_{eq})^2 + \sum_{angles} K_\theta(\theta-\theta_{eq})^2 + \\ & \sum_{dihedrals} \frac{V_n}{2}\left[1+cos(n\varphi-\gamma)\right] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{1}2} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]. \qquad (2.20) \end{aligned}$$

The function describes the equilibrium bond distance r_{eq} and the spring constant of bond oscillation K_r. The angles between three atoms are described using a harmonic angle potential with equilibrium angle θ_{eq} and force constant K_θ. Out of plane dihedral motion of an atom chain *a-b-c-d* describes the rotation of atoms *a* and *d* around the axis of atoms *b* and *c* at angle θ. This motion is described by a Taylor series of a multi well potential whose Fourier coefficients are V_n and the phase is γ. The Van der Waals potential is described by the parameter for repulsion A_{ij} and the attraction B_{ij} over the distance R_{ij}. Finally, the last term of the potential function describes the Coulomb interaction between the atom point charges q_i and q_j at distance R_{ij}.

Notice, how for a given set of protein atoms all possible two, three and four atom connections need to be parametrized. This easily sums up to several ten thousand force field parameters and implicitly explains why there are so many different force fields [33]. The problem of generating a good force field is under determined given the available experimental parameters. This is why after over almost thirty years there are still improvements made to the original parameter set. Force field development is a tedious task that will likely continue for many years to come.

2.3 QM/MM Simulations

Molecular dynamics simulations are a powerful tool to describe equilibrium properties of large molecular systems such as proteins. The approximations made in the molecular dynamics framework completely neglect explicit electrons and reduce the effect of the electronic wave function on the nuclei to an empirical classical potential, the force field. However, the force field approximation will completely fail to describe chemical reactions which involve bond breaking or photon absorption. This renders molecular dynamics useless for the simulation of proton transfer and excited state processes. The idea behind QM/MM simulations is to lift the force field approximations on certain parts of the protein by including the electronic wave function directly through quantum mechanical calculations. This QM/MM approach enables simulations of processes which require explicit electrons while also explicitly including the protein environment at low computational cost.

The QM/MM method is consistent with the three main approximations described above. The efficient propagation of the quantum region (QM) requires the Born-Oppenheimer approximation as well as Newtons equation of motion for the nuclear motion. In addition to these two approximations, the protein environment (MM) is described using also the force field approximation which replaces explicit electrons with an empirical force field. As both regions approximate nuclear motion by Newtons equations of motion, the same time propagation algorithm is used for both the QM and MM region of the system. The MM forces are calculated from the force field while the QM forces are derived from the electronic wave function.

In order to create a physically meaningful simulation, the QM region must be allowed to interact with the MM environment. This can be achieved by directly coupling the quantum region to the environment via embedding of the classical nuclei into the electronic wave function. This method is called electronic embedding and includes the classical point charges from the MM region into the quantum Hamiltonian via [18]:

$$H^{qm/mm} = H^{qm}_{electrons} - \sum_{i}^{n} \sum_{J}^{M} \frac{e^2 Q_J}{4\pi\epsilon_0 r_{iJ}} + \sum_{I}^{N} \sum_{J}^{M} \frac{e^2 Z_I Q_J}{\pi\epsilon_0 R_{IJ}}. \tag{2.21}$$

In this equation, the first double sum describes the coupling of n quantum mechanical electrons to M classical point charges Q_J via Coulomb interaction over the distance r_{iJ}. The second double sum couples the N nuclei with charge Z_I in the QM region to M point charges Q_J in the MM region with distance R_{IJ} via Coulomb interaction.

At the border between QM and MM regions, the angles, dihedrals and impropers connecting the MM to the QM region are taken from the MM force field. A distance constraint is used to replace the bond between QM and MM region and set to the corresponding equilibrium distance. As this bond is missing in the QM calculation a lone electron pair is created. This artifact is removed by placing a virtual hydrogen atom along the constrained bond. The virtual hydrogen is included in the QM calculation and ignored by the MM force field. The resulting force on the hydrogen is evenly distributed among the QM and MM binding partner.

3 Spectral Analysis of QM/MM Trajectories

3.1 The Challenge

The calculation of an-harmonic infrared spectra is a challenging task for large molecules. A good model should be able to describe the system of interest in its natural environment rather than in vacuum [59]. It should also include native support for anharmonicity without having to rely on anharmonic corrections as done in the case of harmonic normal modes . Additionally, the quality of the calculated power density spectra should be high enough to allow the calculation of difference spectra between states of interest. Thus a high resolution in frequency space is desirable. The method of choice should also capture temperature effects and system dynamics at both the ground as well as the excited state. The method presented in the following chapters includes all of the mentioned features as it is based on the time series analysis of finite temperature QM/MM simulations. Here, the system of interest is simulated at room temperature in its natural environment. From the trajectory of this simulation, the dipole moment time series is recorded which is the basis for the calculation of the anharmonic IR spectrum.

3.1.1 IR Transition Probabilities

In short summary, the vibrational spectrum of a molecular system is a representation of periodic motion among the nuclei. The oscillation frequencies of the different collective nuclear motions appear as peaks in the spectrum. This motion behaves quantum mechanically, thus not all vibrations are allowed and their energies are quantized. In theory, these allowed collective frequencies can be obtained by applying perturbation theory to the time dependent Schrödinger Equation:

$$\hat{H}\psi(x,t) = -\frac{\hbar}{i}\frac{\partial\psi(x,t)}{\partial t} \tag{3.1}$$

This section was inspired by the discussion on quantum mechanical transition probabilities in references [58] and [43]. The perturbation $\hat{H}'$ is the effect of the electric field due to infrared light passing the system. This leads to the perturbed Schrödinger equation

$$(\hat{H} + \hat{H}')\psi'(x,t) = -\frac{\hbar}{i}\frac{\partial}{\partial t}\psi'(x,t). \tag{3.2}$$

The resulting perturbed wave function $\psi'(x,t)$ can be expanded in the basis of the eigenfunctions $\psi_n(x,t) = \psi_n(x)e^{-iE_nt/\hbar}$ of the unperturbed Hamiltonian $\hat{H}$ as

$$\psi'(x,t) = \sum_n a_n(t)\psi_n(x)e^{-iE_nt/\hbar}. \tag{3.3}$$

The expansion coefficients $a_n(t)$ contain valuable information about the probability $p_m(t) = |a_m(t)|^2$ of finding the original system $\psi_s(x)$ in the final state $\psi_m(x)$ after a time t. Thus, an expression for the expansion coefficient will be derived in the following paragraphs. First, inserting equation 3.3 into equation 3.2 results in

$$\begin{aligned}(\hat{H} + \hat{H}')\sum_n a_n(t)\psi_n(x)e^{-iE_nt/\hbar} &= -\frac{\hbar}{i}\frac{\partial}{\partial t}\sum_n a_n(t)\psi_n(x)e^{-iE_nt/\hbar} \\ &= -\frac{\hbar}{i}\sum_n \dot{a}_n(t)\psi_n(x)e^{-iE_nt/\hbar} \\ &\quad -\frac{\hbar}{i}\sum_n a_n(t)\frac{\partial}{\partial t}\psi_n(x)e^{-iE_nt/\hbar}. \end{aligned} \tag{3.4}$$

The relation

$$\sum_n a_n(t)\hat{H}\psi_n(x)e^{-iE_nt/\hbar} = -\sum_n a_n(t)\frac{\hbar}{i}\frac{\partial}{\partial t}\psi_n(x)e^{-iE_nt/\hbar} \tag{3.5}$$

is applied to the last term of equation 3.4 resulting in a simplified expression for the derivative of the expansion coefficients a_n,

$$\sum_n a_n(t)\hat{H}'\psi_n(x)e^{-iE_nt/\hbar} = -\frac{\hbar}{i}\sum_n \dot{a}_n(t)\psi_n(x)e^{-iE_nt/\hbar}. \tag{3.6}$$

This equation can be reduced further by left-multiplying $\psi_m(x,t)$, applying

$$E_{m,n} = (E_m - E_n) \; = \; \hbar\omega_{m,n}$$

and switching to $\langle bra|ket\rangle$ notation resulting in

$$\begin{aligned}\sum_n a_n(t) \langle\psi_m| \hat{H}' |\psi_n\rangle e^{i\omega_{m,n}t} &= -\frac{\hbar}{i}\sum_n \dot{a}_n(t) \langle\psi_m|\psi_n\rangle e^{i\omega_{m,n}t} \\ \sum_n a_n(t)\hat{H}'_{m,n}e^{i\omega_{m,n}t} &= -\frac{\hbar}{i}\dot{a}_m(t). \qquad (3.7)\end{aligned}$$

The last step requires the orthogonality relation $\langle\psi_m|\psi_n\rangle = \delta_{mn}$ and introduces a shortened notation for the matrix element $\hat{H}'_{m,n} = \langle\psi_m| \hat{H}' |\psi_n\rangle$. equation 3.7 provides a set of linear differential equations from which a_n can be obtained.

However, the infrared relevant first order term can also be obtained by assuming the perturbation $\hat{H}'$ to be weak and short lived. This implies that the expansion coefficients at time t can be approximated by their value at $t = 0$, analog to a zeroth order Taylor expansion. Further, the system is assumed to be in state $\psi(x, t = 0) = \psi_s(x)$. The $\psi_s(x)$ state has all zero a_n coefficients except for $a_s = 1$. This simplifies equation 3.7 to

$$\dot{a}_m(t) \;=\; -\frac{i}{\hbar}\hat{H}'_{m,s}e^{i\omega_{m,s}t}. \qquad (3.8)$$

For time dependent perturbations $\hat{H}'(t) = \hat{H}'e^{-i\omega t}$ with frequency ω, equation 3.8 changes slightly into

$$\dot{a}_m(t) = -\frac{i}{\hbar}\hat{H}'_{m,s}e^{i(\omega_{m,s}-\omega)t}. \qquad (3.9)$$

Equation 3.9 can be integrated on the interval $t' = [0..t]$ under the assumption that the final state $|m\rangle$ is unoccupied at $t = 0$, $a_m(t = 0) = 0$. The resulting coefficient equation is

$$a_m(t) = \frac{1}{\hbar}\hat{H}'_{m,s}\frac{1 - e^{i(\omega_{m,s}-\omega)t}}{(\omega_{m,s} - \omega)}. \qquad (3.10)$$

The last result is useful for calculating the transition probability

$$p_{m,s}(t) = |a_m(t)|^2 = \frac{2}{\hbar}|\hat{H}'_{m,s}|^2\frac{1 - cos((\omega_{m,s} - \omega)t)}{(\omega_{m,s} - \omega)^2} \qquad (3.11)$$

from state $|s\rangle$ to state $|m\rangle$ via light absorption or emission. This in itself is a very interesting result, as it directly relates the transition probability to the square of the transition matrix elements. Thus, understanding the factors that influence transitions requires further information about the perturbation. One example for these perturbations is the electric field $\vec{E}(x,t) = \vec{E}_0(x)e^{-i\omega t}$ of photons which interact with the electronic dipole μ of the molecule. A simple interaction Hamiltonian

$$\hat{H}' = \vec{\mu} \cdot \vec{E}(x,t) \tag{3.12}$$

can be defined for this interaction. The coupling will be maximal when field and dipole moment are aligned parallel and zero for perpendicular orientations. Inserting equation 3.12 into 3.11 results in

$$p_{m,s}(t) = \frac{4}{\hbar^2} | \langle \psi_m | \, \vec{\mu} \, | \psi_s \rangle \, |^2 |\vec{E}_0|^2 cos^2(\theta_{\mu,E_0}) \frac{1 - cos((\omega_{m,s} - \omega)t)}{(\omega_{m,s} - \omega)^2}. \tag{3.13}$$

At this point the transition to the experiment can be made. Equation 3.13 directly relates the transition probability, or intensity $I_{m,s}$, to the dipole moment operator $\vec{\mu} = \sum_i e_i \vec{q}_i$. In this representation, e_i is the effective charge at atom i and $\vec{q}_i$ is the vector of atom i to the center of mass of the system. Choosing $\vec{q}_i$ this way creates a well defined point of reference even for charged systems, see chapter 3.1.3. The transition intensity $I_{m,s}$ from state $|s\rangle$ to state $|m\rangle$ can be expressed in relation to equation 3.13 as

$$I_{m,s} \propto \left([\hat{\mu}_x]^2_{m,s} + [\hat{\mu}_y]^2_{m,s} + [\hat{\mu}_z]^2_{m,s} \right), \tag{3.14}$$

where

$$[\hat{\mu}_j]_{m,s} = \langle \psi_m | \, \hat{\mu}_j \, | \psi_s \rangle \, . \tag{3.15}$$

Based on this relation for the intensities, transition rules can be obtained for the harmonic case by Taylor expanding the dipole moment operator with respect to the normal coordinates Q_i as

$$\vec{\mu} \approx \vec{\mu}_0 + \sum_{i=1}^{3N-6} \left(\frac{\partial \vec{\mu}}{\partial Q_i} \right)_{\mu_0} Q_i. \tag{3.16}$$

The harmonic approximation results in a finite Taylor series which is convenient but not required if generalized normal coordinates are available. The expression can be inserted back into equation 3.15 which finally leads to the important connection between the change of the dipole moment and vibrational intensities through

$$\begin{aligned}
[\hat{\mu}_j]_{m,s} &= \langle\psi_m|\,\hat{\mu}_j\,|\psi_s\rangle \\
&\approx \langle\psi_m|\,\mu_0 + \sum_{i=1}^{3N-6}\left(\frac{\partial\mu}{\partial Q_i}\right)_{\mu_0} Q_i\,|\psi_s\rangle \\
&= \mu_0\,\underbrace{\langle\psi_m|\psi_s\rangle}_{=0} + \sum_{i=1}^{3N-6}\langle\psi_m|\left(\frac{\partial\mu}{\partial Q_i}\right)_{\mu_0} Q_i\,|\psi_s\rangle \\
&= \sum_{i=1}^{3N-6}\langle\psi_m|\left(\frac{\partial\mu}{\partial Q_i}\right)_{\mu_0} Q_i\,|\psi_s\rangle\,. \qquad (3.17)
\end{aligned}$$

From this equation the important transition condition

$$\langle\psi_m|\left(\frac{\partial\mu}{\partial Q_i}Q_i\right)_{\mu_0}|\psi_s\rangle \neq 0 \qquad (3.18)$$

can be derived. The condition explicitly states that the molecular dipole moment of a given molecule must change with respect to the normal coordinates for allowed transitions from $|s\rangle$ to $|m\rangle$. Of course these statements only hold for the truncated Taylor expansion in equation 3.16 and thus only in harmonic approximation. However, even for the anharmonic case equation 3.18 is expected to be the dominating term in the transition probability and therefore also the transition intensities. Obtaining the required normal coordinates is highly non trivial especially for large systems and anharmonic contributions beyond the truncated Taylor expansion, see section 3.1.5 for further discussion. An additional feature of equation 3.14 is its dependence on all absolute square value of the transition dipole moment. In this formulation, forbidden transitions can exist in either one, two or three dimensions which allows experiments to probe molecules using polarized light. Individual contributions to the total intensity can be probed this way.

3.1.2 Normal Mode Analysis

The Normal Mode Analysis (NMA) provides a framework for obtaining the normal coordinates Q_i in equation 3.16 as well as harmonic frequencies [11, 15, 58]. In this framework the nuclear motion is approximated by harmonic potentials. The motion of the nuclei is described using Newton mechanics and Hook's law $m\ddot{x} = -kx$. The displacement of the atoms from their equilibrium positions is x, $\ddot{x}$ is the acceleration, m the mass and k the harmonic spring constant. Figure 3.1 shows a simple mechan-

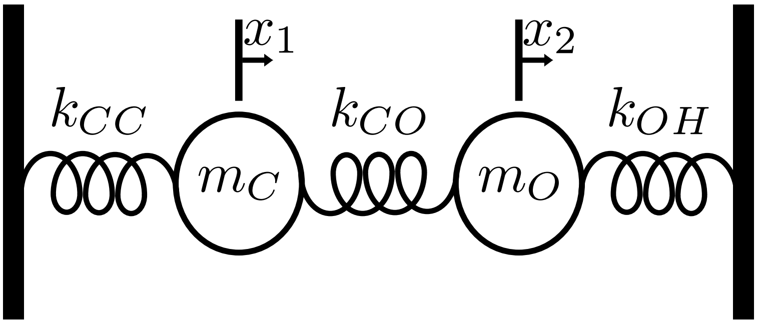

Figure 3.1: Example system with two degrees of freedom: k_i spring constants, m_i masses, x_i equilibrium positions

ical representation of a molecular system similar to a Carbon-(Carbon-Oxygen)-Hydrogen chain in a molecule, however with only two degrees of freedom. The coupled equations of motion for this system can be formulated as:

$$\begin{aligned} m_C\ddot{x}_1 &= -k_{CC}x_1 + k_{CO}(x_2 - x_1) \\ m_O\ddot{x}_2 &= -k_{CO}(x_2 - x_1) - k_{OH}x_2. \end{aligned} \tag{3.19}$$

In NMA, the assumption of an all atom collective motion leads to the following Ansatz:

$$\begin{aligned} x_1 &= A_1 e^{i\omega t} \\ x_2 &= A_2 e^{i\omega t}. \end{aligned} \tag{3.20}$$

Notice how the angular frequency ω is identical for both atoms but the amplitudes or weights A_i are different. Thus, a collective motion of all atoms is implied while not all atoms must participate, i.e. $A_j = 0$. Inserting Ansatz 3.20 into equation 3.19 leads to the set of coupled linear equations

$$\begin{array}{rcr} (k_{CC} + k_{CO} - m_c\omega^2)A_1 & -k_{CO} & A_2 = 0 \\ -k_{CO}A_1 & +(k_{CO} + k_{OH} - m_O\omega^2) & A_2 = 0. \end{array} \tag{3.21}$$

which is defined if the determinant

$$\begin{vmatrix} (k_{CC} + k_{CO} - m_c\omega^2) & -k_{CO} \\ -k_{CO} & +(k_{CO} + k_{OH} - m_O\omega^2) \end{vmatrix} = 0. \tag{3.22}$$

The determinant equation 3.22 can easily to transformed into an eigenvalue problem by substituting ω^2 with λ. For the sake of clarity, the spring constants k_{CC}, k_{CO}, k_{OH} and the masses m_O, m_C are reduced to the unit mass m as well as the arbitrary spring constant k using the following set of substitution rules:

$$k_{OH} = k;\ k_{CC} = 5k;\ k_{CO} = 6k;\ m_C = m;\ m_O = (3/4)m. \tag{3.23}$$

This leads to the two eigenvalues

$$\lambda_1 = \left(\frac{31 - \sqrt{433}}{6}\right)\frac{k}{m} \simeq 1.7\frac{k}{m} \tag{3.24}$$

$$\lambda_2 = \left(\frac{31 + \sqrt{433}}{6}\right)\frac{k}{m} \simeq 8.6\frac{k}{m} \tag{3.25}$$

and thus the frequencies

$$\omega_1 = \pm\sqrt{\lambda_1} \simeq 1.3\sqrt{\frac{k}{m}} \tag{3.26}$$

$$\omega_2 = \pm\sqrt{\lambda_2} \simeq 2.9\sqrt{\frac{k}{m}} \tag{3.27}$$

The set of eiqenvalues now leads to a set of eigenvectors by substituting ω_1 and ω_2 back into equation 3.21

$$\begin{aligned} \frac{A_1}{A_2} &= \tfrac{k_{CO}}{(k_{CC}+k_{CO}-m_c\omega_1^2)} &\approx -0.8 \\ \frac{A_1}{A_2} &= \tfrac{(k_{CO}+k_{OH}-m_O\omega_2^2)}{k_{CO}} &\approx 0.9 \end{aligned} \tag{3.28}$$

and therefore

$$\vec{v}_1 = \begin{pmatrix} -0.8 \\ 1 \end{pmatrix};\ \vec{v}_2 = \begin{pmatrix} 0.9 \\ 1 \end{pmatrix}. \tag{3.29}$$

The simple two DOF system discussed above already shows how amplitudes can only be obtained relative to each other in NMA and not in absolute values. Therefore, $A_2 = 1$ was arbitrarily chosen. It also illustrates how harmonic normal modes can be projected back onto the molecular structure via the original Ansatz of equation 3.20 using eigenfrequencies ω_i and eigenvectors $\vec{v}_i$. This allows for a visual inspection of the different normal modes as well as the assignment of spectral bands to localized groups of atoms. This projection results in a phasic motion (v_2) and an antiphasic motion (v_1) of the two atoms in figure 3.1. In the reference frame of normal coordinates, there is no energy transfer from these two motions as there are no anharmonic off diagonal coupling elements included.

3.1.3 Dipole Moments in QM/MM Simulations

An alternative approach towards calculating vibrational spectra is the direct analysis of system dynamics through time series analysis. In the previous chapters, the expansion of the wave function in eigenfunctions $\psi_n(x)$

$$\psi(x,t) = \sum_n a_n(t)\psi_n(x)e^{-iE_nt/\hbar}$$

was introduced and explicitly solved for the expansion coefficients $a_n(t)$, see equation 3.10. This is not practical for larger systems, as the vibrational wave function is computationally not accessible. In order to derive a more feasible simulation scheme, the coefficients a_n are related directly to the time series of the time depended Schrödinger equation solution. This can be archived by rewriting the Fourier transform of the wave function time correlation function [60]

$$p(\omega) = (2\pi)^{-1}\int_{-\infty}^{\infty} \langle\psi(0)|\psi(t)\rangle\, e^{i\omega t}dt$$

as

$$\begin{aligned} p(\omega) &= (2\pi)^{-1}\iint_{-\infty}^{\infty}\left(\sum_m c_m\psi_m(x)e^{-i\omega_n t}\right)^*\left(\sum_n c_n\psi_n(x)e^{-i\omega_n t}\right)dxe^{i\omega t}dt \\ &= (2\pi)^{-1}\int_{-\infty}^{\infty}\sum_{m,n} c_m^* c_n\delta_{mn}e^{i(\omega-\omega_n)t}dt \\ &= \sum_m |c_m|^2\delta(\omega-\omega_m) \end{aligned} \tag{3.30}$$

using the orthogonality

$$\langle\psi_m|\psi_n\rangle = \delta_{mn} \tag{3.31}$$

and the relation

$$\int_{-\infty}^{\infty} e^{i(\omega-\omega_n)t}dt = 2\pi\delta(\omega-\omega_n). \tag{3.32}$$

Here, equation 3.30 is identical to equation 3.11. Thus the spectrum of a quantum mechanical system can directly be obtained from the system dynamics without the need of Normal Mode Analysis. Instead of relying purely on a harmonic approximation of the underlying energy landscape, the system is allowed to relax and explore the energy landscape at finite temperature. This intrinsically includes anharmonic effects and temperature. Unfortunately, the time dependent Schrödinger equation can currently not be solved for system sizes of interest in this thesis. Instead, the cheaper QM/MM simulation scheme is used to propagate the wave function in time.

In this section, the dipole moment operator is applied to the QM/MM wave function and the resulting dipole moment time series is related to the spectrum via the Wiener−Khinchin Theorem. By choosing the dipole moment as the observable, the IR activity criterion from equation 3.18 is explicitly met as only dipole active modes can be identified. Modes which do not change the dipole moment will not contribute to the dipole moment time series and thus will not be resolved in the dipole time series analysis.

First, the concept of dipole moment analysis is introduced and sequentially extended towards charged molecules. The concept of dipole moments arises from the multipole expansion of the Coulomb potential. In the multipole expansion, the dipole moment $\vec{p}$ of a continuous charge distribution $\rho(\vec{r_0})$ in the volume V is defined as

$$\vec{p} = \int_V \rho(\vec{r_0})(\vec{r_0} - \vec{r}_{ref})d^3r_0. \tag{3.33}$$

In addition to the dependence of the dipole moment on the charge density ρ and the position r, the dipole moment also depends on a reference point $\vec{r}_{ref}$. It can easily be shown that this explicit dependence only holds for charged systems. For systems with a neutral overall charge $Q = \int_V \rho(\vec{r_0})d^3r_0 = 0$, the explicit dependence vanishes as

$$\begin{aligned}\vec{p} &= \int_V \rho(\vec{r_0})(\vec{r_0} - \vec{r}_{ref})d^3r_0 \\ &= \int_V \rho(\vec{r_0})\vec{r_0}d^3r_0 - \int_V \rho(\vec{r_0})\vec{r}_{ref}d^3r_0 \\ &= \int_V \rho(\vec{r_0})\vec{r_0}d^3r_0 - Q\cdot\vec{r}_{ref} \\ &= \int_V \rho(\vec{r_0})\vec{r_0}d^3r_0\end{aligned} \tag{3.34}$$

However, the analysis of neutral systems is not sufficient for the analysis of neither green fluorescent nor photoactive yellow protein as both systems have charged active sites. It is therefore desirable to find a good reference point. There are a number of possible candidates for reference points and intuitively, the origin seems like a good choice. With $r_{ref} = (0,0,0)$, the second term of the right hand site in equation 3.34

vanishes. However, this is not sufficient as the integral $\int_V \rho(\vec{r}_0)\vec{r}_0 d^3r_0$ will not only depend on position differences but also on the absolute position of the system for $\int_V \rho(\vec{r}_0)d^3r_0 \neq 0$. In this case the dipole moment will no longer be translation or rotation invariant as global motion changes the absolute distance towards the origin. As translational and rotational invariance is desired, the center of mass

$$r_{COM} = \frac{\sum_{n=1}^{N} m_n r_n}{\sum_{n=1}^{N} m_n} \tag{3.35}$$

can be chosen as reference point for all dipole time series. This effectively shifts the problem of reference into the internal coordinate frame of the molecule. A reference bias in the spectrum can only be reduced and will likely not be excluded this way. To approach this problem, the center of reference is explicitly included and monitored during all simulations conducted as part of this thesis. Additionally, the spectral contribution of r_{COM} motion is calculated in analogy to the dipole time series analysis. This enables band specific bias estimation which is expected to become less important for increasing system size as more atoms are included in the r_{COM} average.

Choosing a proper reference point is not the only obstacle when applying dipole time series analysis (DTSA) to large QM/MM simulation trajectories. The large computational cost of quantum calculations in the QM/MM scheme effectively limits the number of vibrations that can be identified. For medium sized QM systems, 1000-1500 fs trajectory lengths are feasible which also equals about 1500 samples of the underlying dynamics. It therefore is no surprise that not all $3N$-6 modes for the entire simulation box can be determined. For GFP and PYP the total number of DOFs surpasses 10^5 and can no longer be handled at the QM/MM level.

Figure 3.2 illustrates three ways of calculating the dipole moments for the QM region in QM/MM simulations. Figure 3.2c) represents the simplest solution which only includes static MM charges for the QM region. This ensures compatibility with the MM charges as there are no contributions from induced dipole moments. There is also no explicit dependence on the electron dynamics which is also missing in the MM force field. Figure 3.2a) extends this simple picture by refitting the electron generated electrostatic potential at each time step. This can be done in analogy to the original parametrization of the MM force field [33] which ensures consistency with the dipole contribution from the MM region. Both QM methods for calculating the dipole moment are in good agreement with the MM description. Thus

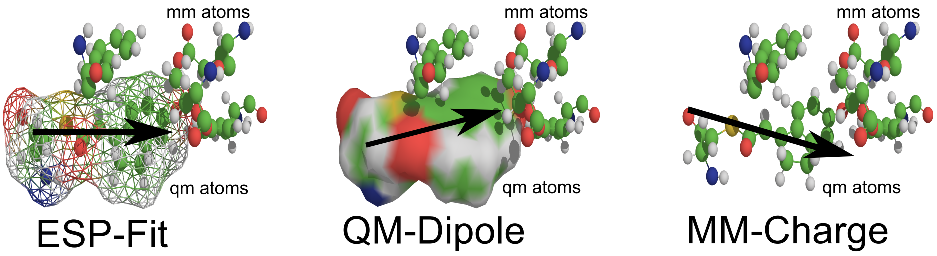

Figure 3.2: QM/MM dipole moment simulation schemes with implicit dependence on MM atoms; left: dynamic fitting of electrostatic qm potential onto nuclear centers; middle: dipole moment operator applied to qm wave function; right: static charge scheme using mm charges for qm atoms

frequency offsets should be comparable. This is not the case for figure 3.2b). Here, the QM wave function with implicit MM atom contributions is used to calculate the dipole moment expectation value. This quantity will most likely experience different frequency shifts and cannot easily be added to the MM contribution of the dipole moment. Taken for the QM region alone, 3.2b) is the most accurate treatment of the dipole moment.

The sampling problem reduces the need for a coherent QM/MM description of the dipole moment as the trajectory length is insufficient to include the MM contribution directly. Additionally, the force field approximation in the MM region is most reasonable for vibrational modes below $k_B T/h \approx 200\ cm^{-1}$ [5]. However, the infrared region of interest lies above 1000 cm^{-1} and is therefore not described sufficiently at the MM level of theory. Therefore, the fast degrees of freedom in form of bond vibrations are excluded from the MM region via constraints [31]. The dipole time series analysis is therefore limited to only the quantum region of the QM/MM simulation. However, the constrained MM region is still present implicitly through point charges in the quantum Hamiltonian. This procedure effectively reduces the number of modes that need to be fitted and additionally also excludes the low quality MM modes. For this reason, short QM/MM simulations mostly benefit from the scheme in figure 3.2b).

At this point a procedure for calculating dipole moment time series from QM/MM simulations has been introduced. This time series is now related to the power spectral density using the Wiener–Khinchin theorem.

3.1.4 Wiener-Khinchin Theorem

The Wiener-Khinchin Theorem is a fundamental part of signal analysis. It relates the time autocorrelation function $R(\tau)$ of a wide sense stationary (WSS) signal $\mu(t)$ to its power spectral density $S(\omega)$. However, it was also shown to hold for deterministic signals as well [16]. The Wiener-Khinchin Theom will be derived in analogy to derivation by Leon Cohen [16] with the extension of equation block 3.41 for clarity. For the case of random, zero mean, stationary time dependent signals

$$\mu(t) = \begin{cases} \mu(t) & \text{|t|<T} \\ 0 & \text{|t|>T.} \end{cases} \tag{3.36}$$

the signal Fourier transform $F(\omega)$ and inverse Fourier transform can be defined as

$$F(\omega) = \int_{-\infty}^{\infty} \mu(t)e^{-i\omega t}dt \tag{3.37}$$

$$\mu(t) = (2\pi)^{-1}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega. \tag{3.38}$$

The process autocorrelation function $R(t,\tau)$ is defined as the ensemble average $E[]$ of the signal as

$$R(\tau) = E[\mu^*(t)\mu(t+\tau)]. \tag{3.39}$$

Under the assumption of ergodicity, the ensemble averaged autocorrelation function equals the time autocorrelation function.

The power spectral density $S(\omega)$ is defined as

$$S(\omega) = \lim_{T\to\infty}\frac{1}{2T}E[|F_T(\omega)|^2] \tag{3.40}$$

and can be rewritten to result the Wiener-Khinchin theorem since

$$\begin{aligned} S(\omega) &= \lim_{T\to\infty} \frac{1}{2T} E[|F_T(\omega)|^2] \\ &= \lim_{T\to\infty} \frac{1}{2T} E\left[\int_{-T}^{T} \mu(t) e^{-i\omega t} dt\right]^2 \\ &= \lim_{T\to\infty} \frac{1}{2T} E\left[\iint_{-T}^{T} \mu^*(\tau)\mu(t) e^{-i\omega(t-\tau)} dt d\tau\right] \\ &= \lim_{T\to\infty} \frac{1}{2T} \iint_{-T}^{T} E[\mu(\tau)\mu(t)] e^{-i\omega(t-\tau)} dt d\tau \\ &= \lim_{T\to\infty} \frac{1}{2T} \iint_{-T}^{T} dt R(t-\tau) e^{-i\omega(t-\tau)} dt d\tau \\ &= \underbrace{\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} dt}_{=1} \int_{-\infty}^{\infty} R(\tau') e^{-i\omega\tau'} d\tau'. \end{aligned} \tag{3.41}$$

Here, the autocorrelation function $R(\tau)$ does no longer depend on the time t but rather on the lag window length τ and therefore the time integral drops out.

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau. \tag{3.42}$$

Equation 3.42 is the Wiener-Khinchin theorem. It enables direct calculation of power spectra from time autocorrelation functions. In strict mathematical terms, the autocorrelation function detour is not necessarily required to calculate power spectra, as can be seen from equation 3.40. The time autocorrelation function is smooth, time independent and can also exist for signals which cannot be Fourier transformed directly. This is the case for signals which do not fulfill the condition of finite energy. Additionally, autocorrelation based spectral estimation has lower variance than direct spectra, which is desirable [39]. The reason for discussing the Wiener-Khinchin theorem will become more imminent when introducing the maximum entropy spectral estimation. Here, the autocorrelation function for short signals is extended under the boundary condition of maximum entropy.

3.1.5 Assumptions, Approximations and Limitations

The first section of this chapter promised a method for calculating infrared spectra from QM/MM simulations providing the following feature set:

- **Native support for anharmonicity and temperature effects**
- **High frequency resolution for difference spectra**
- **Natural protein environment is included**
- **Covers both the ground as well as the excited state**

These features have all been provided, except for the high resolution in frequency space which is discussed in chapter 3.2. However, the efficient implementation of the described method requires several implicit assumptions and approximations which are important in order to understand the limitations of this approach. Figure 3.3 illustrates the underlying physics of the three different methods discussed so far. The three vibrational degrees of freedom $\vec{v}_i$ are shown for a single water molecule. Additionally, the classical potential functions as well as the vibronic states s_i, s_i' and s_i'' of the quantum wave function are drawn. In this picture, upon infrared light absorption of energy $E = \hbar\omega_i$ (purple) the discrete vibronic quantum number s_i is increased and the vibronic wave function shifts to a higher energy level. This process is experimentally observable. The transition probabilities as derived in section 3.1.1 determine the intensity of this light absorption process. At first sight, the QM region in a QM/MM simulation might seem to cover the physics of this process. However, this is not completely true as the motion of the nuclei is assumed to be classical in QM/MM. This assumption allows the nuclei and therefore wave function to be propagated using the MM propagation scheme and thus Newton mechanics. The

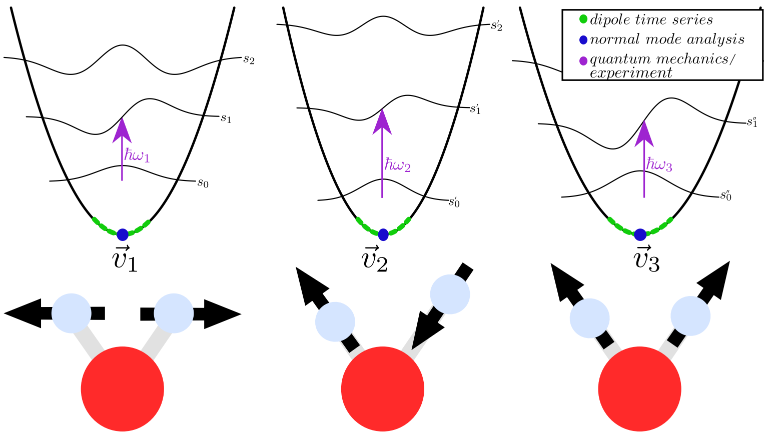

Figure 3.3: Collective vibrational modes of water; s_i: vibrational energy levels of the wave function; green dots: range of QM/MM simulation sampling; blue dot: idealized location of the NMA structure

consequences of this approximation are severe as the vibrational energy states are approximated classically and the energy dissipation among the vibrational modes is no longer quantized. Vibrational degrees of freedom can have continuous energy levels as opposed to discrete quantum levels in the experiment. With respect to this energy partition issue, the dipole time series analysis (sampled region shown in green) cannot be expected to significantly improve on the Normal Mode Analysis results.

However, the dipole time series analysis does have other clear advantages over NMA. Figure 3.3 illustrates the strong dependence of Normal Mode Analysis on finding a well energy minimized structure. The blue dot represents the idealized single starting point configuration for Normal Mode Analysis. Obtaining this beforehand unknown point is not trivial especially for large QM systems in a protein environment such as the GFP active pocket. With dipole time series analysis, the close proximity of this point is also taken into account through sampling of the dynamics. A more severe limitation of Normal Mode Analysis is the assumption of an harmonic potential energy landscape around the equilibrium position. In physical terms this can be seen as a very low order Taylor expansion to the potential energy landscape. Deviations from this set of harmonic orthogonal potentials are not allowed and energy transfer between different normal coordinates is not possible. This effectively biases the identified modes towards their closest harmonic equivalent. The time series analysis is more general as it identifies periodic contributions to a signal without making assumption about the underlying energy landscape.

3.2 Estimating Power Spectral Densities

3.2.1 Fourier Spectra

The relation between the power spectral density and the Fourier transform was derived in section 3.1.4. The Wiener-Khinchin Theorem was derived under the assumption of infinite sampling of a periodic signal $s(t)$. Therefore, an exact autocorrelation function $R(\tau) = E[s(t)s^*(t+\tau)]$ and infinite boundaries in the Fourier integral were assumed for the power spectral density

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau. \tag{3.43}$$

In real world applications such as simulated IR spectra of proteins, this assumption is not valid as only a finite set of sampling points is available. The high cost of performing QM/MM simulations severely shortens the available trajectory lengths to about 1000 sampling points. The transition between the infinite S_∞ and finite S_T time series on the interval $[-T..T]$ can be formulated as a multiplication of S_f with a rectangular window function

$$\varsigma_{rect}(t) = \begin{cases} 1 & \text{|t|≤|T|} \\ 0 & \text{|t|>|T|} \end{cases} \tag{3.44}$$

as

$$S_T = S_\infty * \varsigma_{rect}(t). \tag{3.45}$$

The window multiplication or convolution operation also affects the Fourier representation of the signal. The Fourier spectrum of the finite length signal is smoothened with decreasing window length and spectral resolution is lost. In the ideal case of infinite sampling, the window function has infinite width and its Fourier transform, the delta function, does not influence the spectrum. A second effect of finite window lengths is caused by the truncation of the maximum lag length τ. This results in an increase in variance of the autocorrelation function estimate for long lag times. The uneven variance of the estimate is due to the nature of the finite autocorrelation function in which short lag times τ are better sampled that long ones.

The undesirable variance can be reduced by introducing a lag dependent weighting function $\varsigma(\tau)$ into the autocorrelation function estimate $R(\tau)$. This method was introduced by R.B. Blackman and J.W. Tukey [7] in 1959. The Blackman and Tukey method describes the weighted power spectral density

$$S(\omega) = \int_{-\infty}^{\infty} \varsigma(\tau) R(\tau) e^{-i\bar{\omega}\tau} d\tau. \tag{3.46}$$

As a consequence, the weighted power spectrum effectively becomes the convolution of the true spectral estimate and the Fourier transform of the weighting window. This convolution can have severe effects on the spectrum as the positivity is not necessarily conserved and spectral amplitude can leak into neighboring peaks. Spectral positivity can be conserved by using a triangular window of the form

$$\varsigma_{triang}(t) = \begin{cases} \frac{2\tau}{L+1} & 1 \leq \tau \leq \frac{L+1}{2} \\ \frac{2(L-\tau+1)}{L+1} & \frac{L+1}{2} < \tau \leq L \end{cases} \tag{3.47}$$

at the cost of effectively reducing the information used to obtain the spectrum in half. Triangular windows are a simple but wasteful way of reducing spectral variance. Additionally, triangular windows increase the spectral bias through the mentioned sidelobe leakage of power as a result of the convolution procedure.

An alternative way of reducing spectral variance is the Bartlett method. The Bartlett method is simply an average over N statistically independent spectral estimates $S_i(\omega)$ as

$$S(\omega) = \frac{1}{N}\sum_{i=1}^{N} S_i(\omega). \tag{3.48}$$

This approach decreases the spectral resolution but also decreases the variance by a factor of $1/N$. It is therefore desirable to have high resolution spectral estimates $S_i(\omega)$ before averaging. The Bartlett method is especially appealing for QM/MM trajectories as many short trajectories are significantly cheaper to calculate than single long trajectories.

3.2.2 Maximum Entropy Method

The windowing methods discussed so far all assume the autocorrelation estimate to be zero outside the interval $[-2T..2T]$. However, this assumption can be lifted by extending the signal autocorrelation function under the condition of maximum entropy. This approach is called Maximum Entropy Method (MEM). The MEM generates additional information on the autocorrelation function which can then be used to calculate a Fourier spectrum. The extended autocorrelation function from approximated data has a higher spectral resolution but will not significantly decrease the spectral variance. Additionally, the extension of the autocorrelation function may introduce pseudo peaks into the spectrum due to over fitting for large extension ranges. MEM should therefore be considered a parametric method for spectral estimation in contrast to the parameter free Fourier spectrum. The parameter estimation and optimization is matter of section 3.3.

In this section, the basic mathematical concepts of the MEM according to the work of J.P. Burg [12] are introduced and connected to the all pole autoregressive (AR) filter in analogy to A. van den Bos [64] and D.A. Gray [26]. The definition of entropy

$$H_N = log(2\pi e)^{N/2} det\{R\}^{1/2} \tag{3.49}$$

in the MEM originates from statistical physics but is equally important in information theory where it can be seen as a measure of spectral uncertainty. Applied to a power spectrum, the spectral entropy is maximal when all frequencies are equiprobable which is defined as white noise. In the context of the MEM, the best spectral estimate is assumed to be the spectrum which has maximum entropy under the boundary condition of reproducing the N measured lags r_i of the signal autocorrelation function

$$R_N = \begin{pmatrix} r_0 & r_1 & \cdots & r_{N-1} \\ r_1 & r_0 & & \vdots \\ \vdots & & \ddots & r_1 \\ r_{N-1} & r_{N-2} & \cdots & r_0 \end{pmatrix}. \tag{3.50}$$

The MEM now extends R_N with the estimated lag r_N to

$$\overline{R}_{N+1} = \begin{pmatrix} r_0 & r_1 & \cdots & r_{N-1} & r_N \\ r_1 & r_0 & & & \vdots \\ \vdots & & \ddots & & r_1 \\ r_N & r_{N-1} & \cdots & & r_0 \end{pmatrix} \tag{3.51}$$

under the boundary condition of maximizing the estimated entropy

$$H_{N+1} = log(2\pi e)^{(N+1)/2} det\{\overline{R}\}^{1/2} \tag{3.52}$$

as

$$\frac{\partial H_{N+1}}{\partial r_N} = 0. \tag{3.53}$$

Conveniently, r_N was shown [64] to be obtainable from the solution of the determinant equation

$$det \begin{vmatrix} r_1 & r_0 & \cdots & r_{N-2} \\ r_2 & r_1 & & r_{N-3} \\ \vdots & & \ddots & \\ r_N & r_{N-1} & \cdots & r_1 \end{vmatrix} = 0. \tag{3.54}$$

Based on the $N+1$ autocorrelation estimate, the values for $N+2$, $N+3$, etc. can be calculated iteratively.

3.2.3 Burgs Method

In the previous section, the Maximum Entropy Method was introduced which allowed the calculation of power spectra by Fourier transforming an extended autocorrelation function.

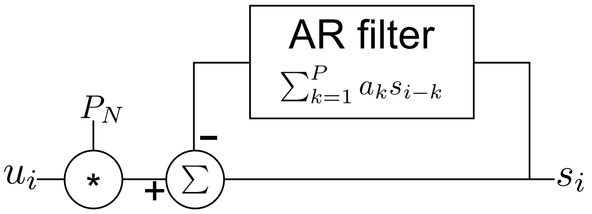

Figure 3.4: All pole autoregressive filter in time domain.

This is just one example of how the spectral resolution can be increased. An alternative way is to rewrite a signal s_i in terms of its last P samples and a random noise input u_i as

$$s_i = -\sum_{k=1}^{P} a_k s_{i-k} + P_N u_i. \tag{3.55}$$

This model is known as the all-pole model [47], see figure 3.4. The weights a_k are called filter coefficients. P_N is the gain or output power obtained from the resulting linear prediction model [26], or autoregressive filter. The frequency representation of this model can be written in terms of a z transform $\sum_{n=0}^{N} a_n z^{-n}$ as

$$S_{out}(z) = \frac{P_N}{[\sum_{n=0}^{P} a_n z^{-n}][\sum_{n=0}^{P} a_n^* z^n]}. \tag{3.56}$$

No Fourier transform is needed and the spectrum $S_{out}(z)$ is obtained in a P poles polynomial representation.

The work of J.P. Burg [10, 12, 51] extended this approach by relaying the power spectrum directly to the prediction error filter coefficients (p.e.f.c.s.). Burg decomposed the $N+1$ signal samples into an all pole autoregressive filter of order P. The measured signal s_i is approximated by the estimate $\overline{s_i}$ as

$$\overline{s}_i = -\sum_{k=1}^{P} a_k s_{i-k}. \tag{3.57}$$

A forward p.e.f.c.s. estimator $f_i(n) = s_i - \overline{s}_i$ is constructed for all P filter coefficients a_i as

$$\begin{aligned}
f_0(n) &= x_n \\
f_1(n) &= x_n + a_1 x_{n-1} \\
&\vdots \\
f_P(n) &= x_n + a_1 x_{n-1} + a_2 x_{n-2} + \cdots + a_P x_{n-P}. \qquad (3.58)
\end{aligned}$$

By reversion of time, a backward p.e.f.c.s. estimator $b_i(n)$ is constructed analogously as

$$\begin{aligned}
b_0(n) &= x_n \\
b_1(n) &= x_{n-1} + a_1 x_n \\
&\vdots \\
b_P(n) &= x_{n-P} + a_1 x_{n-P+1} + a_2 x_{n-P+1} + \cdots + a_P x_n. \qquad (3.59)
\end{aligned}$$

The filter coefficients a_i are then obtained by recursively minimizing the sum of squares for both forward and backward estimators as

$$RSS(P) = \sum_{n=P+1}^{N} (f_P^2(n) + b_P^2(n)) \qquad (3.60)$$

A more in depth discussion of the Levinson-Durbin recursion for calculating higher model orders can be found in the literature [12, 51, 62, 63]. Burgs method is a very powerful tool for calculating the AR filter coefficients a_i which also guarantees positivity of the calculated spectra.

3.2.4 Burg Alternatives for Autoregressive Filters

Burg's method is just one among many different ways of obtaining autoregressive filter coefficients a_i. Differences arise in how the prediction error is minimized and what algorithm is used to obtain the filter coefficients [25, 52]. A prominent alternative to the Burg method is the Yule-Walker method [78] which was originally applied to investigate periodicity in Wolfer's sunspot numbers. Instead of minimizing a forward and backward prediction error, the Yule-Walker approach tries to directly optimize all AR coefficients simultaneously in a least squares fashion [51] by minimizing

$$RSS_{yule}(P) = \sum_{n=-\infty}^{\infty} (x_n + a_1 x_{n-1} + \cdots + a_P x_{n-p})^2. \tag{3.61}$$

The Mathworks MATLAB [49] implementation of the MEM algorithm is equivalent to this Yule-Walker approach and the spectra are identical. Forward only and forward/backward least square methods also exist which minimize the prediction errors according to

$$RSS_{forward}(P) = \sum_{n=P+1}^{N} (x_n + a_1 x_{n-1} + \cdots + a_P x_{n-p})^2 \tag{3.62}$$

and

$$\begin{aligned} RSS_{forw.\&back.}(P) &= \sum_{n=P+1}^{N} (x_n + a_1 x_{n-1} + \cdots + a_P x_{n-p})^2 \\ &+ \sum_{n=1}^{N-K} (x_n + a_1 x_{n-1} + \cdots + a_P x_{n-p})^2 \end{aligned} \tag{3.63}$$

respectively. The main advantage of Burg's method over the Yule-Walker method is its efficient usage of information. The Yule-Walker algorithms uses N-P data points for each filter coefficient while Burg's method used N-1 for the first, N-2 for the second, N-P for the P'th filter coefficient. For large number of filter coefficients this is a severe limitation of the Yule-Walker and other least squares algorithms.

3.3 Simulation Scheme

A major challenge in preparation of the thesis was the development of a state of the art simulation scheme for calculating excited state anharmonic vibrational spectra. The steps performed for the calculation of anharmonic spectra are the following

First, a QM/MM simulation environment has to be set up for the states of interest. States can consist of different molecular conformations or excitation levels and are used to calculate the difference spectra. It is important to use the same QM method for all simulated states as each QM method will likely experience different frequency shifts. The shifts would otherwise severely influence the quality of the difference spectrum. In the classical MM region, all bond vibrations have to be constrained in order to remove their frequency contribution to the QM dipole moment time series.

The motivation for this is the large number of 3N-6 vibrational degrees of freedom. Only the smaller subset of vibrations in the QM region can be resolved from the small number of samples generated in QM/MM simulations. A suitable algorithm for constraining bonds is the LINCS algorithm by B. Hess et al. [31]. The QM/MM simulation system should be small enough to propagate it for around 2000-3000 time steps of 1 fs but shorter trajectories lengths are possible for very small QM regions. Longer trajectories will increase the quality of the spectra and increase the number of vibrational degrees of freedom which can be identified. During the Simulation, the coordinates of all QM atoms are recorded each time step together with the dipole moment vector from the QM wave function. A set of multiple simulation trajectories is required to reduce the spectral variance.

Second, the QM/MM dipole moment time series is analyzed to obtain the spectrum. The post processing of the coordinate and dipole information consists of removing the QM system center of mass motion as described in section 3.1.3. The Fourier and the parametric Burg method, as described in section 3.2.3, are used to calculate the spectrum. The number of AR parameters needs to be determined beforehand to reduce artifacts of line splitting and zombie peaks in the Burg spectrum. Therefore, a normal mode spectrum is calculated beforehand for the system of interest, see section 3.1.2. This normal mode spectrum is assumed to have a similar but different peak distribution and peak number compared to the spectrum generated from the QM/MM trajectory. Therefore, the NMA frequency distribution is used to generate a stationary signal composed of sinusoids in additive white noise. The initial phase of each sine functions corresponding to a NMA frequency is chosen randomly. The Burg method is applied to the generated spectrum and the optimal model order P is determined based on the quality of the identified peaks. Afterward, the Burg method of model order P is applied to the measured signal from the QM/MM simulation trajectories. A variance reduced averaged vibrational spectrum is calculated from multiple QM/MM trajectories as described in equation 3.48. Finally, the difference spectrum can be calculated from the averaged single state spectra by subtraction of the spectra.

3.4 Assumptions, Approximations and Limitations

The method presented in this chapter greatly improves the sampling of the energy landscape compared to the Normal Mode Analysis. It also removes the harmonic approximation and directly connects QM/MM simulation trajectories to spectral analysis. This makes experimental infrared spectra a new, valuable source of information about the quality of simulated protein dynamics. A simulated infrared spectrum which agrees with the experimental data greatly improves the confidence in the physical correctness of the simulation.

However, the method does make several approximations and assumptions about the system and its environment which inevitably lead to limitations in applicability. First, all frequency analysis methods discussed in section 3.2 require stationary signals. This explicitly limits the presented method to static processes. During the simulation, no transitions into other states may occur and large conformational changes must be avoided. Conformational changes can be described by simulating the initial and the final state in two separate simulations. The initial velocity distribution should already be converged to the target temperature and must not be coupled to a thermostat during the simulation. The rescaling of velocities in thermostats has unpredictable effects on the spectrum as it introduces discontinuities in the dipole moment time series. The spectra generated from time series should be averaged over multiple simulation trajectories to reduce artifacts of poor starting conformations. The number of required trajectories is not well defined beforehand but can be estimated by monitoring the variance of the identified peaks as a function of included trajectories.

Second, the nuclear motion is assumed to be classical in QM/MM simulations. This introduces errors in the power distribution as the energy stored in each mode is not discrete but continuous. This should in principle also effect normal mode spectra. The common frequency scaling of harmonic normal modes to correct for anharmonic effects [34] is not required for time series based spectra. Normal mode frequencies are too high because the harmonic approximation to the true potential is too narrow. Therefore, the resulting spectrum is commonly scaled by a factor of around of 0.9 for DFT based spectra. Time series based spectra do not make harmonic approximations to the underlying potential and identify periodic motion directly which makes the scaling obsolete.

Third, systematic errors in the simulation setup may greatly reduce the quality of the calculated spectra. QM/MM simulations usually describe electrostatic interactions using a cutoff scheme which is known to produce artifacts on long time scales. For short trajectory lengths the error is assumed to be small. The motion of bond vibrations is not well approximated by harmonic MM bond potentials which is why the bond motion is constrained to the equilibrium value. This also reduces artifacts in the bond vibrational part of the final spectrum. However, errors in the QM region itself are likely more severe than in the MM region. The excited state simulations performed using the CASSCF method with a reduced active space may introduce further errors. It was not investigated how the reduction of the active space effects the dynamics of the dipole moment. Only the static case was considered which showed a considerable change in the overall dipole moment for different reduced active spaces.

4 Applications

4.1 Green Fluorescent Protein

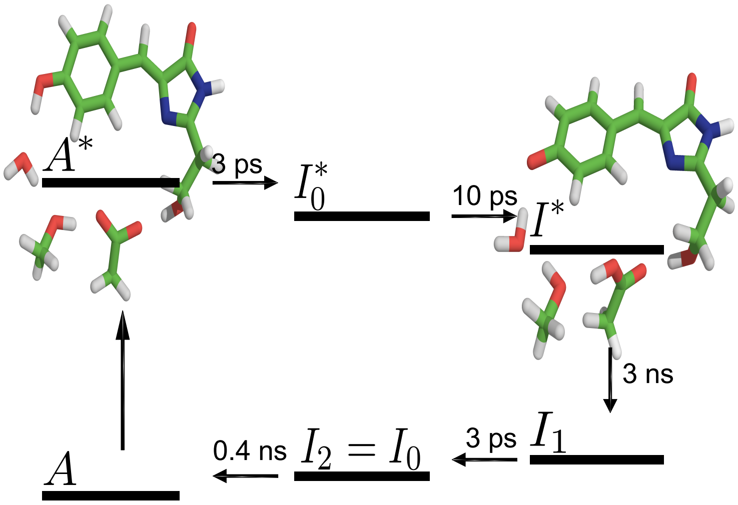

Figure 4.1: Proposed sequential GFP photo cycle [20]; A) ground state; A^*) first excited state; I_0^* and I^*) excited state intermediates; I_1 and I_0) ground state intermediates.

The Green Fluorescent Protein (GFP) is a barrel shaped protein with a photoactive chromophore in its center. The protein absorbs light in the UV range and emits green light in the visual part of the spectrum. GFP has become a widely used tool for fluorescent labeling as the protein can be fused into other proteins by inserting the GFP gene into the host DNA. The protein and its chromophore self-assemble autocatalytically. The chromophore is formed from the tripeptide Ser65-Tyr66-Gly67 without enzymatic assistance[50]. GFP is also an interesting candidate for super resolution imaging which is no longer bound to the Abbe diffraction limit [27, 30] However, despite the existing experimental tools to produce new GFP variants, the precise molecular mechanism of the GFP active pocket after photo excitation is still an active matter of debate [20, 65, 66]. Figure 4.1 illustrates the sequential model proposed by M.L. Groot et al. and shows six proposed stages of the photo cycle. Simulated protein dynamics may help to identify the proposed states by relating calculated difference spectra to the experimental data.

4.1.1 GFP Model Order Prediction

Frequency analysis is a tricky business with considerable pitfalls to consider. Section 3.3 introduced the simulation scheme which is used to generate the data and section 3.2 introduced three main spectral analysis methods. These are the Fourier transform, the Burg autoregressive filter analysis and the Yule-Walker or Maximum Entropy Method. The predictive quality with respect to the GFP chromophore hydrogen bond network was investigated.

The main motivation of this analysis is to overcome the uncertainty principle of the Fourier transform which relates the spectral resolution $\Delta\omega$ to the trajectory length ΔT as

$$\Delta\omega_{max} \leq \frac{1}{\Delta T}. \tag{4.1}$$

However, the higher resolution does come at the price of introducing autoregressive parameters p which may cause severe artifacts if not chosen carefully. These artifacts are line splitting, false positives, initial phase dependence of the peak location and high variance in the peak heights. Therefore, the test system was carefully designed to help identify theses artifacts on data similar to the expected simulation trajectory data. First, the normal mode frequencies ω_i were used as input for generating the signals $s_i(t)$ of length L as

$$s_i(t) = Asin(\omega_i t + \varphi). \tag{4.2}$$

The amplitude A was set to one and the phase φ was drawn randomly from the interval $[-2\pi..2\pi]$. The unit amplitudes distribute the power equally over the signal and therefore all spectral peaks are expected to have the same height. The random phase recreates the actual simulation conditions. The final signal $S(t)$ is the sum of all normal mode contributions $s_i(t)$ and additive white noise $\xi(t)$ as

$$S(t) = \sum_{j}^{3J-6} s_i(t) + \xi(t). \tag{4.3}$$

The white noise was generated with a signal to noise ratio of $s2n = 20$ dB according to the root mean square deviations (rms) of signal S and noise ξ

$$s2n = 20log_{10}\left(\frac{rms(\sum_j^{3J-6} s_i(t))}{rms(\xi(t))}\right) \tag{4.4}$$

which corresponds to a signal which is 10 times stronger than the additive noise. This model does not include non additive noise from a thermostat or bond vibrations in the MM region. The simulation setup was carefully designed to avoid such non additive contributions by constraining MM bond vibrations and deactivating the thermostat in the QM region. However, non additive noise can, at the present stage of development, not fully be excluded.

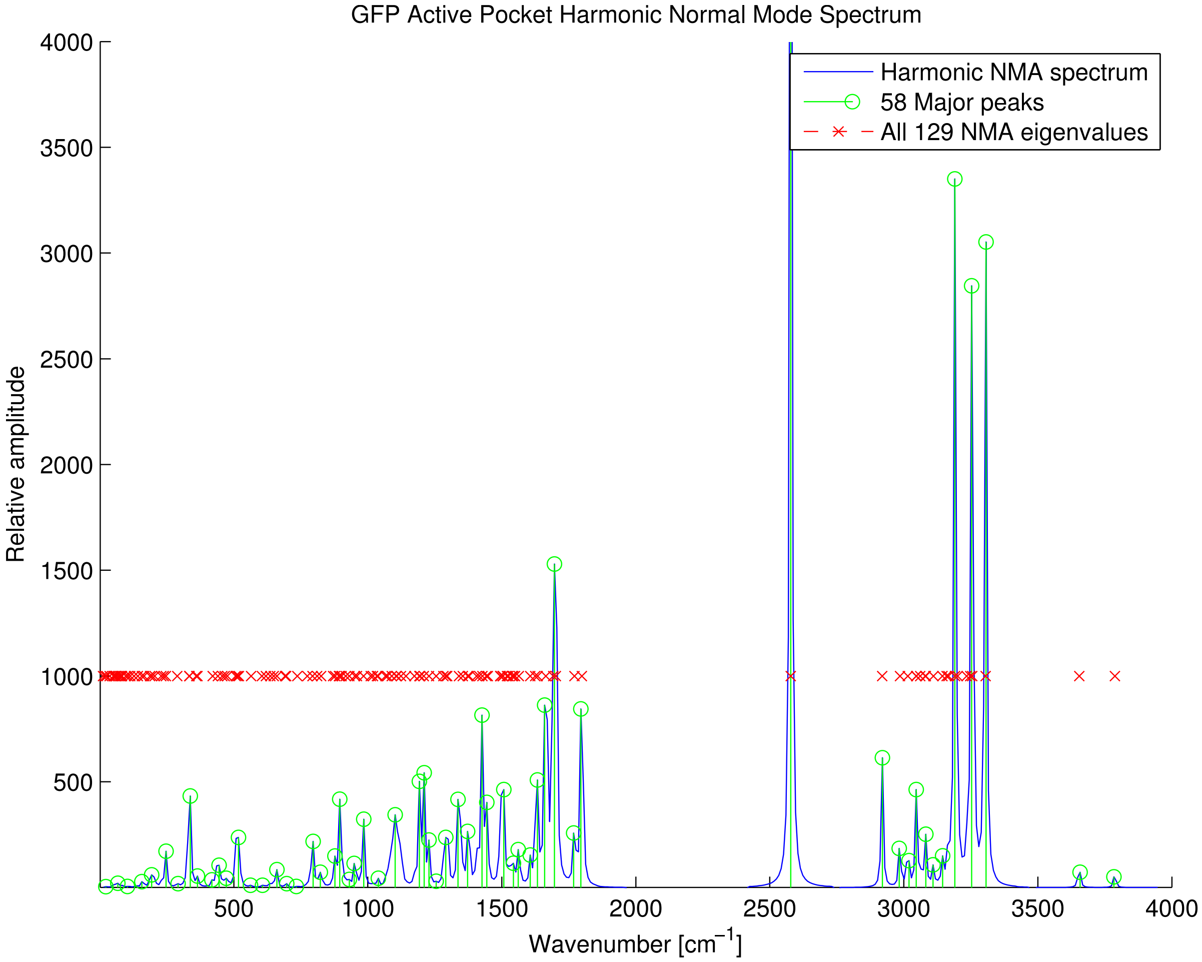

Figure 4.2: normal mode spectrum of the GFP chromophore hydrogen bond network; (blue) NMA spectrum; (green) 58 major frequency peaks; (red) full NMA frequency density

The GFP chromophore hydrogen bond network has $J = 45$ atoms which results in

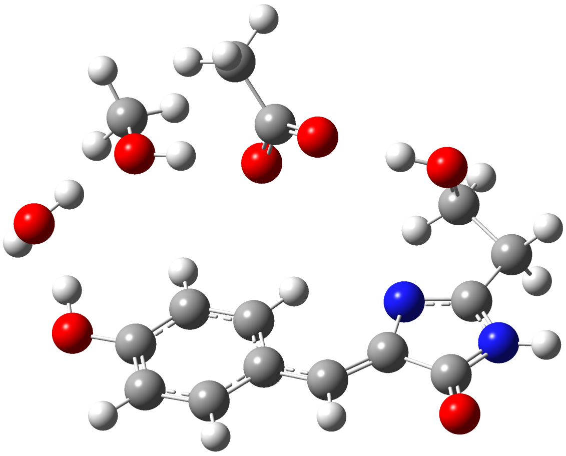

Figure 4.3: Stucture of GFP hydrogen bond network for Normal Mode Analysis

$3J - 6 = 129$ degrees of freedom and an 129 possible frequencies. Figure 4.2 shows the frequency density and the 58 main frequency peaks of the harmonic normal mode spectrum in vacuum. As the cost of GFP QM/MM simulations is very high, a simulation trajectory length of $L = 1000$ fs was used to generate the GFP test signal. The result is shown in figure 4.4.

The spectral information of the GFP test system is only poorly recovered from the 1000 fs trajectory. The signal was generated such that all peak heights are expected to be equal in the unscaled Fourier spectrum. For the Burg spectrum, the area under the peaks is expected to be constant. Neither is the case as can be seen in figures 4.4 and 4.5. The maximum resolution estimate from equation 4.1 for a 1000 fs trajectory is $\Delta\omega \approx 34\ cm^{-1}$ and 500 autoregressive parameters are not sufficient to reduce this below the desired $< 10\ cm^{-1}$ experimental resolution. The GFP model system is expected to reach single wavenumber resolution at roughly 32000 fs but already 3000 fs yield usable frequency resolution. The 3000 fs example is shown in figure 4.6 and 4.7. The data shown in figure 4.4 helps to put the quality of spectra calculated in section 4.1.3 into perspective. Due to the large size of the GFP chromophore system, excited state QM/MM calculations of 3000 fs are out of reach given the computational resources present at the time of writing this thesis. The presented 1000 fs trajectories have very poor frequency resolution and should only be seen as proof of principle for the developed simulation method.

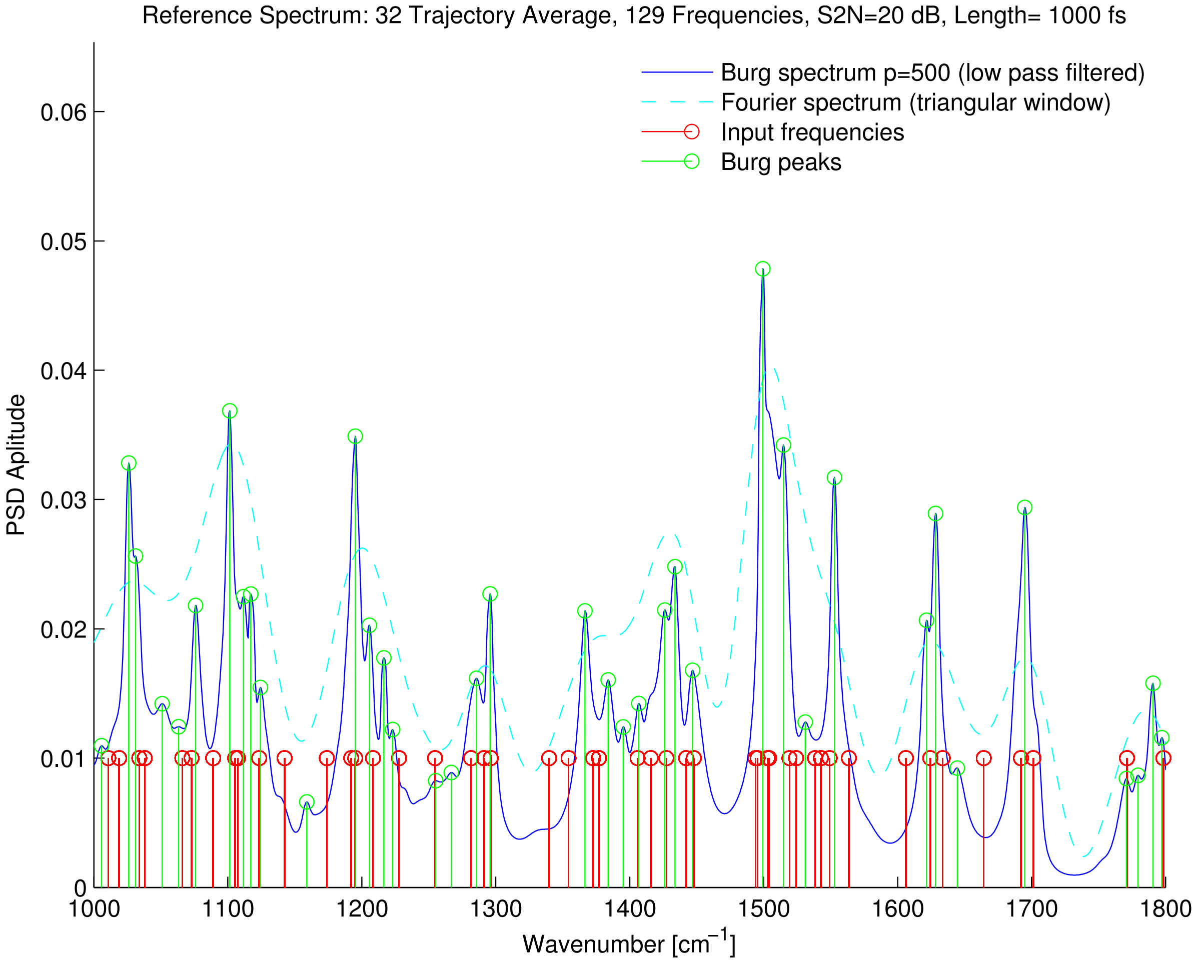

Figure 4.4: Portion of GFP like 1000 fs spectrum with 129 NMA frequencies (red); Burg peaks (green); Fourier spectrum (cyan); Burg spectrum (blue); peak heights are not well recovered, target frequencies are not matched.

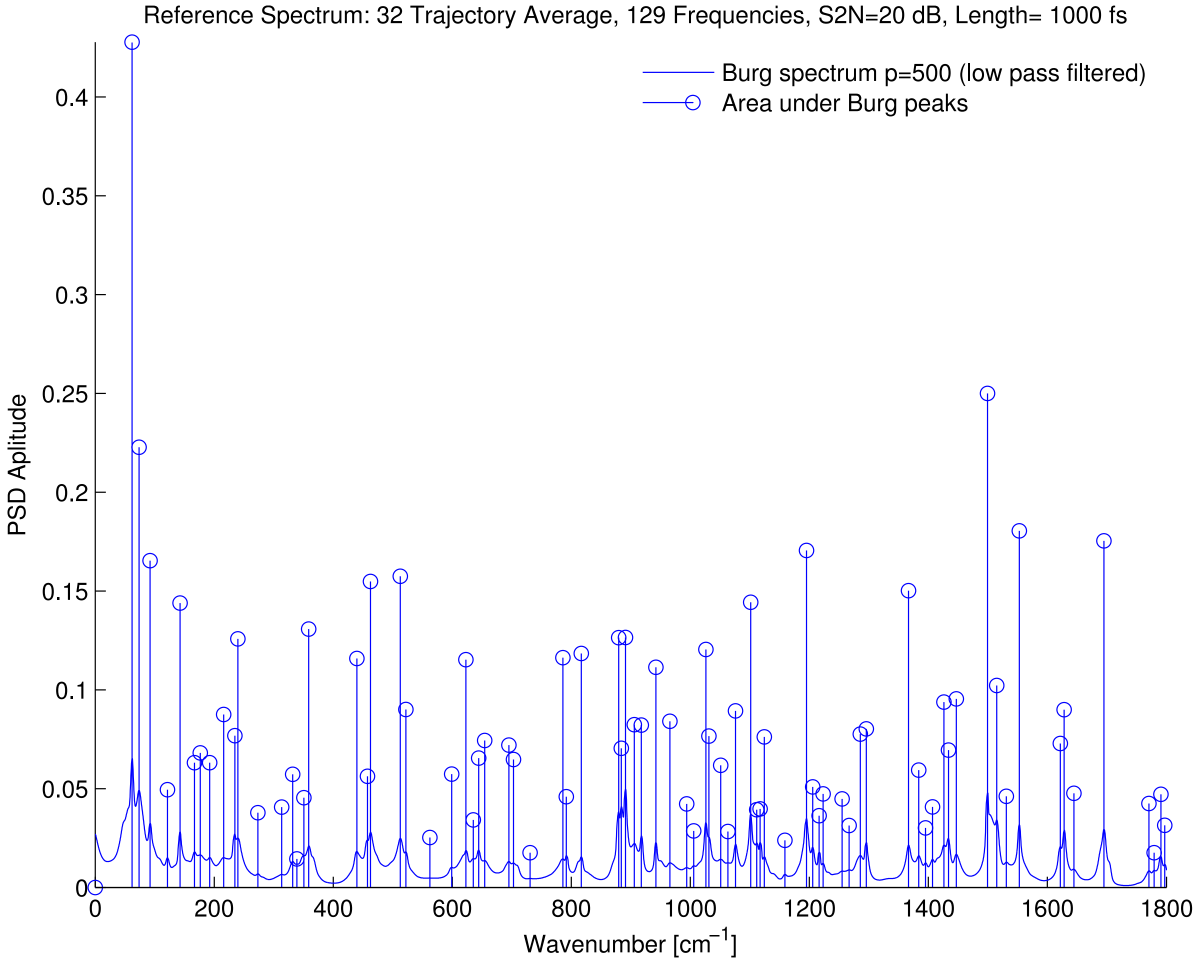

Figure 4.5: GFP like 1000 fs spectrum from figure 4.4; Burg spectrum (blue lines); integrated area under peaks (blue dots). The area under the Burg peaks is not constant.

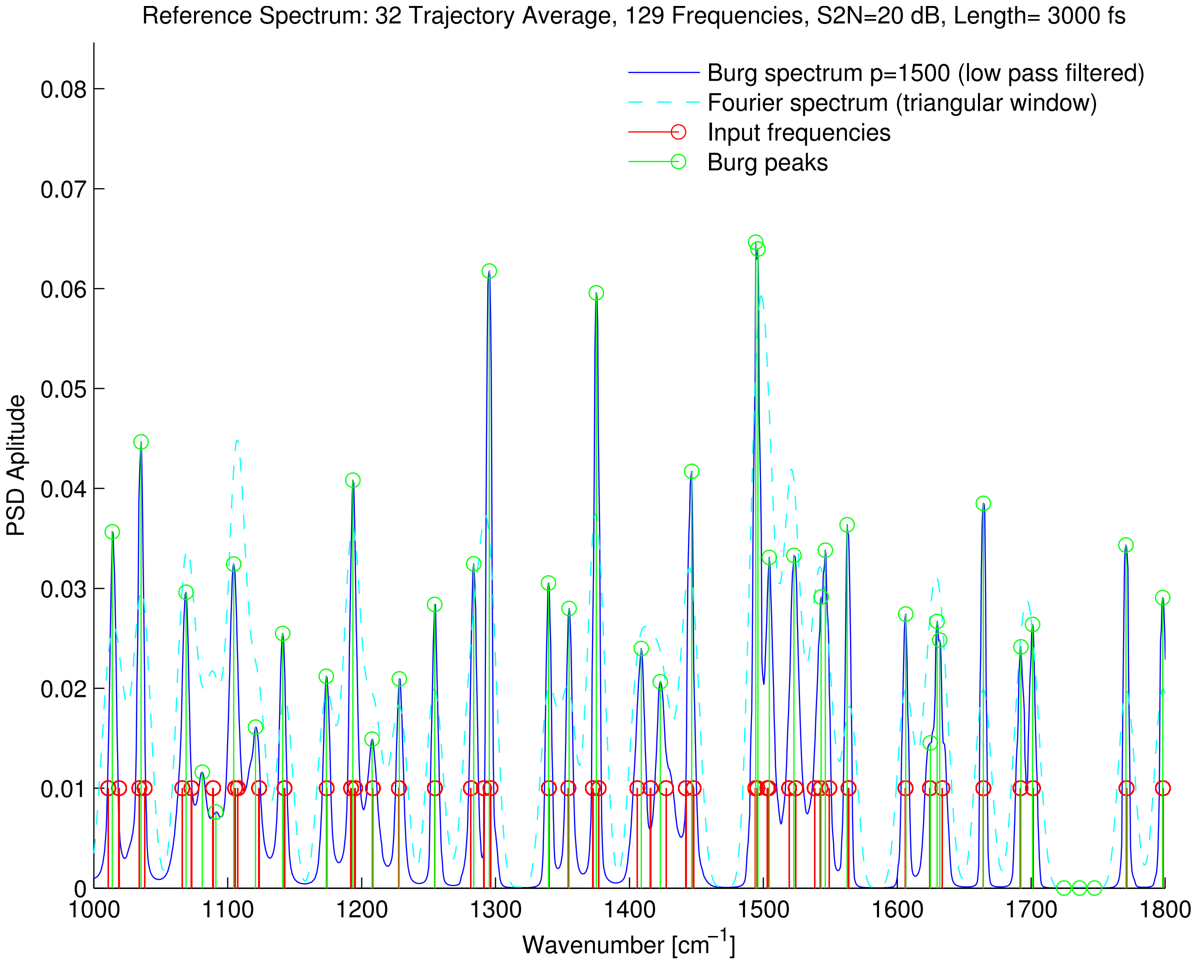

Figure 4.6: Portion of GFP like 3000 fs spectrum with 129 NMA frequencies (red); Burg peaks (green); Fourier spectrum (cyan); Burg spectrum (blue); target frequencies are well matched and peak heights greatly improved over 1000 fs spectrum.

4.1.2 Simulation Setup

Before discussing the quality of the simulated QM/MM spectra a short introduction into the simulation setup is given at this point. The ground state GFP x-ray structure 1GFL [76] was equilibrated and used as the starting point for generating a structure for the proposed I^* state. The I^* structure was generated by constraining all heavy atoms in the active pocket and only optimizing the hydrogen atoms for the deprotonated chromophore. Figure 4.1 shows a stick representation of the resulting conformation as well as the ground state conformation with the protonated chromophore. The MM part of the system was parametrized using the AMBER03 force field together with a custom parameter set for the protonated and unproto-

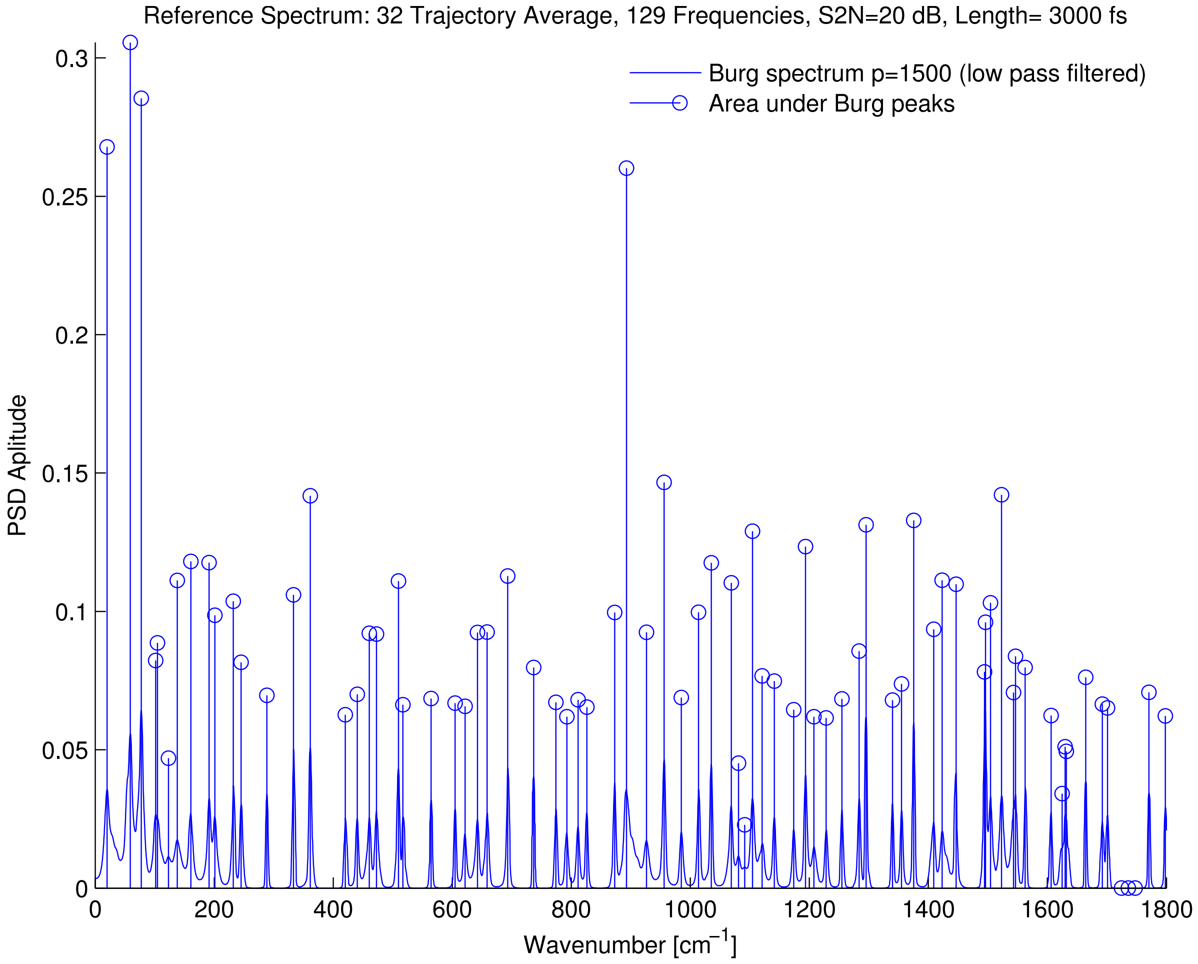

Figure 4.7: GFP like 3000 fs spectrum from figure 4.6; Burg spectrum (blue lines); integrated area under peaks (blue dots). The area under the peaks greatly improved over 1000 fs spectrum but not optimal.

nated gfp chromophore as kindly provided by Gerrit Groenhof. The AMBER99sb [33] parametrization procedure described for the PYP chromophore in section 4.2 was also applied to the GFP chromophore but did not result in a stable hydrogen bond network which is why the existing AMBER03 parameter set was used. A total of 48 snapshots were forked off a 30 ps MD simulation trajectory for both the A and I_1 state. The Gromacs 4.5.1 QM/MM simulation software was modified to record the dipole moment vector time series as well as the coordinates of the chromophore pocket. The QM/MM simulations were run for 1000 fs each. From this data, the dipole moment time series was corrected for the center of mass motion of the chromophore pocket. To illustrate what the result of these simulations is, the resulting time series for the ground state is shown in Appendix figure 7.9.

Reduction of CASSCF active space

The CASSCF method was introduced in section 2.2.4. CASSCF is a truncated version of the full CI method which produces good results for excited states. It only includes selected electrons into the CI expansion while all other electrons are described at the lower Hartree-Fock level of theory. The selection of the CASSCF active space is crucial for the quality of the resulting wave function. This makes CASSCF one of the few quantum chemical methods which are not 'black box'. Instead, the chemically important delocalized π-orbitals are selected individually by hand. Due to the very high cost of large active spaces, the full GFP π electron system cannot be included. Therefore, a reduction of the active space is unavoidable.

The reduction is usually performed by removing orbitals piecewise based on their electron occupation numbers. The reduction process requires careful orbital evaluations and a great portion of 'chemical intuition'. Therefore, the reduction from CASSCF(12/12) to CASSCF(6/6) may require up to 12 reduction steps. As there are only

$$N_{(12/12)\rightarrow(6/6)} = \binom{3}{6}_{bind} \binom{3}{6}_{anti} = 400 \tag{4.5}$$

possible candidates to reduce the CASSCF(12/12) active space to CASSCF(6/6) a brute force algorithm was developed to try them all. Thus, the reduction process would no longer require chemical intuition. A selection rule was designed such that the physically important quantities of dipole moment and excitation energy are conserved in the reduction process. All 400 active space reductions were scored according to the function

$$C = \Delta p_{gs} + \Delta p_{s1} + 100 * \Delta E_{vert}. \tag{4.6}$$

In this approach, the difference in dipole moment for the ground Δp_{gs} and first excited state Δp_{s1} between the larger CASSCF(12/12) and each smaller CASSCF(6/6) active space was used. Additionally, the difference in vertical excitation energies ΔE_{vert} was added to the two dipole moment differences to obtain the total deviation measure C. The choice of the two dipole moment difference parameters is based on the assumption that a reduced active space should reproduce a similar wave function of only the most relevant correlated orbitals which is expected to results in only a small deviation of the total dipole moment. The vertical excitation energy was included as a quality measure for the description of the transition to the first excited state. This energy was added in units of Hartree and has therefore only a small contribution to the overall score. The dipole moment is considered to be more important for the subsequent spectral analysis. The result of this reduction for the

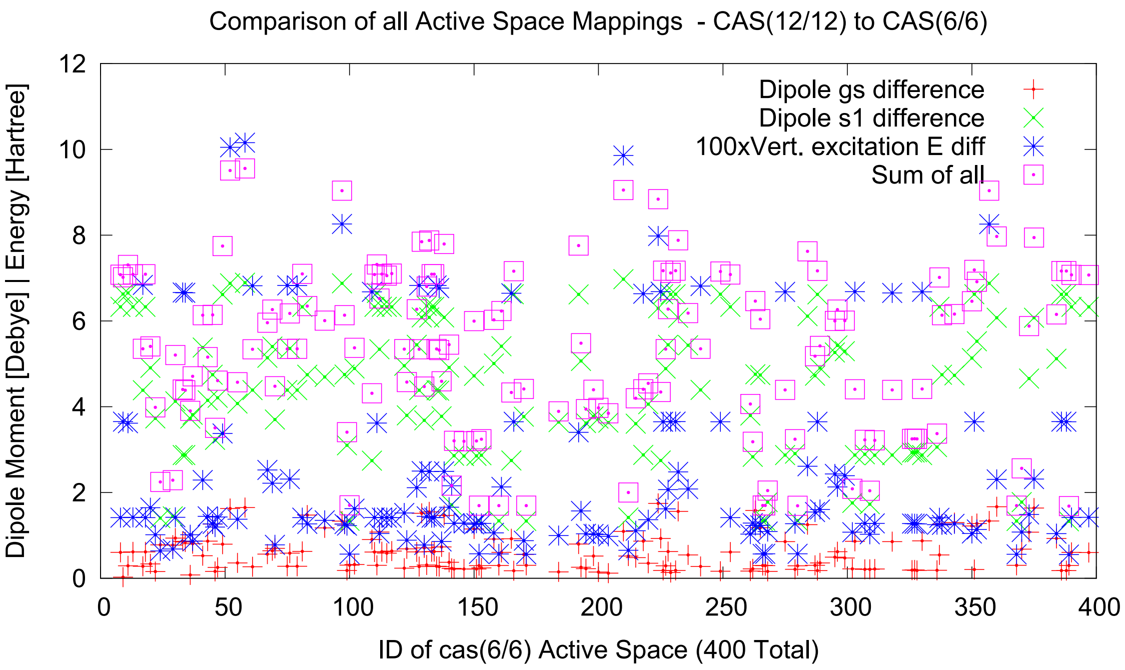

Figure 4.8: Representation of all converged candidates out of the 400 possible CASSCF(12/12) to CASSCF(6/6) reductions; (magenta) total score C for each conformation $C = \Delta p_{gs} + \Delta p_{s1} + 100 * \Delta E_{vert}$ (lower is better); (red) Δp_{gs} difference in ground state dipole moment CASSCF(12/12)-CASSCF(6/6); (green) Δp_{s1} difference in exited state dipole moment CASSCF(12/12)-CASSCF(6/6); (blue) $100\Delta E_{vert}$ vertical excitation energy difference CASSCF(12/12)-CASSCF(6/6). The best candidate is shown in figure 4.9.

neutral GFP chromophore pocket is shown in figure 4.8. An interesting feature of this diffuse plot is the large spread in dipole moment differences. This is an indirect measure of the quality of the underlying wave function which strongly depends on the orbitals chosen in the reduced space. In order to determine the best reduced active space, the energies of the eight lowest overall scores where examined. These eight candidates all have an overall score of less than two and seven of which also have degenerate total energies. This implies that there is more than one best active space and several reduction choices correspond to the same reduced wave function. Out of the seven degenerate reductions the one with a slightly lower overall score

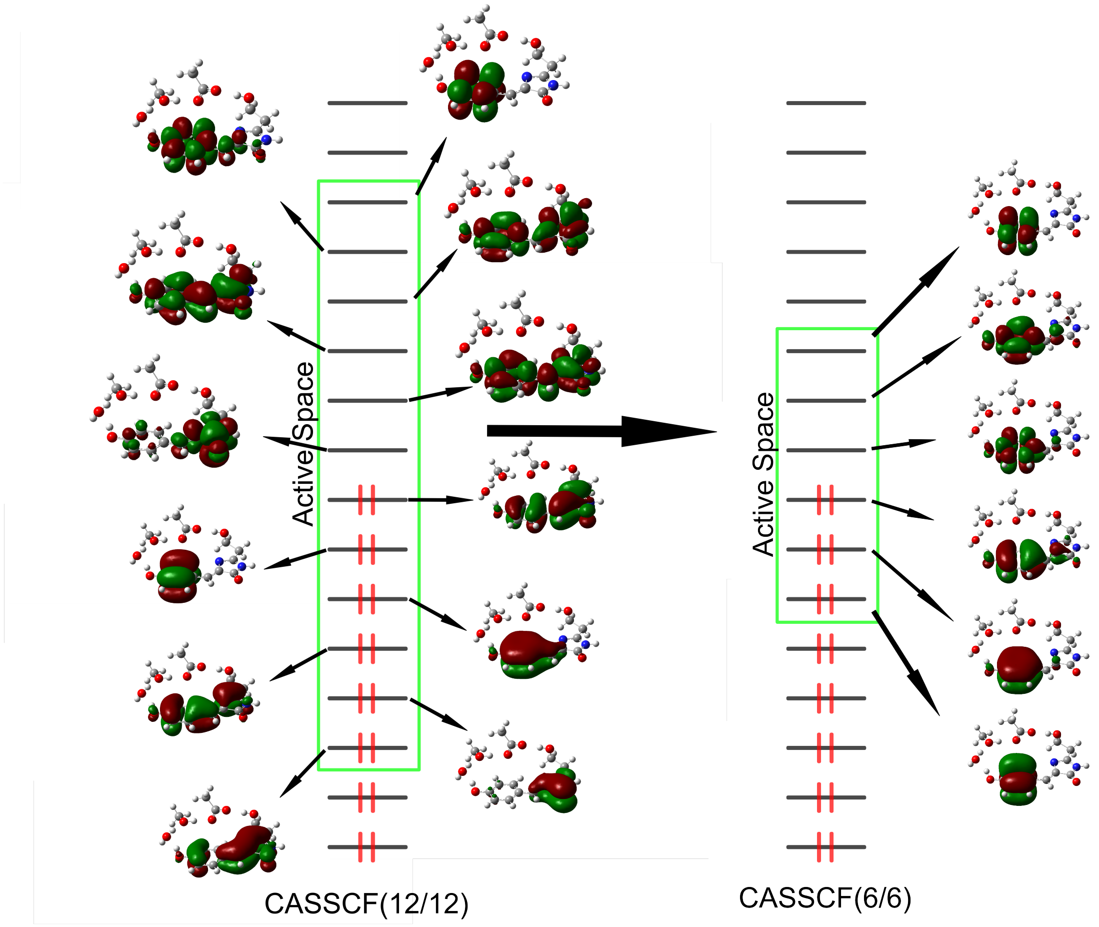

Figure 4.9: Orbital representation for the optimal reduction candidate of the GFP CASSCF(12/12) active space: (right) best CASSCF(6/6) reduced active space out of all 400 possible combinations.

was chosen for the final CASSCF reduced active space. The orbitals corresponding to this best candidate are presented in figure 4.9

Unfortunately, the reduced CASSCF(6/6) wave function still required 65 minutes of computing time per time step. The cluster node was a 2Ghz Magny-Cour AMD machine with 8GB of RAM using 4 cores and the Gaussian03 [23] quantum package and a convergence parameter of 10^{-8}. The Molpro [73] quantum package was also tested but did not improve the Gaussian result. In conclusion, the required 3000 fs of simulation trajectory for the A^* and I^* states in GFP would consume around 4 months of computing time which is not an option at the time of writing this thesis. In order to still test the applicability of the developed method to GFP, ground state trajectories were generated using the very efficient density functional theory (DFT) in combination with the B3LYP functional.

4.1.3 Anharmonic GFP Spectra from QM/MM Simulation

The GFP protein was simulated in the A and I_1 state using the ground state DFT method and the B3LYP functional together with a 6-31g* Gaussian type basis set. A trajectory length of 1000 fs was chosen as this is currently the limit for excited state calculations. However, ground state trajectories can in principle be extended to the required 3-4 ps trajectory length. From the resulting dipole moment time series, both the Burg and the Fourier spectra were calculated. The maximum entropy spectral data is not shown due to its inferior performance in the model order prediction calculations. The ground state raw data for all 48 trajectories is shown in Appendix figure 7.9 and the I_1 state raw data is shown in Appendix figure 7.10. The raw data is included to illustrate that the time series are stationary as required by the analysis methods. Not all generated time series were used for signal analysis. The selected subset is described in the raw data figure captions. The selection was made to exclude short trajectories due to hardware defects, slow node performance and anomalies observed in the visual trajectory inspection. Trajectories in which the hydrogen bond network broke were excluded as well as trajectories in which the starting conformation showed a twisted GLU222 residue. In the I_1 state trajectories, GLU222 twisting was observed in five cases out of the 24 1 ps trajectories. The twisting did not result in a breaking of the hydrogen bond network. The simulation data roughly corresponds to a 4 ps^{-1} decay process. However, a more reliable estimate of the observed process as well as the possible identification of the I_0 state requires longer trajectories and better statistics.

In total 33 individual vibrational spectra were used to calculate averaged Fourier and Burg spectra. The overlay of both spectra is shown in figure 4.10. The same procedure was applied to the Fourier and Burg spectra for the I_1 state. The resulting averaged spectra are shown in figure 4.11.

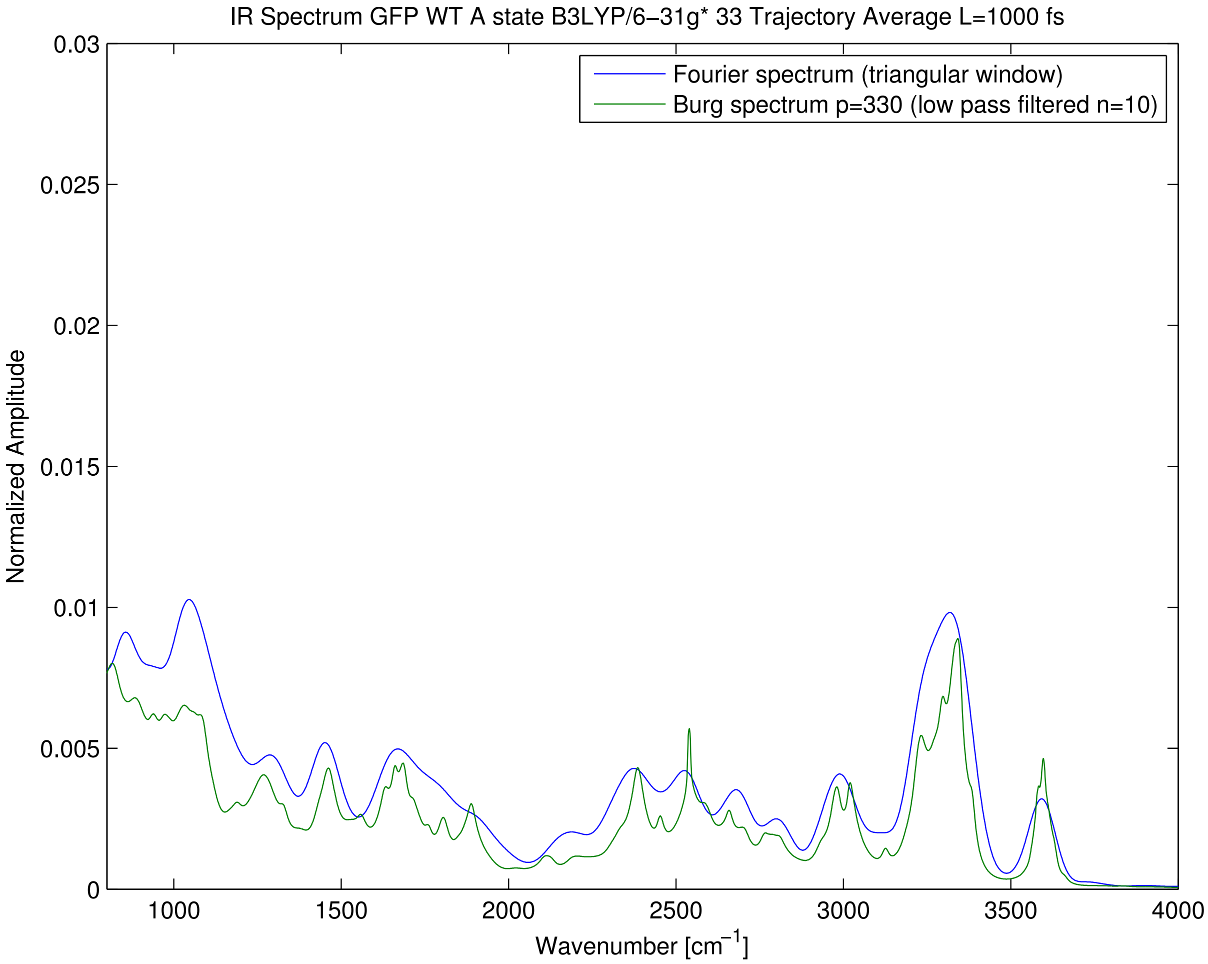

Figure 4.10: Anharmonic GFP spectrum of the ground A state from 33 1000 fs QM/MM simulations (unscaled frequencies).

The results are in agreement with the expected accuracy from the model order estimation process in section 4.1.1. The model order estimation calculations predicted that 1000 fs sampling of the underlying GFP potential is not sufficient to distinguish individual peaks in the spectrum. Additionally, the Burg peak locations do most likely not correspond to the expected frequencies nor do the peak heights correspond to the correct power distribution. The predicted resolution limit of the Fourier transform cannot be improved due insufficient trajectory lengths. The resolution does also not increase through averaging of multiple trajectories. However,

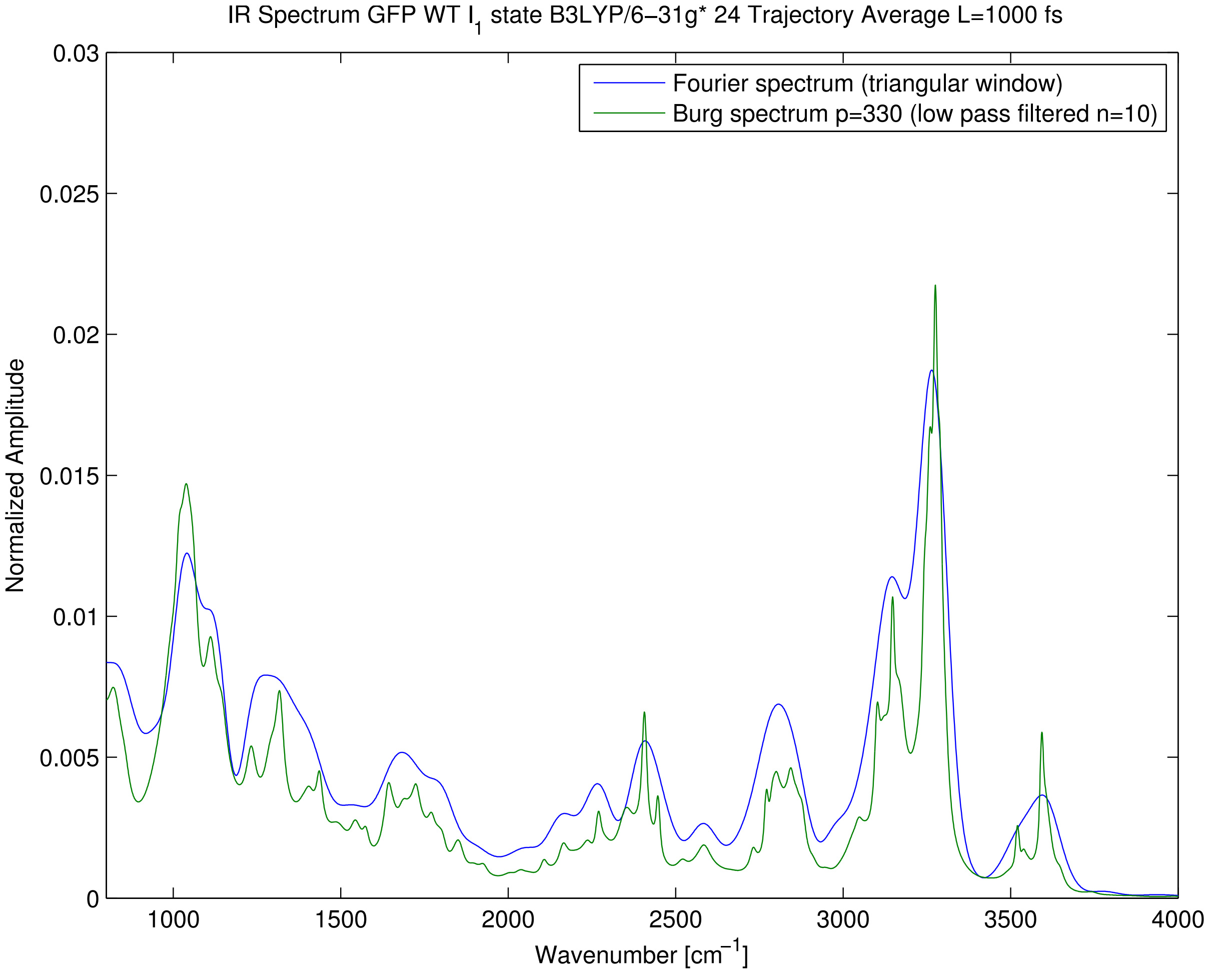

Figure 4.11: Anharmonic GFP spectrum of the I_1 state from 24 1000 fs QM/MM simulations (unscaled frequencies).

the simulations do reveal an interesting effect of the MM environment on the QM dipole moment time series which the model order prediction calculations in section 4.1.1 do not consider. The effect becomes visible when the full spectrum is examined. Figure 4.12 illustrates how the MM environment contributes to the low frequency part of the spectrum below 1000 cm^{-1}. The contribution is likely caused by angle and dihedral vibrations. The bond vibrations from the MM region were constrained during the QM/MM simulation and should not appear in the spectrum. However, the angle and dihedral degrees of freedom were not constrained to ensure a physically correct behavior of the protein. The center of mass motion of the active pocket which was removed beforehand also contributes to this part of the spectrum. It becomes eminent that the low frequency bias to the spectrum is the price to pay when including the protein explicitly into the analysis of the chromophore spectrum. The MM contribution also increases the total number of degrees of freedom which have to be identified. The total required trajectory length and autoregressive parameter number may increase beyond the predictions made in section 4.1.1 due to this effect. The spectral analysis of GFP is concluded at this point, as longer simulation trajectories are required for further analysis.

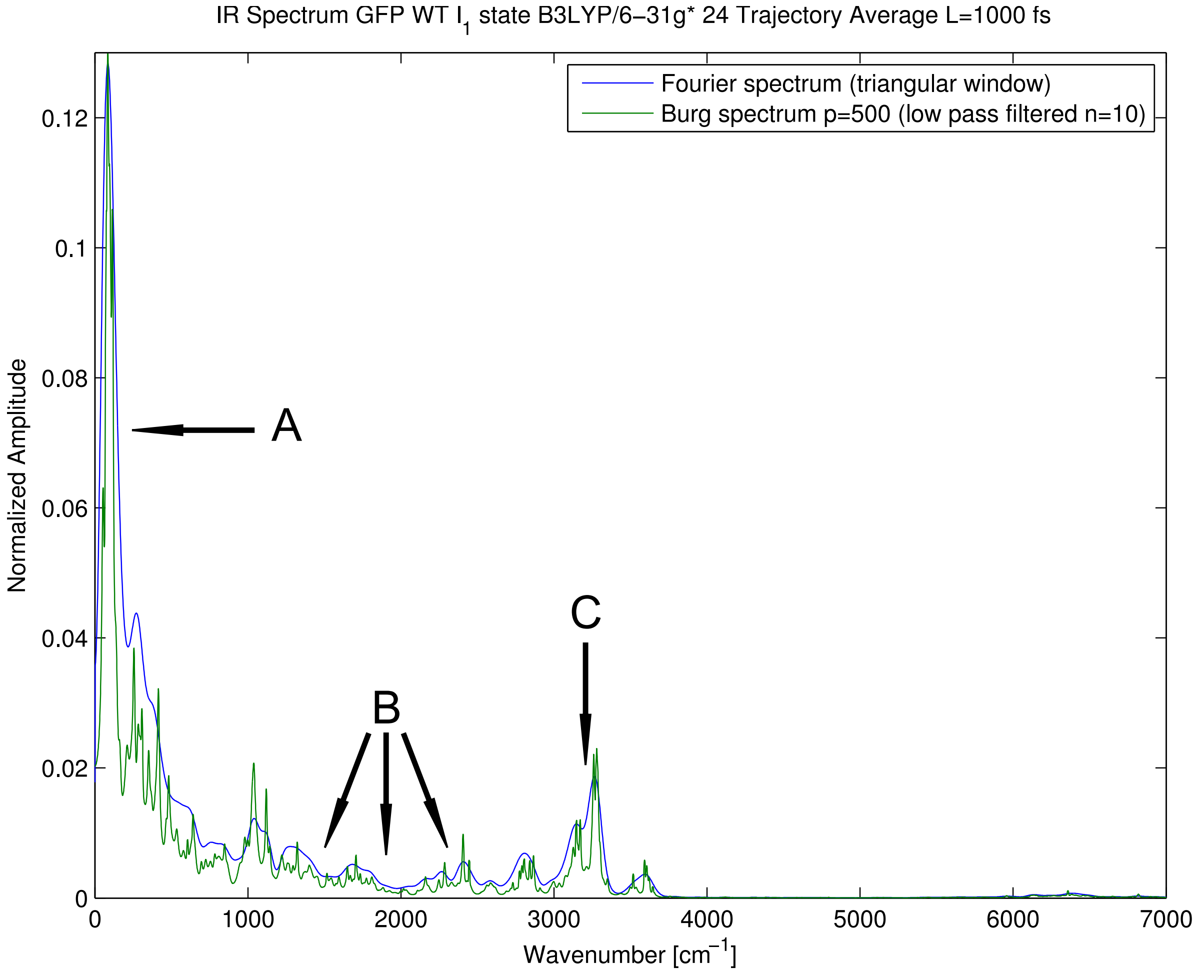

Figure 4.12: Full GFP I_1 state spectrum (unscaled frequencies), 24 1000 fs QM/MM simulations; A) slow dihedral and angle vibrational contributions from MM region (MM bond vibrations constrained to zero during simulation); B) carbon and oxygen bond vibrations QM region; C) fast hydrogen bond vibrations.

4.2 Photoactive Yellow Protein

The method testing now continues with the spectral analysis of the Photoactive Yellow Protein (PYP) chromophore. This system was predominantly chosen for its accessibility in both simulations and experiments. The QM region in this system can be chosen much smaller than in GFP which enables longer trajectories and higher spectral resolution due to a decrease in total vibrational degrees of freedom. The reader is reminded that the question to be answered at this point is not about the biology of the PYP system. Two open questions remain. First, it is not clear whether

QM/MM trajectory lengths are sufficient to identify individual contributions to the spectrum. Second, it will be investigated whether or not the Burg method can be used to improve the resolution limit set by the Fourier transform of short QM/MM simulation time series.

4.2.1 PYP Model Order Prediction

The model order prediction scheme described for GFP in section 4.1.1 was also applied to the smaller locked PYP chromophore, figure 4.14. The QM/MM simulations of GFP included the complete proton wire surrounding the chromophore with a total of 45 atoms (182 electrons). The PYP QM region was reduced to only the locked chromophore without including any further protein residues into the QM region. This effectively reduces the number of QM atoms to $N = 26$ (116 electrons) and the total degrees of freedom to 3N−6=72. Given the same size of CASSCF active space, the PYP QM calculation can be expected to be around four times faster than the GFP system due to the reduced number of electrons. Figure 4.13 shows the normal mode spectrum and the frequency density for all 72 harmonic frequencies. The locked chromophore has two additional carbon atoms. Therefore, a parameter set which sufficiently describes the locked chromophore is assumed to also perform well for the smaller wild type chromophore.

The spectrum for a 2500 fs trajectory generated from the locked PYP chromophore normal mode frequencies is shown in figures 4.15 and 4.16. The identification of frequency peak locations is significantly better than for the 3000 fs GFP trajectory. The input frequencies are recovered within less than 10 wavenumber deviation. Also, the recovery of spectral power is significantly better than for the GFP system. The Burg spectrum separates peaks which cannot be distinguished in the Fourier spectrum and overcomes the resolution limit described in formula 4.1. The raw data for the Burg, Fourier and maximum entropy spectra are shown in appendix figures 7.1, 7.2 and 7.3. The maximum entropy spectrum only marginally improves the frequency resolution compared to the Fourier spectrum. This barely justifies the drawbacks on including a parameter based model. Therefore, the parameter based maximum entropy method is no alternative to the parameter free Fourier transform. The Burg spectrum is also parameter dependent but noticeably improves the Fourier frequency resolution.

Based on these model prediction order calculations, Burg's method is the preferable choice for PYP spectral analysis given a sufficiently long trajectory of 2500

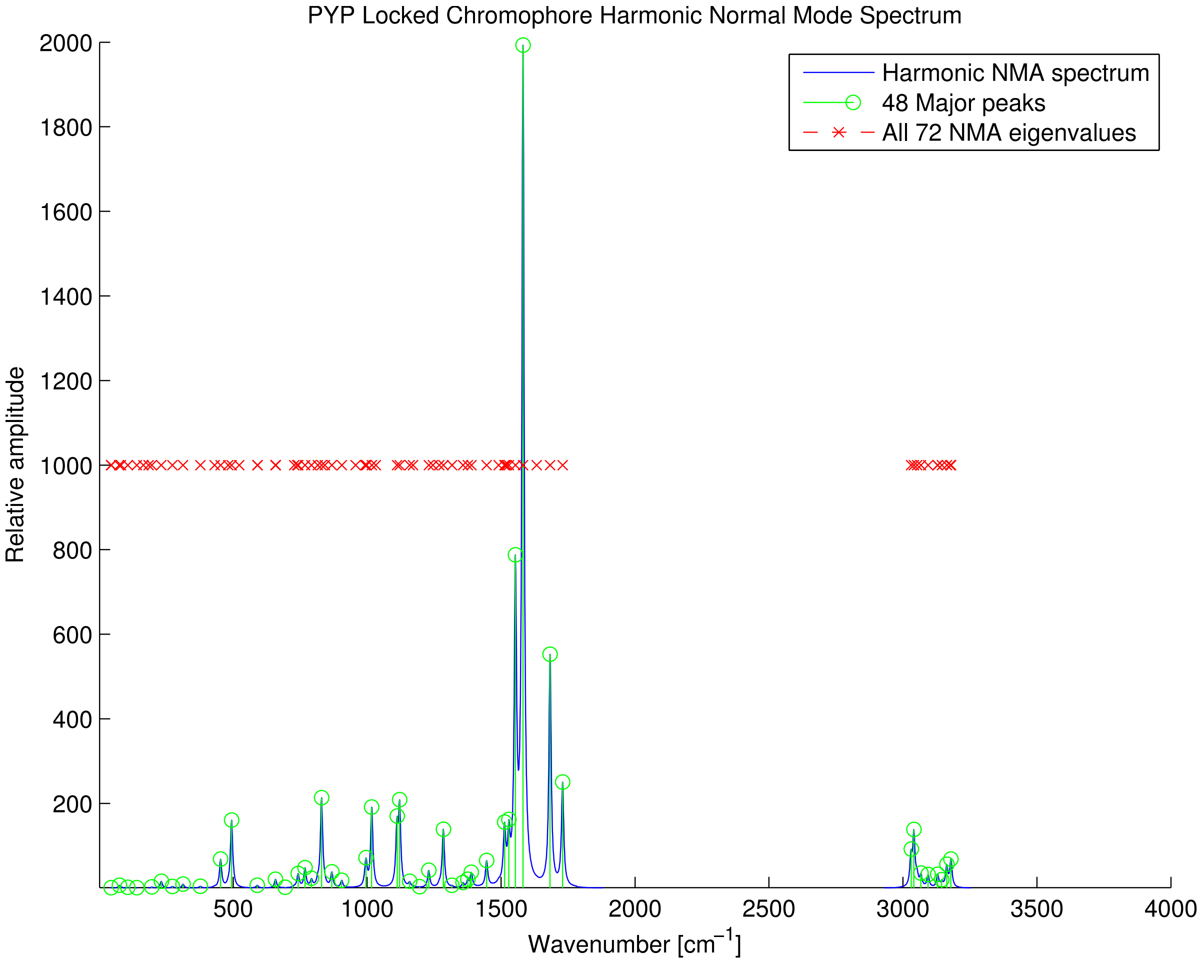

Figure 4.13: normal mode spectrum of the locked PYP chromophore; (blue) NMA spectrum; (green) 48 major frequency peaks; (red) full NMA frequency density

fs. The performance for shorter trajectories was also examined and 1000 fs spectra are shown in appendix figures 7.4, 7.5, 7.6, 7.7 and 7.8. The quality of the 1000 fs spectrum is sightly better than the 1000 fs GFP spectrum but not sufficient for the calculation of accurate difference spectra. At 1000 fs simulation length, only qualitative spectral differences and large shifts are visible.

4.2.2 Simulation Setup

Simulations of the Photoactive Yellow Protein (PYP) wild type and three selected mutants, see figure4.17, were performed. All systems were simulated in the ground state at the DFT-B3LYP/6-31g* level of theory while the locked mutant was also

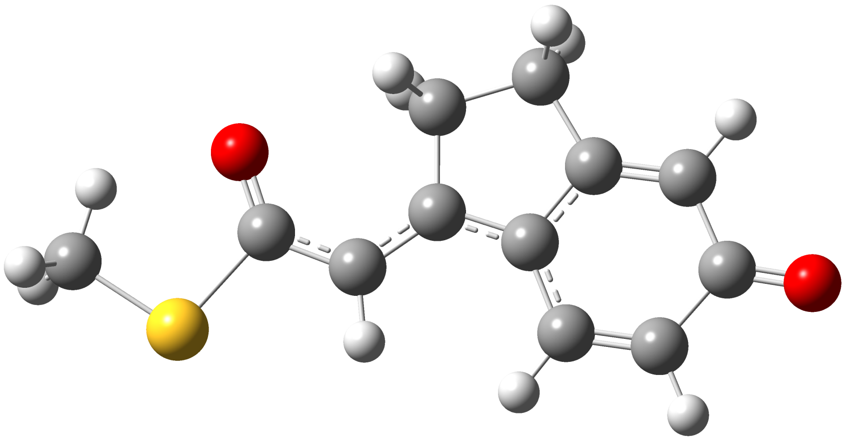

Figure 4.14: Stucture of locked PYP chromophore for Normal Mode Analysis

simulated at the CASSCF(6/6) level of theory.

The wild type PYP x-ray crystal structure, pdb code 2ZOH [75], was used as the starting conformation. Based on this conformation, three mutants were generated. First, the wild type chromophore was replaced by a locked PYP chromophore which disables one of the two known isomerization channels by locking the single bond dihedral. Second, two additional mutants were generated by replacing the arginine residue 52 by alanine (R52A) in the wild type and locked protein. The Gromacs molecular dynamics program [18] was used to assign AMBER99sb [33] force field parameters to the amino acid residues. The AMBER99sb force field was chosen for its good description of protein interactions together with the computationally efficient tip3p[36] water model. The AMBER99sb force field was also chosen for its strong dependence on quantum mechanical calculations. This greatly simplifies the procedure of extending the force field with new residues. However, force field parametrization remains a tedious process that also requires experimental validation. The setup up to this point covers most of the protein and water dynamics but does not yet describe the dynamics of the most important part of the PYP protein, the chromophore.

The chromophore of the PYP protein is not part of any modern force field. This implies that this part of the protein cannot be simulated using the molecular dynamics methods. To resolve this issue, a custom force field parameter set was generated.

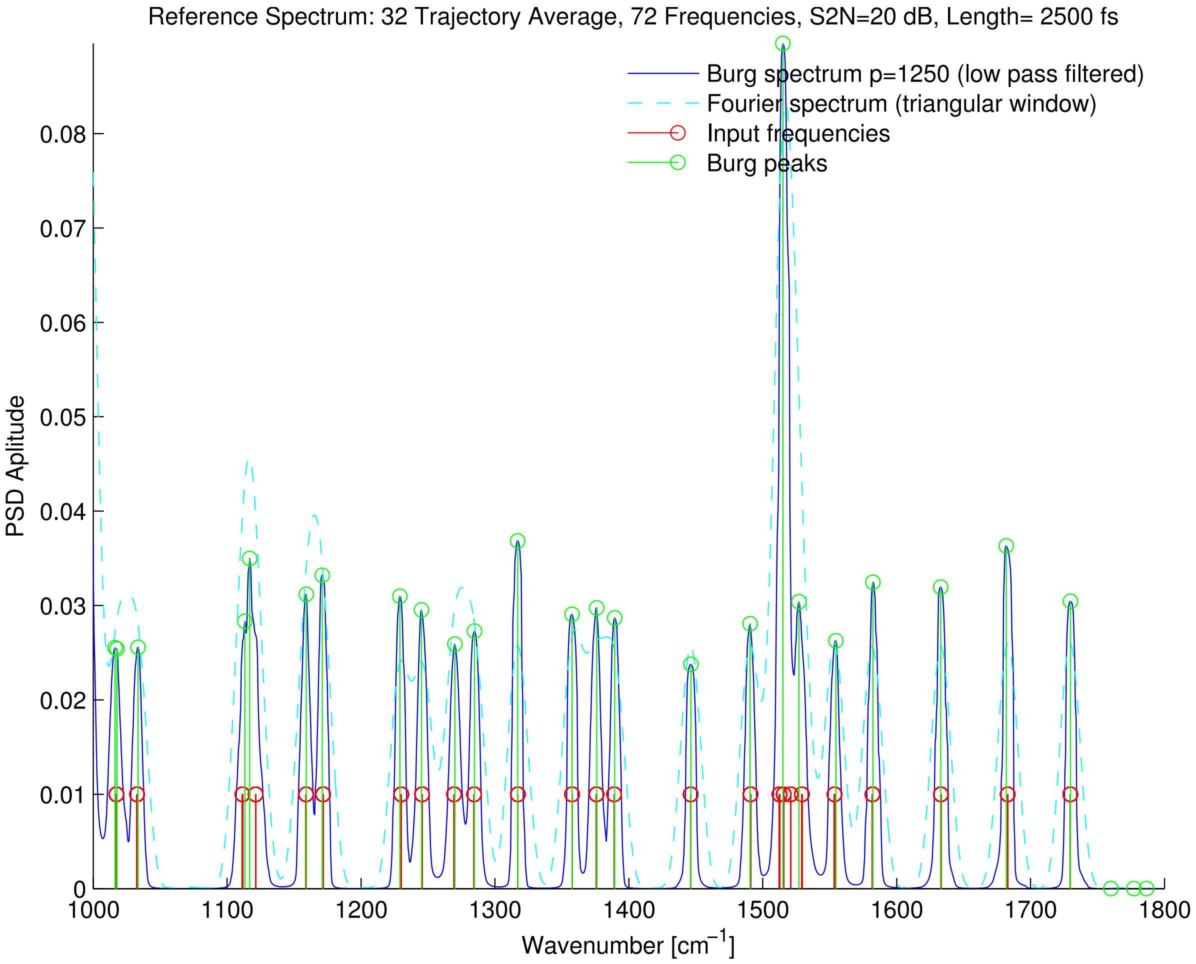

Figure 4.15: Portion of PYP like 2500 fs spectrum with 72 NMA frequencies (red); Burg peaks (green); Fourier spectrum (cyan); Burg spectrum (blue); Burg peak heights and frequency identification improves Fourier spectral estimates.

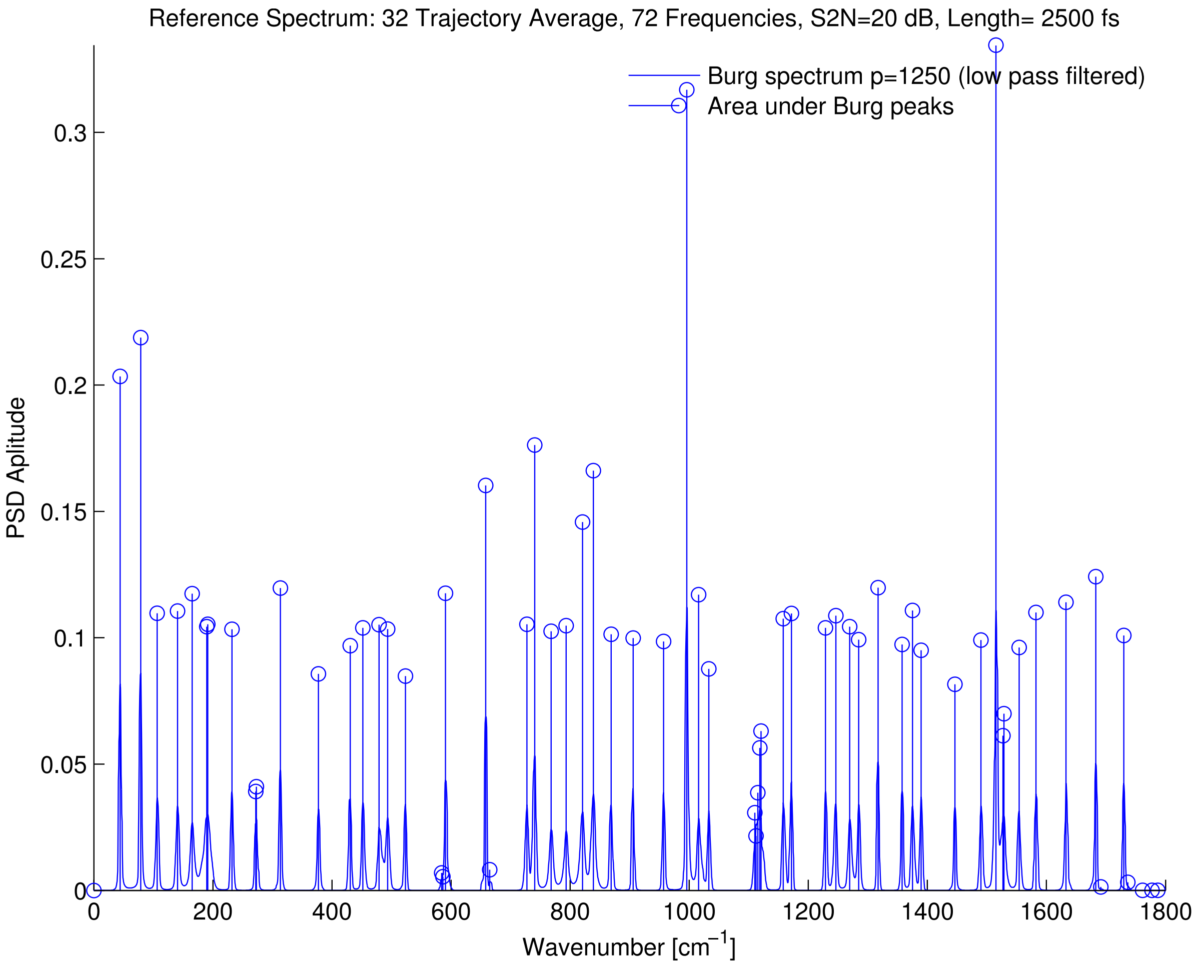

Figure 4.16: PYP like 2500 fs spectrum from figure 4.15; Burg spectrum (blue lines); integrated area under peaks (blue dots); variance in power (area under peaks) in good agreement with the expected flat power distribution

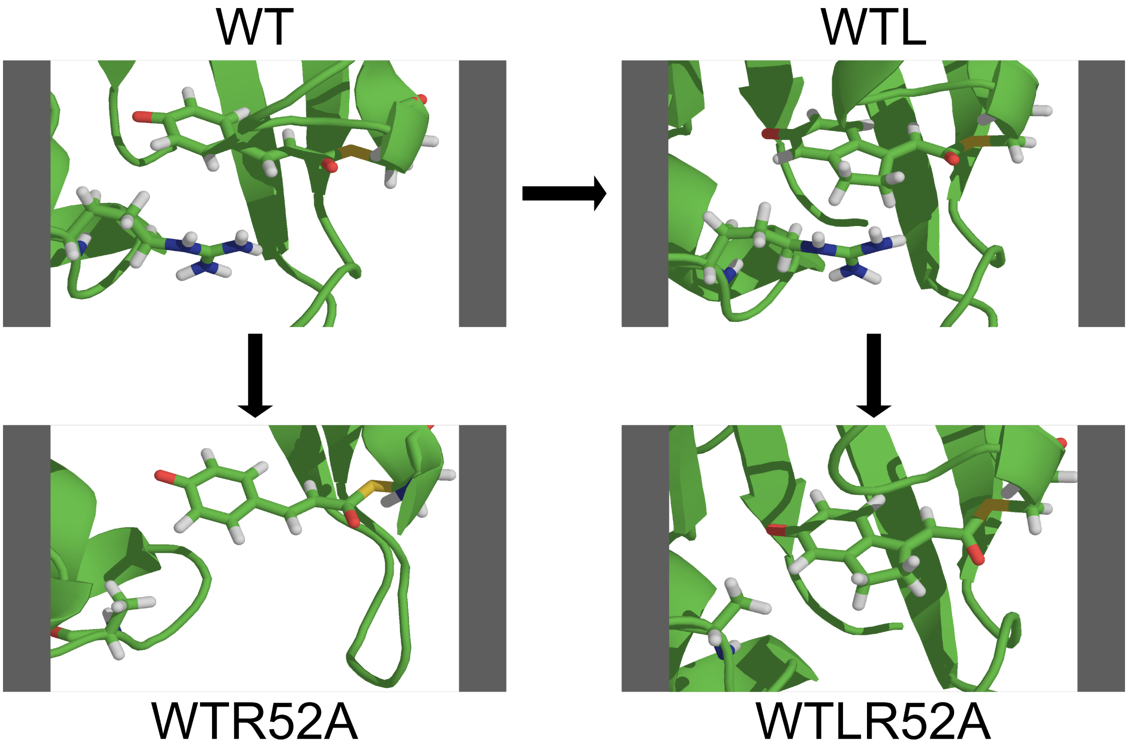

Figure 4.17: PYP wild type and three selected mutants; top left: PYP wild type (WT); top right: PYP locked WTL mutant; bottom left: PYP WTR52A mutant; bottom right: PYP locked WTLR52A mutant;

This was done following the parametrization philosophy of the previously applied AMBER99sb forcefield. However, the generated parameter set for the WT and locked PYP chromophore should be considered of comparable but not of equal quality as AMBER99sb. Modern protein force fields heavily rely on experimental input in form of configurational energies and solvation free energies, among many others. At the time of writing, this data was not available for the two PYP chromophore variants parametrized in this thesis.

The chromophore parameters were generated using the Generalized Amber Force Field (GAFF) [70] . The initial guess was generated using the automated atom and bondtype assignment procedure [71] developed in the lab of D. Case and the former Kollman group. The automatically generated parameters were subsequently checked and corrected by hand to match the known chemical topology. At this stage, the bonded parameters, i.e. bond stretching, angle bending, dihedral torsions and the ring planarity are covered. The Lennard-Jones parameters were taken as found in the AMBER99sb and GAFF parameter set.The electrostatic potential was reduced onto the nuclear centers using the Restrained Electrostatic Potential Fit

(RESP)[3]. Here, the backbone atomic charges were restrained to match the standard AMBER values to avoid incompatibilities. At this point the standard AMBER99sb parametrization philosophy was left. Instead of using the Hartree-Fock RESP charges from a vacuum calculation, the chromophores were placed inside the protein and the RESP charges were calculated using Density Functional Theory DFT and the B3LYP exchange/correlation functional. This procedure was chosen to better mimic the charge polarization of the chromophore inside the protein pocket.

The generated parameter sets for the WT and the locked WTL mutant were used to generate 6 ns of classical MD trajectories. From each trajectory, 48 snap shots were forked off as input for subsequent ground state QM/MM simulations. The DFT B3LYP/6-31g* QM/MM simulations were setup to run 1000 fs. However, the calculations for the model prediction error suggested that much longer trajectories are required to achieve reasonably high frequency resolutions. Therefore, the PYP WT as well as the PYP WTL mutant simulation trajectories were extended to up to 4 ps. In the following, only the data for these extended trajectories is shown. The quality of the 1000 fs spectra for the WTR52A and WTLR52A mutants is slightly better than those for the short GFP trajectories but still not sufficient to calculate difference spectra.

4.2.3 CASSCF Results

Much of the interesting PYP dynamics takes place in the excited state. Therefore, CASSCF simulations were prepared in order to simulate the excited state dynamics. This short section summarizes the problems in generating dipole time series for the excited chromophore. The active space of the PYP WTL chromophore was reduced in the same way as the GFP chromophore in section 4.1.2. The chromophore orientations from 48 DFT/B3LYP ground state frames were used to create CASSCF(6/6) guess wave functions. The chromophore structures from the snap shots were optimized at the HF/6-31g* level of theory followed by a CASSCF(12/12) single point calculation and the subsequent reduction to CASSCF(6/6). The same active space orbital selection was generated for each frame. Similar to the GFP reduction, the CASSCF(6/6) active space with the lowest overall score was used. The simulations were started in the ground state for each frame.

The resulting trajectories clearly speak against the described automated active space reduction. Out of 48 trajectories, only three trajectories reached 300 fs trajectory length without convergence failure. The proposed reduction is appealing as it does not require chemical intuition. However, this is also what seems to be missing. Usage of the proposed scheme is therefore not advised.

4.2.4 PYP WTL-WT Difference Spectrum

The extended ground state simulation trajectories for the PYP wild type and the locked mutant were used to generate averaged anharmonic Fourier and Burg spectra. The raw time series data is shown in Appendix figures 7.11 and 7.12. From the set of all trajectories, only a small subset of 11 long trajectories of at least 2800 fs were selected for both systems. Longer trajectories are expected to result in higher frequency resolution. From these subsets, averaged spectra for the WT and WTL chromophore were calculated which are shown in figures 4.18 and 4.19. Both figures also include the corresponding normal mode spectra for the isolated chromophores in vacuum. In addition, the WTL spectrum highlights three important spectral features.

First, the Burg spectrum contains several smaller side peaks. These side peaks were not present in the model order estimation calculations for the PYP WTL system. However, the nature of these peaks is likely not related to the vibrations of the chromophore but rather an artifact of the high model order. The number of suspected artifacts reduced significantly by increasing the trajectory lengths from 2300 to 2800 fs. A commonly applied upper bound for the Burg parameter number p in the literature is $L/3$ for L samples. In both of the PYP spectra, model orders as large as $L/2$ were used in combination with a low pass filter to smooth the resulting closely split lines. In the model order estimation section, this procedure was found to be in good agreement with the Fourier results as the closely split lines gathered around the expected frequencies and their convoluted heights corresponded to the Fourier amplitudes. However, the low pass filter also causes sightly lower Burg peak amplitudes in the PYP spectrum. An alternative way to reduce the side peaks could also be to include the full trajectory set into the average instead of just the small subset given sufficiently long simulation times.

Second, the normal mode spectrum fails to correctly predict the peak amplitudes at 300K. Instead, the NMA predicts a high spectral amplitude for a WTL peak located at 1590 cm^{-1} which is much lower in the actual simulated spectrum. Normal mode spectra are by design calculated at 0K which makes it difficult to predict temperature effects. The B) label in the WTL spectrum also shows a peak without

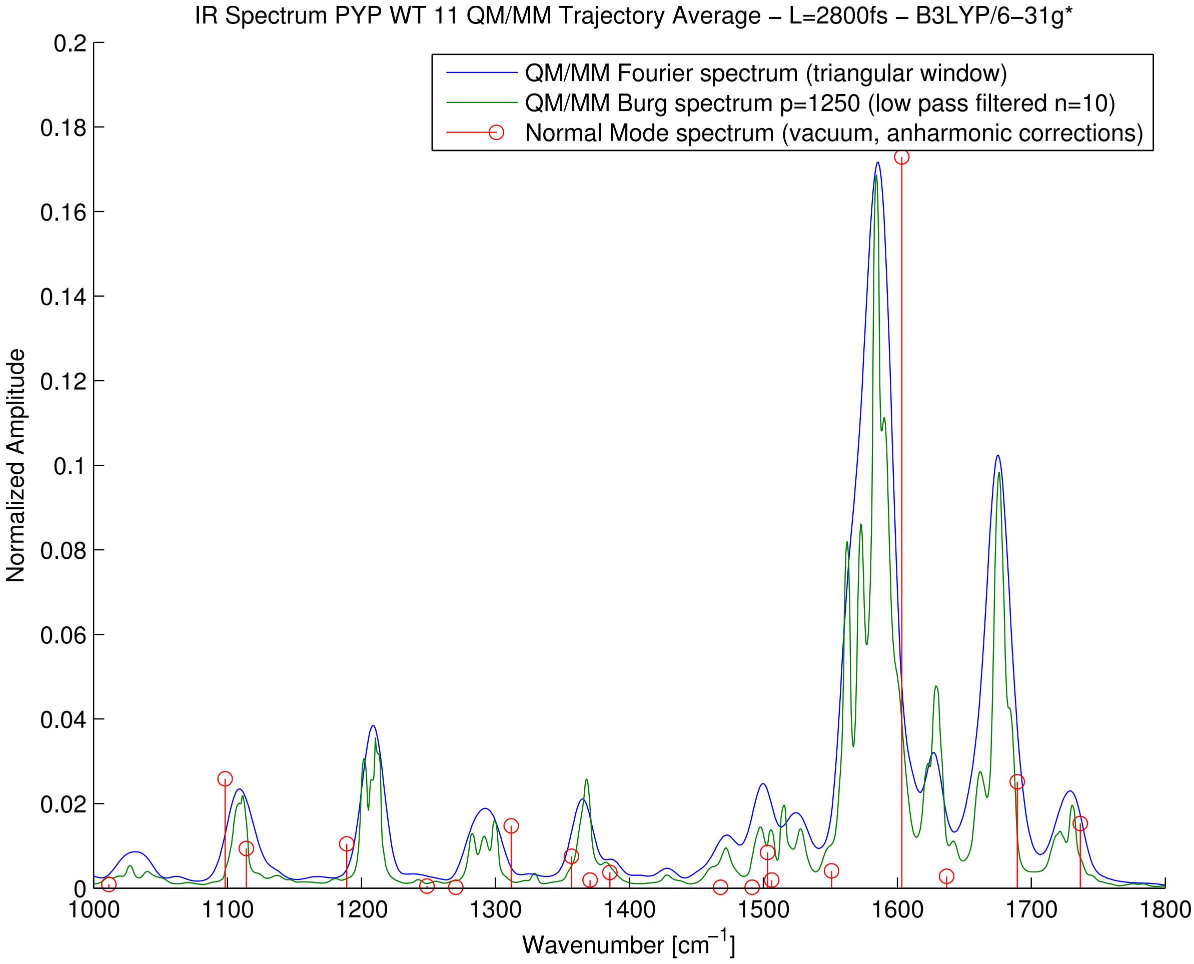

Figure 4.18: Anharmonic PYP WT ground state spectrum from 11 QM/MM simulations of length 2800 fs; blue line: average Fourier spectrum; green line: average Burg spectrum; red dots: normal mode spectrum of isolated chromophore in vacuum.

NMA amplitude that clearly contributes in the simulated spectrum.

Third, there are clear frequency shifts between the normal mode and time series spectra. From this data alone, it cannot be distinguished whether the frequency shifts result from the protein environment or from the difference in methodology. To separate the two possible sources, a small test system was simulated in vacuum. The spectrum was calculated using both time series analysis and NMA. The resulting spectra are shown in Appendix figure 7.14. However, this small test system also shows slight shifts in frequency which makes it difficult to determine the effect of the PYP protein environment without performing QM simulations of only the chromophores in vacuum.

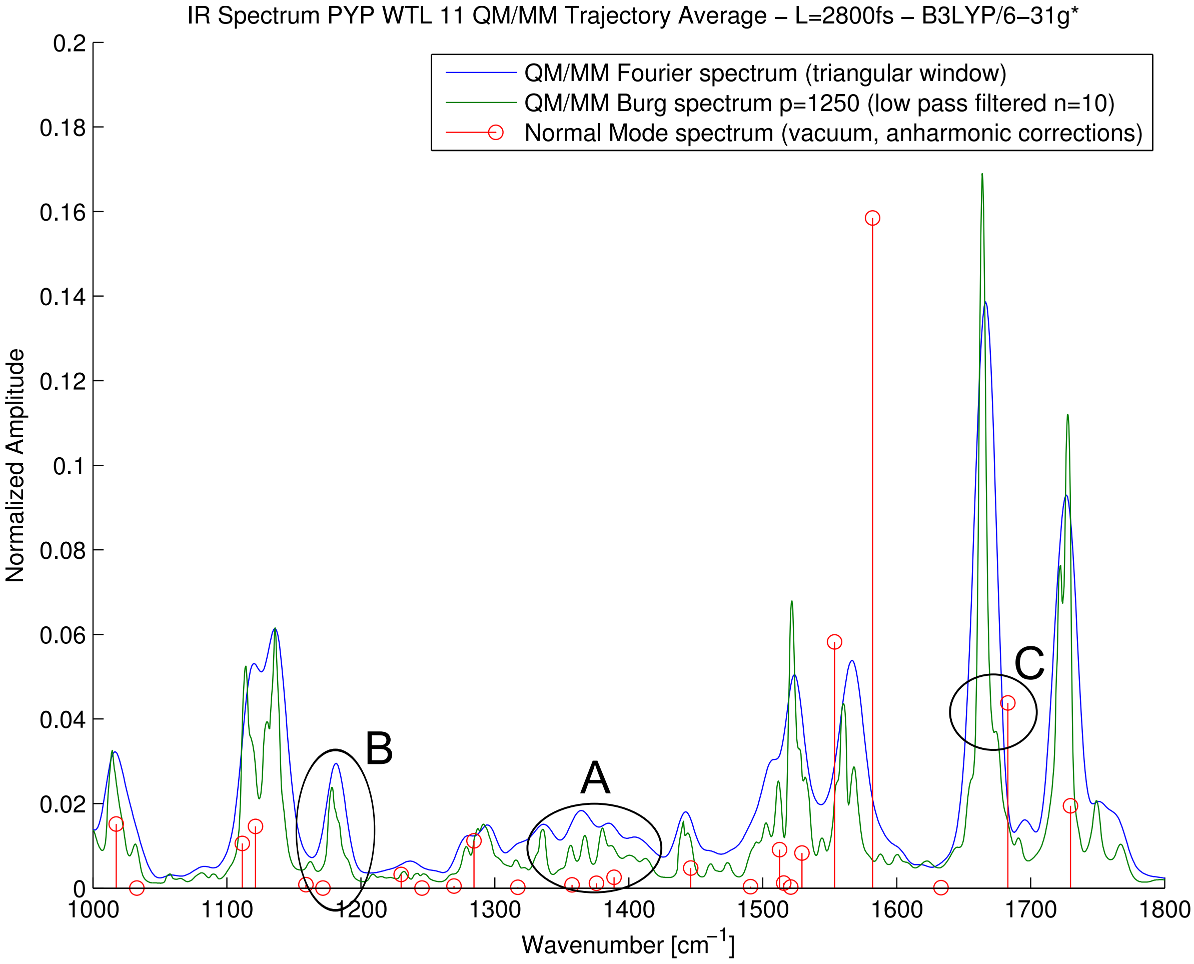

Figure 4.19: Anharmonic PYP WTL ground state spectrum from 11 QM/MM simulations of length 2800 fs; blue line: average Fourier spectrum; green line: average Burg spectrum; red dots: normal mode spectrum of isolated chromophore in vacuum. A) Burg line splitting due to over fitting. B) Failure of NMA to determine IR amplitude of the circled mode. C) Frequency shift between NMA and QM/MM spectrum.

Finally, the subset of long simulation trajectories was used to calculate a truely anharmonic vibrational difference spectrum between the WTL and the WT ground state spectrum. The resulting difference spectrum is especially exciting as it can directly be related to future experimental data which was not available at the time of writing this thesis. The calculated spectra reflect the frequency resolved average power distribution over the full trajectory length. An increase in trajectory length will also increases the accuracy at which power contributions can be located in the spectrum. This results in higher but narrower peaks which has to be considered when

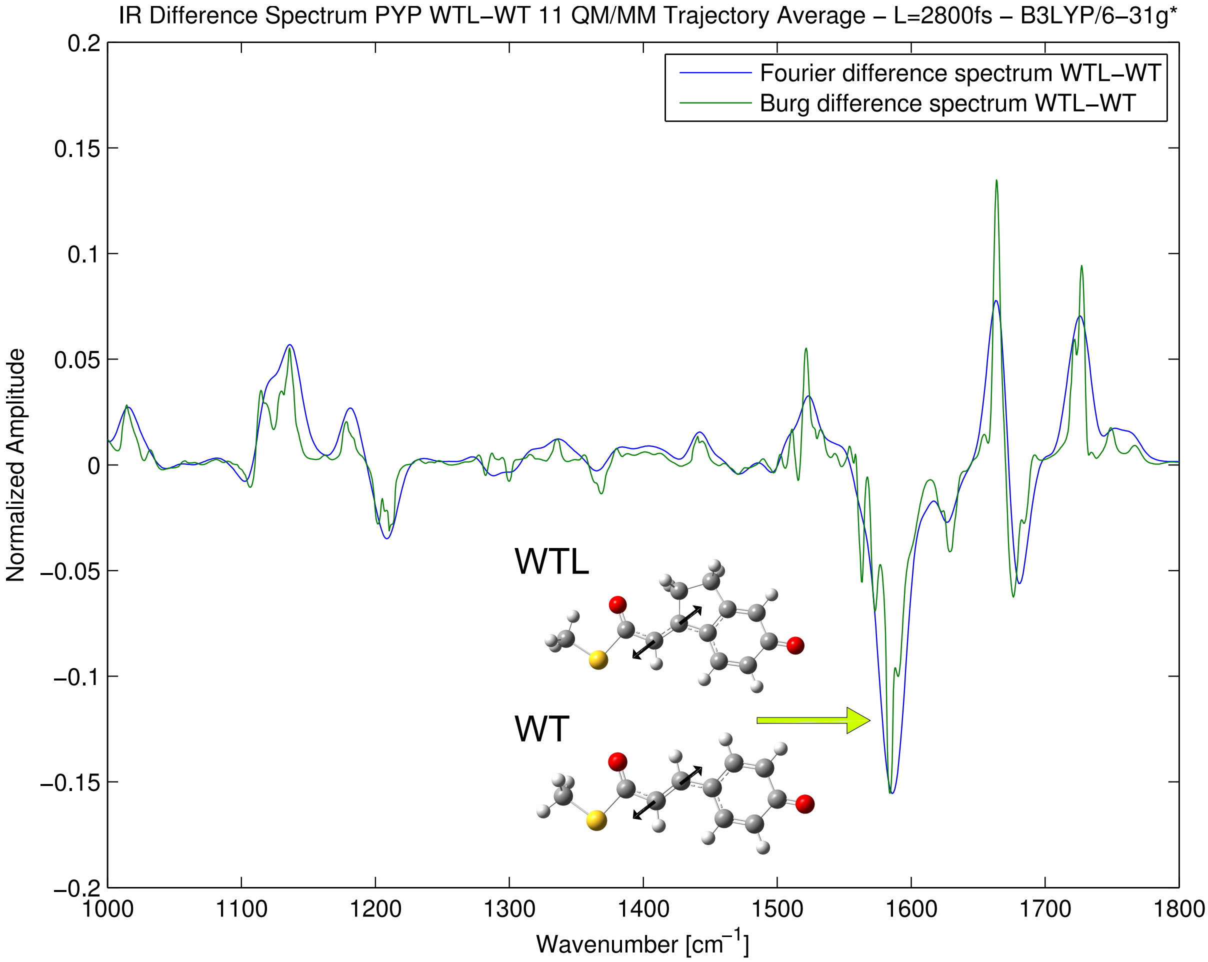

Figure 4.20: Anharmonic PYP WTL-WT ground state difference spectrum from 11 QM/MM simulations of length 2800 fs ; green arrow indicates a large difference in the WTL spectrum at around 1590 cm^{-1}; black arrows visualize the corresponding Normal Coordinate in the WT and WTL chromophore. The large negative peak in the spectrum indicates that the illustrated motion is absent in the WTL chromophore.

applying the presented method to calculate difference spectra. Difference spectra will have the highest accuracy when both systems have similar overall degrees of freedom and equal trajectory lengths.

The WTL-WT spectrum is shown in figure 4.20. A trajectory length of L=2800 fs was used for all analyzed time series. The spectrum clearly shows a large negative peak at 1590 cm^{-1}. Based on the normal mode data calculated for both chromophores in vacuum, the closest corresponding vibrational frequency to the 1590 cm^{-1} peak was identified for each. Both frequencies also correspond to the same nuclear motion which is illustrated by the black arrows. The absence of this motion in the WTL chromophore could be explained by the mechanical restrain on this motion introduced by the additional two carbon atoms in the WTL chromophore.

To avoid overinterpretation of the simulated spectrum, only the most intense difference is considered for now but further analysis is possible. Hopefully, experimental data will shed light on the true predictive power of the smaller peaks.

5 Discussion

The attempt to increase the spectral resolution using parameter based methods revealed many pitfalls which turned out to be difficult to control. The quality of model based spectra strongly depends on the model order. It is tempting to increase the subjective quality of a spectrum by tuning the Burg model order. However, the predictive quality of these tuned spectra will be meaningless as any result can be obtained this way. Therefore, the model order was carefully estimated using a reference signal generated from the normal mode frequency density. The model order obtained this way was not changed during data analysis which restores the objectivity of the spectra.

It is noteworthy that the estimated optimal Burg model orders were fairly high with p=1250 for the PYP and p=1500 for the GFP case. This is due to the fact that all-pole models, such as Burg's method, only poorly describe gaps in the spectrum which greatly increases the required model order. Combined all-pole|all-zero ARMA models also exist in the literature which combine autoregressive filters with moving average models at the cost of introducing another model parameter. However, a second model parameter only complicates the already difficult process of determining the right models order. Therefore, high Burg model orders are likely unavoidable to increase the spectral resolution.

Burg's method does increase the resolution limit compared to Fourier spectra. However, parameter based methods should not be seen as a replacement for the Fourier transform. Instead, Burg's method offers an alternative and independent way of analyzing the QM/MM time series. Fourier spectra of short trajectories are very smooth and tend to combine closely separated peaks into one. Burg spectra on the other hand tend to overshoot due to their all-pole nature and quickly separate broad Fourier peaks into two or more peaks. When long simulation trajectories are available, it turned out to be very useful to analyze also shorter versions of the generated time series while increasing the amount analyzed data stepwise. This revealed that the Burg spectrum already identifies features in short data sets which later also become visible in the Fourier spectrum. The sequential analysis helps to identify regions were the peak identification is converged as well as rapidly changing regions. The confidence in the accuracy of the peaks in the difference spectrum can be increased this way. Results obtained from the Yule-Walker or Maximum Entropy method do not justify the drawbacks of introducing a parameter based model.

Ground state QM/MM simulations of GFP in combination with the model order prediction scheme revealed that excited state simulations of at least 3 ps are required to achieve acceptable frequency resolution. This can currently not be achieved using the available software and hardware resources at the time of writing. Considerable effort is required before anharmonic excited state GFP difference spectra can be calculated using the presented approach. Additionally, there is currently no affordable QM method which correctly describes short hydrogen bonds. This makes it especially difficult to calculate spectra for the I_0^* state in the GFP photo cycle. The 1 ps ground state simulation trajectories from figure 4.10 and 4.11 are not sufficient to compare the simulations to experimental difference spectra which is why no difference spectrum is shown. It is also unclear whether the structure generated for the I_1 state is close enough to the true I_1 state to match experimental difference spectra. It might be worth trying a multilayer ONIOM [19] approach to record one time series for the hydrogen bond network using a fast ground state method and one time series for the excited state chromophore. However, the accuracy of this approach will be lower than the presented full electronic embedding of the active pocket.

The PYP simulation results are promising. However, the calculated difference WTL-WT spectrum in figure 4.20 only considers differences in the chromophore. Changes in Protein modes are only included indirectly and will likely not be visible in the difference spectrum. The difference spectrum in figure 4.20 also shows two corresponding normal mode eigenvectors which match the target frequency. The identified mode is reasonable as the lock in the WTL mutant might suppress the illustrated motion. However, it is not guaranteed to actually correspond to the true motion. This is due to the fact that the normal mode eigenvectors were calculated in vacuum using the harmonic approximation. Experimental data is highly desirable to see if the dominant peak at 1590 cm^{-1} is also present in the experiment.

The author of this thesis was not satisfied with the way CASSCF active spaces are identified and reduced using mostly chemical intuition. Therefore, considerable effort was put into automatically reducing the CASSCF active space by calculating all possible one step reductions. The resulting active spaces for PYP and GFP suffered from convergence problems and poor performance. The developed scoring function is clearly missing part of the information required to select good active space reductions. Considering that the reduction process has to be performed once per system and takes less than one day, it is likely not worth the effort to pursuit this automated reduction scheme any further.

The calculated vibrational spectra take the full protein environment into account without relying on harmonic approximations. Spectral amplitudes are recovered from the simulated protein dynamics in absolute values. This is a clear improvement over the relative normal mode amplitudes. No modifications to the QM/MM code other than recording the dipole moment time series with respect to the center of mass motion of the QM region are required. The only approximation made for calculating dipole time series based spectra is that the electric field of the infrared photon is only a small perturbation to the molecular wave function. The identified minimum required trajectory lengths of 2-3 ps are barely reachable with post Hartree Fock methods but certainly not impossible to simulate. Vibrational spectra should be representative and should not be calculated from single trajectories but rather consist of a statistically independent ensemble of many trajectories. The progress in computer performance does no longer justify using single trajectory data for quantitative predictions. The calculation of high level QM/MM difference spectra was shown to be a practical method that can be applied using today's computing resources.

Normal mode analysis remains a popular tool for calculating vibrational spectra because it is simple to use not because it is very accurate. The comparative results presented for PYP in figures 4.18 and 4.19 clearly show that vacuum NMA spectra do not reflect the in protein behavior of the chromophore. Comparison to the experimental difference spectra will likely support this statement further.

6 Conclusions

Anharmonic vibrational difference spectra can be calculated from QM/MM simulations including the full protein environment. The developed simulation scheme is capable of calculating vibrational difference spectra for charged QM regions and excited states.

The parametric Burg method was identified as a valuable alternative to state of the art Fourier dipole time series analysis. An objective Burg model order estimation method was developed based on the normal mode frequency density.

6.1 Outlook

Comparison to experimental IR difference spectra is the next logical step in the development of the presented method. Without experimental data, the quality of the calculated results is difficult to evaluate.

The presented model order estimation scheme would greatly benefit from including the large number of protein degrees of freedom in the low frequency range below 1000 cm^{-1}. This can easily be included into the test signal in terms of many additional low frequency sinusoid.

Currently, the presented time series analysis of QM/MM trajectories does not allow a visual inspection of the motion corresponding to the identified peaks in the spectrum. Promising methods for calculating normal coordinates without the harmonic approximation exist in the literature [48] but have not been included in this thesis. The short QM/MM trajectory lengths will likely reduce the quality of the identified modes.

Bibliography

[1] A. Amadei, I. Daidone, A. Di Nola, and M. Aschi. Theoretical-computational modelling of infrared spectra in peptides and proteins: a new frontier for combined theoretical-experimental investigations. *Current opinion in structural biology*, 20(2):155–161, 2010. ISSN 0959-440X.

[2] P.W. Atkins, R.S. Friedman, and Inc NetLibrary. Molecular quantum mechanics. 3, 1997.

[3] C.I. Bayly, P. Cieplak, W. Cornell, and P.A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 1993.

[4] A.D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *Chem. Phys*, 98(1):5648–5652, 1993.

[5] H.J.C. Berendsen. *Simulating the physical world.* Cambridge University Press Cambridge, UK, 2007.

[6] P.H. Berens and K.R. Wilson. Molecular dynamics and spectra. i. diatomic rotation and vibration. *The Journal of Chemical Physics*, 74:4872, 1981.

[7] R.B. Blackman, J.W. Tukey, J.W. Tukey, and J.W. Tukey. *The measurement of power spectra from the point of view of communications engineering.* Dover New York., 1959.

[8] M. Born and R. Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[9] B. Brauer, M. Pincu, V. Buch, I. Bar, J.P. Simons, and R.B. Gerber. Vibrational spectra of α-glucose, β-glucose, and sucrose: Anharmonic calculations and experiment. *The Journal of Physical Chemistry A*, 2011.

[10] P.M.T. Broersen. *Automatic autocorrelation and spectral analysis.* Springer-Verlag New York Inc, 2006.

[11] M. Buchner, B.M. Ladanyi, and R.M. Stratt. The short-time dynamics of molecular liquids. Instantaneous-normal-mode theory. *The Journal of chemical physics*, 97:8522, 1992.

[12] John Parker Burg. *Maximum Entropy Spectral Analysis.* PhD thesis, Stanford University, 1975.

[13] J. Chai and M. Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics*, 10:6615, 2008.

[14] B. Champagne, E.A. Perpete, D. Jacquemin, S.J.A. van Gisbergen, E.J. Baerends, C. Soubra-Ghaoui, K.A. Robins, and B. Kirtman. Assessment of Conventional Density Functional Schemes for Computing the Dipole Moment and (Hyper) polarizabilities of Push- Pull [pi]-Conjugated Systems†. *J. Phys. Chem. A*, 104(20):4755–4763, 2000.

[15] M. Cho, G.R. Fleming, S. Saito, I. Ohmine, and R.M. Stratt. Instantaneous normal mode analysis of liquid water. *The Journal of chemical physics*, 100: 6672, 1994.

[16] L. Cohen. Generalization of the wiener-khinchin theorem. *Signal Processing Letters, IEEE*, 5(11):292–294, 1998.

[17] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[18] B. Hess A. R. van Buuren E. Apol P. J. Meulenhoff D. P. Tieleman A. L. T. M. Sijbers K. A. Feenstra R. van Drunen D. van der Spoel, E. Lindahl and H. J. C. Berendsen. *Gromacs User Manual version 4.5.4*, 2010.

[19] S. Dapprich, I. Komáromi, K.S. Byun, K. Morokuma, and M.J. Frisch. A new oniom implementation in gaussian98. part i. the calculation of energies, gradients, vibrational frequencies and electric field derivatives. *Journal of Molecular Structure: THEOCHEM*, 461:1–21, 1999.

[20] Mariangela Di Donato, Luuk J.G.W. van Wilderen, Ivo H.M. Van Stokkum, Thomas Cohen Stuart, John T.M. Kennis, Klaas J. Hellingwerf, Rienk van Grondelle, and Marie Louise Groot. Proton transfer events in gfp. *submitted*, 2011.

[21] A. Dreuw and M. Head-Gordon. Single-reference ab initio methods for the calculation of excited states of large molecules. *Chemical reviews*, 105(11): 4009–4037, 2005. ISSN 0009-2665.

[22] P.L. Freddolino, C.B. Harrison, Y. Liu, and K. Schulten. Challenges in protein-folding simulations. *Nature physics*, 6(10):751–758, 2010.

[23] M.J. Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JR Cheeseman, JA Montgomery, T. Vreven, KN Kudin, JC Burant, et al. Gaussian 03, revision c. 02. 2008.

[24] M.P. Gaigeot, R. Vuilleumier, and M. Martinez. Infrared spectroscopy in the gas and liquid phase from first principle molecular dynamics simulations-application to small peptides. 2008.

[25] J.F. Giovannelli, G. Demoment, and A. Herment. A Bayesian method for long AR spectral estimation: A comparative study. *Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions on*, 43(2):220–233, 1996. ISSN 0885-3010.

[26] DA Gray. Maximum entropy spectrum analysis technique-a review of its theoretical properties. *Technical Report 1812*, 1977.

[27] S. Habuchi, R. Ando, P. Dedecker, W. Verheijen, H. Mizuno, A. Miyawaki, and J. Hofkens. Reversible single-molecule photoswitching in the gfp-like fluorescent protein dronpa. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9511, 2005.

[28] M. Head-Gordon. Systematic optimization of long-range corrected hybrid density functionals. *The Journal of chemical physics*, 128:084106, 2008.

[29] Dermot Hegarty and Michael Robb. Application of unitary group methods to configuration interaction calculations. *Molecular Physics*, 38:1795–1812, 1979. doi: 10.1080/00268977900102871.

[30] S.W. Hell and J. Wichmann. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics letters*, 19(11):780–782, 1994.

[31] B. Hess, H. Bekker, H.J.C. Berendsen, and J.G.E.M. Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[32] P. Hohenberg, W. Kohn, et al. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964.

[33] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65 (3):712–725, 2006.

[34] K.K. Irikura, R.D. Johnson III, and R.N. Kacker. Uncertainties in scaling factors for ab initio vibrational frequencies. *The Journal of Physical Chemistry A*, 109(37):8430–8437, 2005.

[35] F. Jensen. *Introduction to computational chemistry*. Wiley, 2007.

[36] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79:926, 1983.

[37] A. Kaczmarek, M. Shiga, and D. Marx. Quantum effects on vibrational and electronic spectra of hydrazine studied by “on-the-fly” ab initio ring polymer molecular dynamics†. *The Journal of Physical Chemistry A*, 113(10):1985–1994, 2009.

[38] M. Kaledin, A.L. Kaledin, J.M. Bowman, J. Ding, and K.D. Jordan. Calculation of the vibrational spectra of h5o2+ and its deuterium-substituted isotopologues by molecular dynamics simulations†. *The Journal of Physical Chemistry A*, 113 (26):7671–7677, 2009.

[39] E.R. Kanasewich. *Time sequence analysis in geophysics*. University of Alberta Press, 1975.

[40] J.L. Klepeis, K. Lindorff-Larsen, R.O. Dror, and D.E. Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Current opinion in structural biology*, 19(2):120–127, 2009.

[41] W. Kohn, LJ Sham, et al. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

[42] W. Kohn, A.D. Becke, and R.G. Parr. Density functional theory of electronic structure. *The Journal of Physical Chemistry*, 100(31):12974–12980, 1996.

[43] E. Krügel. *The physics of interstellar dust.* Taylor & Francis, 2003.

[44] H.P. Lamichhane and G. Hastings. Calculated vibrational properties of pigments in protein binding sites. *Proceedings of the National Academy of Sciences*, 108(26):10526, 2011.

[45] A.R. Leach. *Molecular modelling: principles and applications.* Addison-Wesley Longman Ltd, 2001.

[46] T. Leininger, H. Stoll, H.J. Werner, and A. Savin. Combining long-range configuration interaction with short-range density functionals. *Chemical Physics Letters*, 275(3-4):151–160, 1997.

[47] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63 (4):561–580, 1975.

[48] G. Mathias and M.D. Baer. Generalized normal coordinates for the vibrational analysis of molecular dynamics simulations. *Journal of Chemical Theory and Computation.*

[49] U. Matlab. The mathworks. *Inc., Natick, MA*, 1992, 1760.

[50] G.U. Nienhaus. The green fluorescent protein: a key tool to study chemical processes in living cells. *Angewandte Chemie International Edition*, 47(47): 8992–8994, 2008.

[51] D.B. Percival and A.T. Walden. *Spectral analysis for physical applications: multitaper and conventional univariate techniques.* Cambridge Univ Pr, 1993.

[52] AR Rao and A. Durgunoglu. AR and ARMA spectral estimation. *Stochastic Hydrology and Hydraulics*, 2(1):35–50, 1988. ISSN 0931-1955.

[53] E.A. Robinson. A historical perspective of spectrum estimation. *Proceedings of the IEEE*, 70(9):885–907, 1982.

[54] M.A. Rohrdanz and J.M. Herbert. Simultaneous benchmarking of ground-and excited-state properties with long-range-corrected density functional theory. *The Journal of chemical physics*, 129:034107, 2008.

[55] M.A. Rohrdanz, K.M. Martins, and J.M. Herbert. A long-range-corrected density functional that performs well for both ground-state properties and time-dependent density functional theory excitation energies, including charge-transfer excited states. *The Journal of chemical physics*, 130:054112, 2009.

[56] C. C. J. Roothaan. New developments in molecular orbital theory. *Rev. Mod. Phys.*, 23(2):69–89, Apr 1951. doi: 10.1103/RevModPhys.23.69.

[57] E. Schrödinger. Quantisierung als eigenwertproblem (series of four publications). *Annalen der Physik*, 79-81(multiple):361–76,489–527,437–90,109–39, 1926.

[58] F. Siebert and P. Hildebrandt. *Vibrational spectroscopy in life science*, volume 3. Vch Verlagsgesellschaft Mbh, 2008.

[59] P.J. Steinbach, R.J. Loncharich, and B.R. Brooks. The effects of environment and hydration on protein dynamics: a simulation study of myoglobin. *Chemical physics*, 158(2-3):383–394, 1991.

[60] D.J. Tannor. *Introduction to quantum mechanics.* University Science Books, 2007.

[61] M.D. Towler. *Computational Methods for Large Systems*, chapter Quantum Monte Carlo, or, how to solve the many-particle Schrödinger equation accurately whilst retaining favourable scaling with system size. Wiley, 2010.

[62] T.J. Ulrych and T.N. Bishop. Maximum entropy spectral analysis and autoregressive decomposition. *Reviews of Geophysics*, 13(1):183–200, 1975. ISSN 8755-1209.

[63] T.J. Ulrych and R.W. Clayton. Time series modelling and maximum entropy. *Physics of the Earth and Planetary Interiors*, 12(2-3):188–200, 1976. ISSN 0031-9201.

[64] A. Van den Bos. Alternative interpretation of maximum entropy spectral analysis (corresp.). *Information Theory, IEEE Transactions on*, 17(4):493–494, 1971.

[65] J.J. van Thor, G. Zanetti, K.L. Ronayne, and M. Towrie. Structural events in the photocycle of green fluorescent protein. *The Journal of Physical Chemistry B*, 109(33):16099–16108, 2005.

[66] J.J. Van Thor, K.L. Ronayne, M. Towrie, and J.T. Sage. Balance between ultrafast parallel reactions in the green fluorescent protein has a structural origin. *Biophysical journal*, 95(4):1902–1912, 2008.

[67] O.A. Vydrov and G.E. Scuseria. Assessment of a long-range corrected hybrid functional. *The Journal of chemical physics*, 125:234109, 2006.

[68] O.A. Vydrov, G.E. Scuseria, and J.P. Perdew. Tests of functionals for systems with fractional electron number. *The Journal of Chemical Physics*, 126:154109, 2007.

[69] J. Wang, P. Cieplak, and P.A. Kollman. How well does a restrained electrostatic potential(resp) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000.

[70] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.

[71] J. Wang, W. Wang, P.A. Kollman, and D.A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling*, 25(2):247–260, 2006.

[72] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3): 765–784, 1984.

[73] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, et al. Molpro, version 2010.1, a package of ab initio programs. 2010. see.

[74] J. Xu, Y. Zhang, and G.A. Voth. Infrared Spectrum of the Hydrated Proton in Water. *The Journal of Physical Chemistry Letters*, 2:81–86, 2011. ISSN 1948-7185.

[75] S. Yamaguchi, H. Kamikubo, K. Kurihara, R. Kuroki, N. Niimura, N. Shimizu, Y. Yamazaki, and M. Kataoka. Low-barrier hydrogen bond in photoactive yellow protein. *Proceedings of the National Academy of Sciences*, 106(2):440, 2009.

[76] F. Yang, L.G. Moss, G.N. Phillips Jr, et al. The molecular structure of green fluorescent protein. *Nature biotechnology*, 14(10):1246–1251, 1996.

[77] S. Yang and M. Cho. IR spectra of N-methylacetamide in water predicted by combined quantum mechanical/molecular mechanical molecular dynamics simulations. *The Journal of chemical physics*, 123:134503, 2005.

[78] G.U. Yule. *On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers*, volume 226. The Royal Society, 1927.

[79] M.C. Zerner. Perspective on "new developments in molecular orbital theory". *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 103(3):217–218, 2000.

7 Appendix

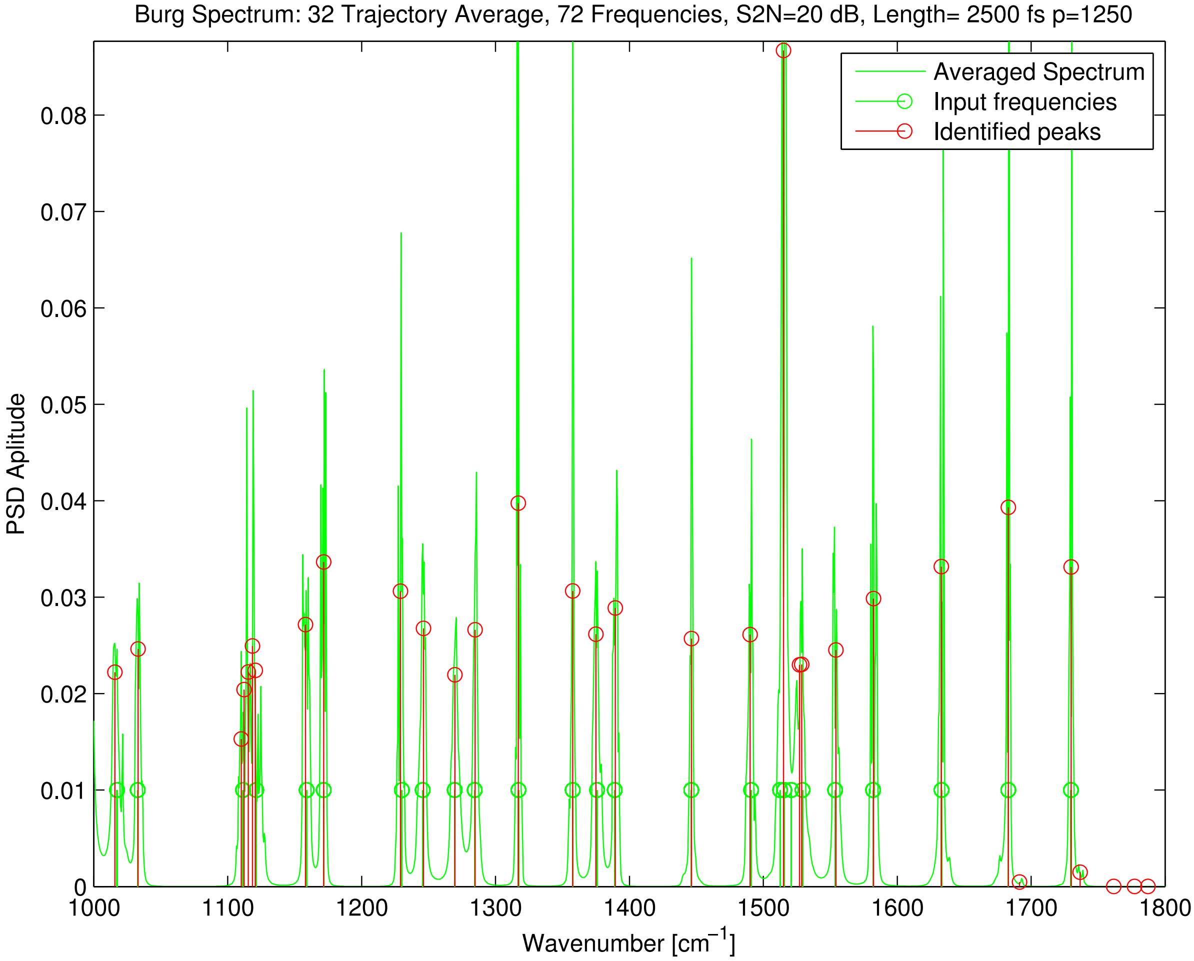

Figure 7.1: Portion of Burg PYP like 2000 fs spectrum with 72 NMA frequencies (green); Burg peaks (red); Burg spectrum (green line); no low pass filter was applied, raw spectrum shown.

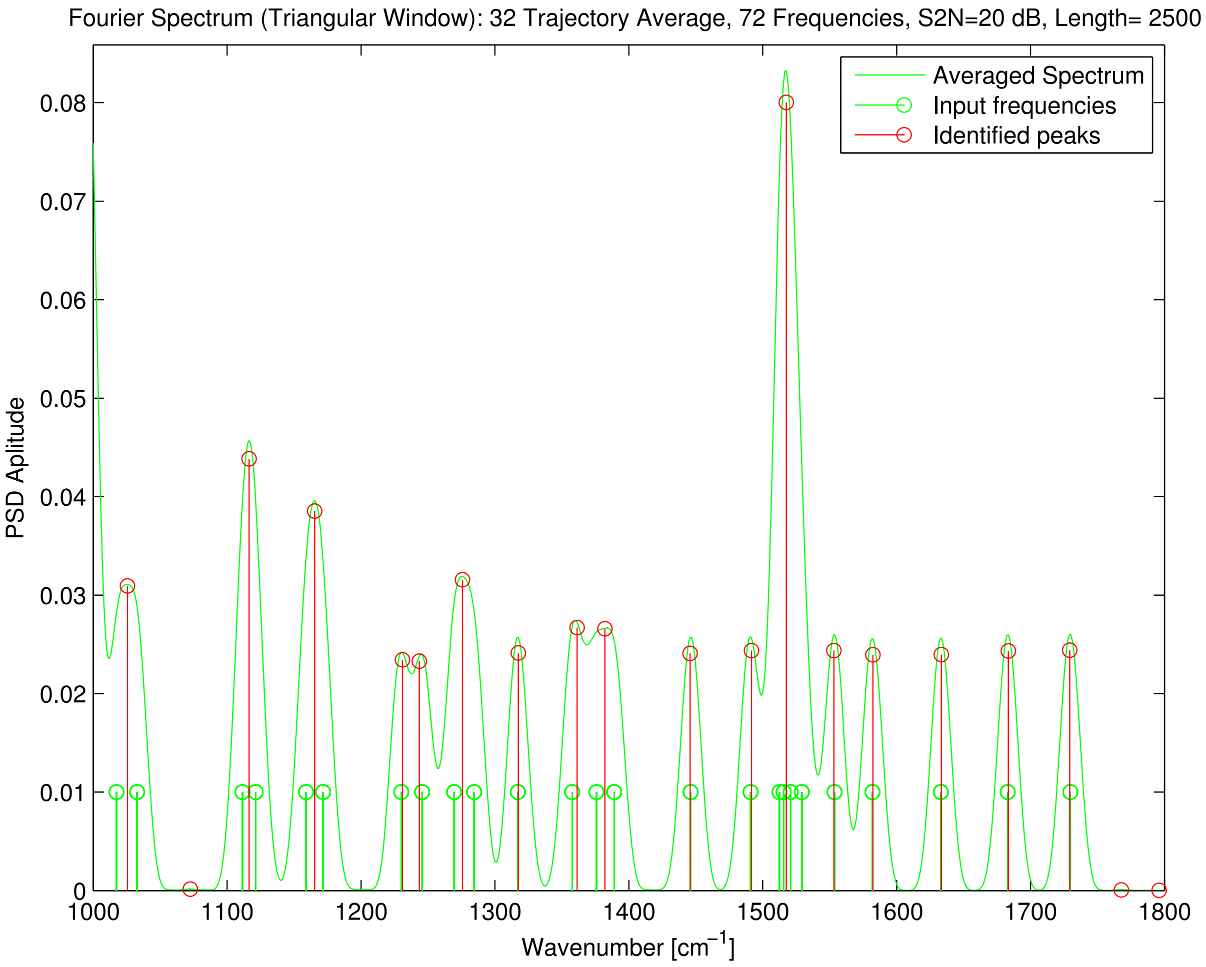

Figure 7.2: Portion of Fourier PYP like 2000 fs spectrum with 72 NMA frequencies (green); Fourier peaks (red); Fourier spectrum (green line).

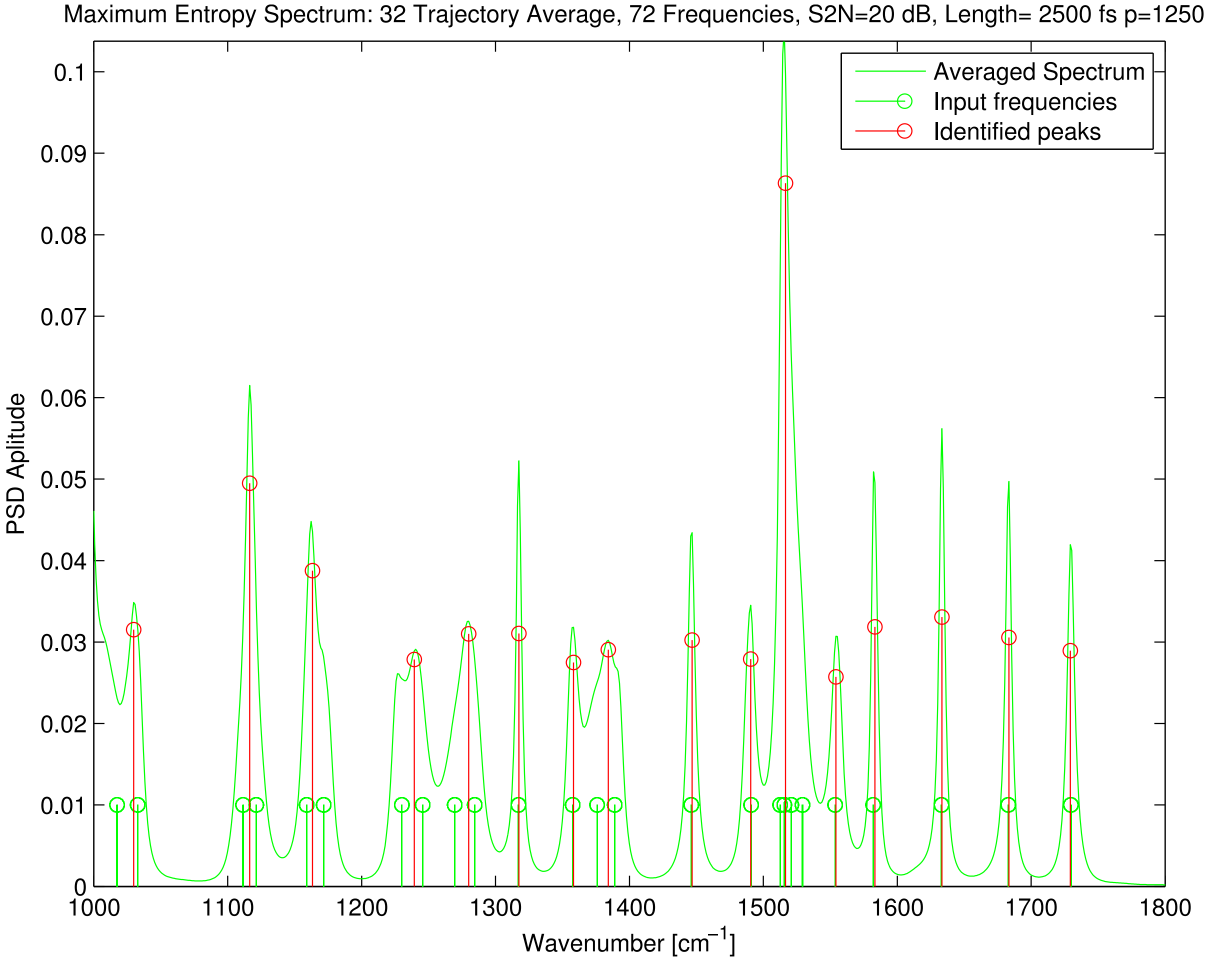

Figure 7.3: Portion of maximum entropy PYP like 2000 fs spectrum with 72 NMA frequencies (green); maximum entropy peaks (red); maximum entropy spectrum (green line).

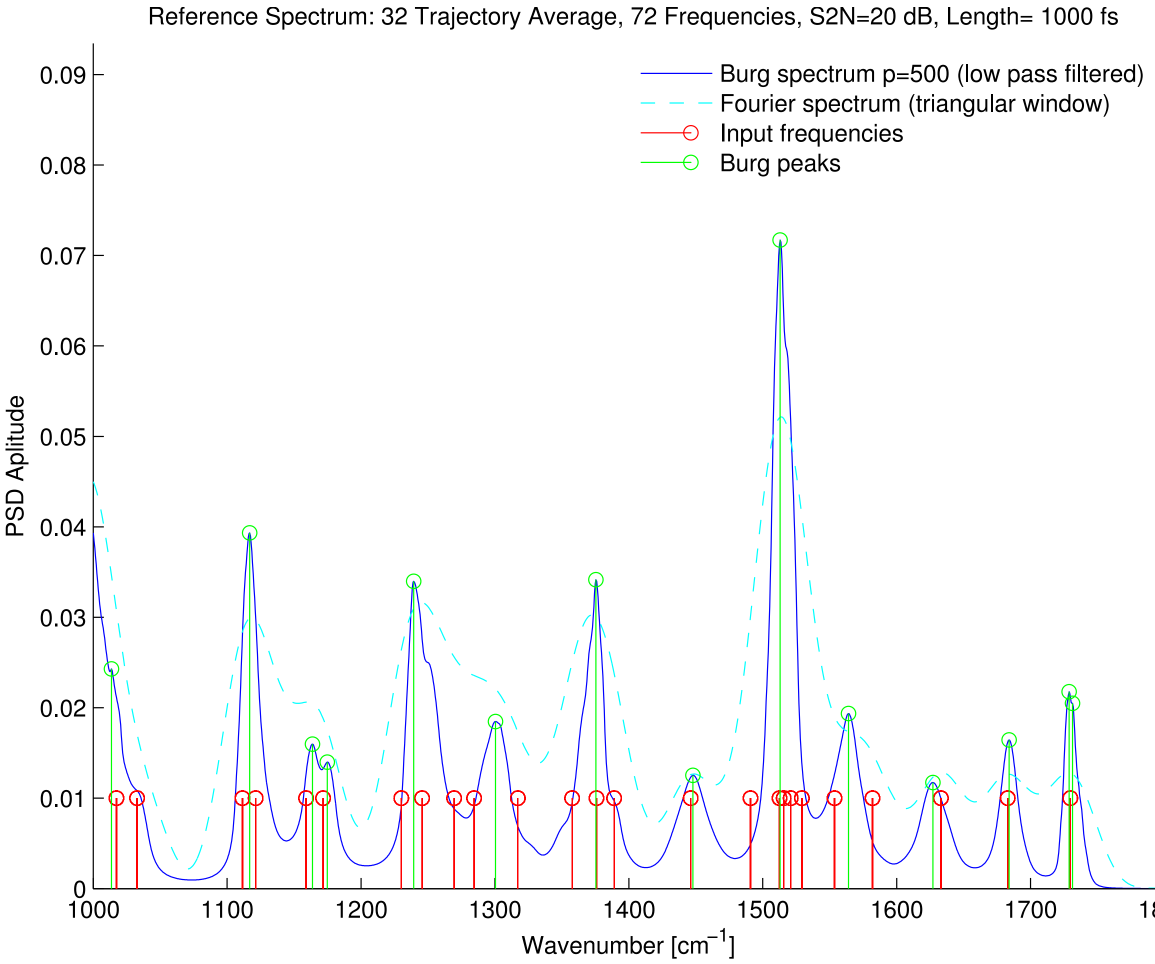

Figure 7.4: Portion of PYP like 1000 fs spectrum with 72 NMA frequencies (red); Burg peaks (green); Fourier spectrum (cyan); Burg spectrum (blue); Burg peak height prediction and frequency identification is poor; Fourier spectral resolution is low.

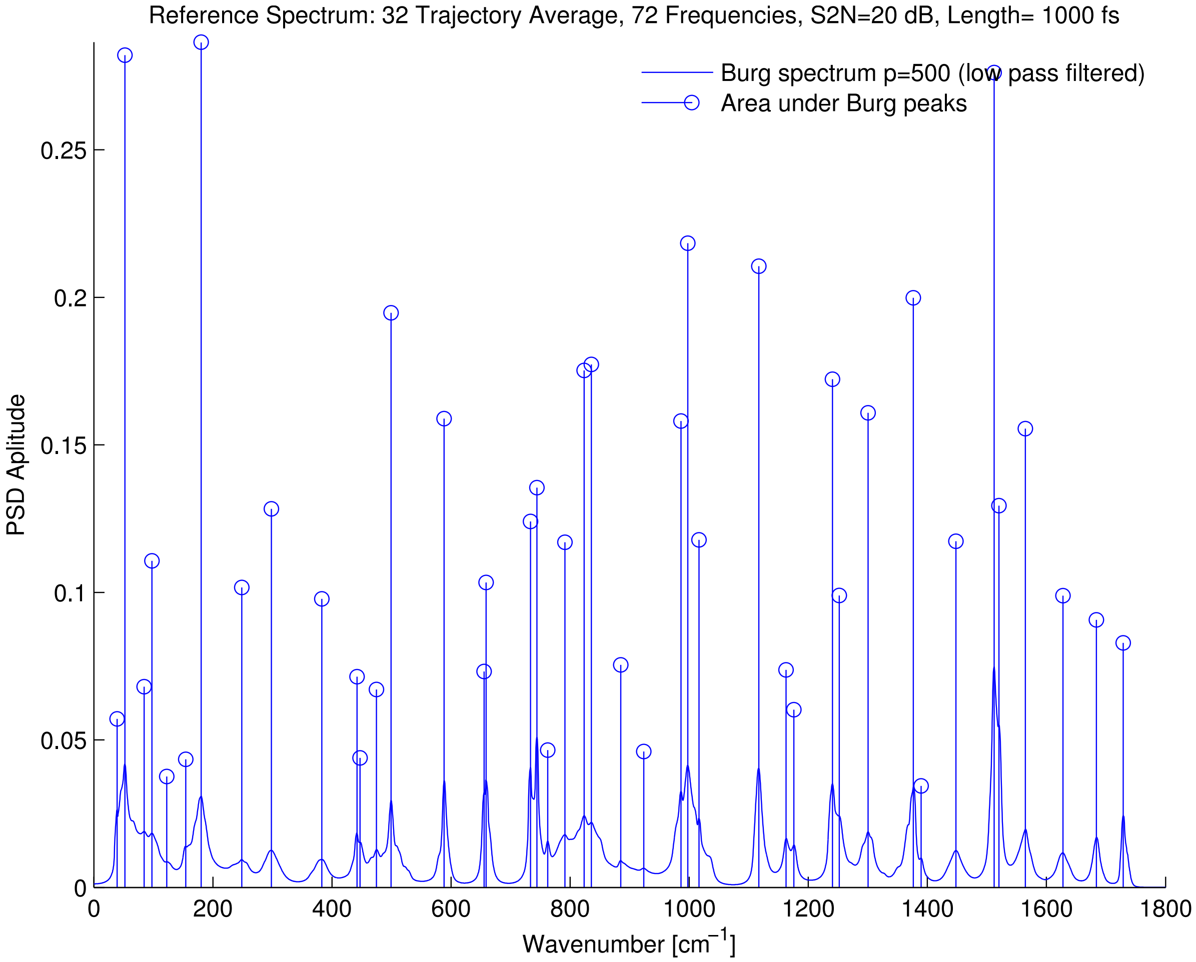

Figure 7.5: PYP like 1000 fs spectrum from figure 7.4; Burg spectrum (blue lines); integrated area under peaks (blue dots); variance in power (area under peaks) in too large for accurate spectral estimation.

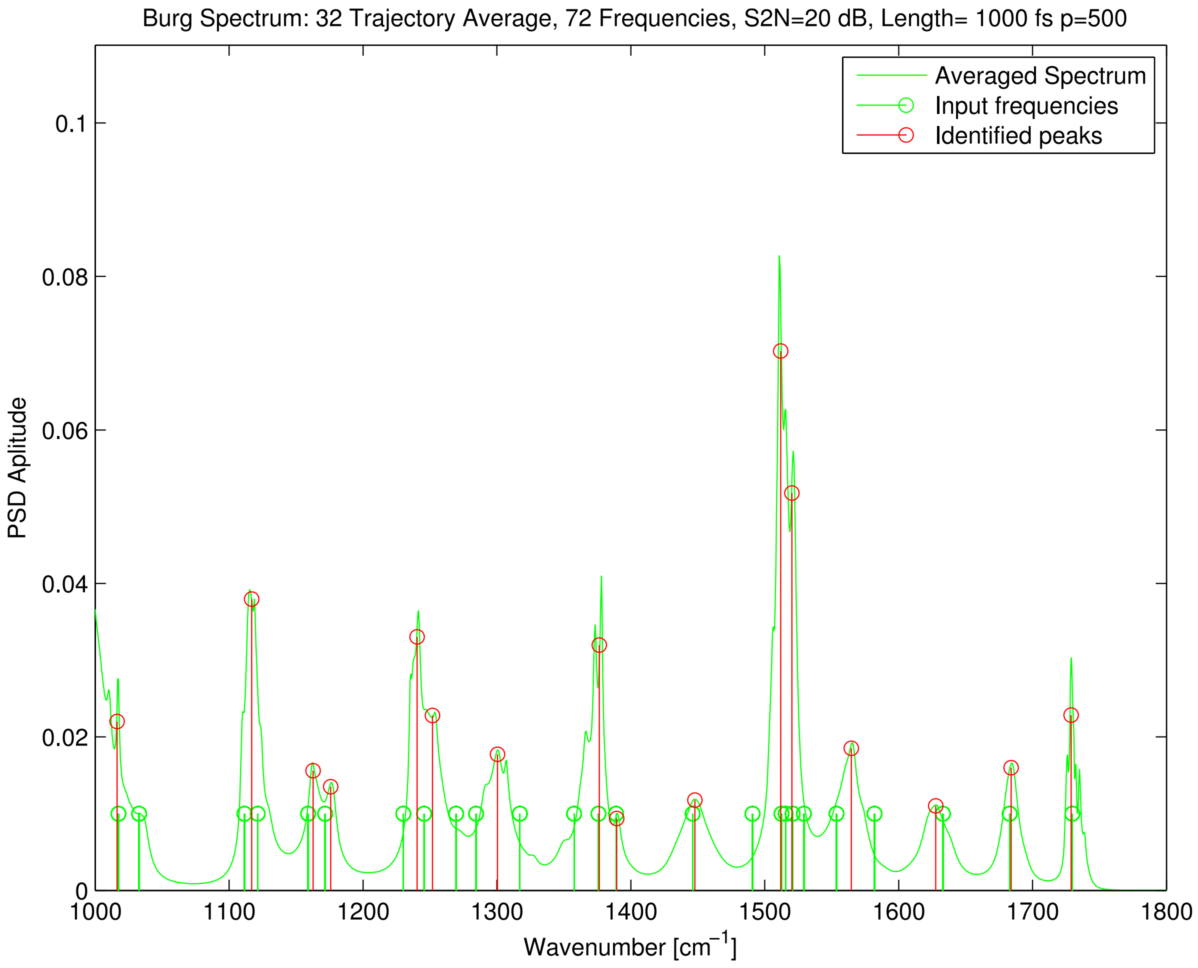

Figure 7.6: Portion of Burg PYP like 1000 fs spectrum with 72 NMA frequencies (green); Burg peaks (red); Burg spectrum (green line); no low pass filter was applied, raw spectrum shown.

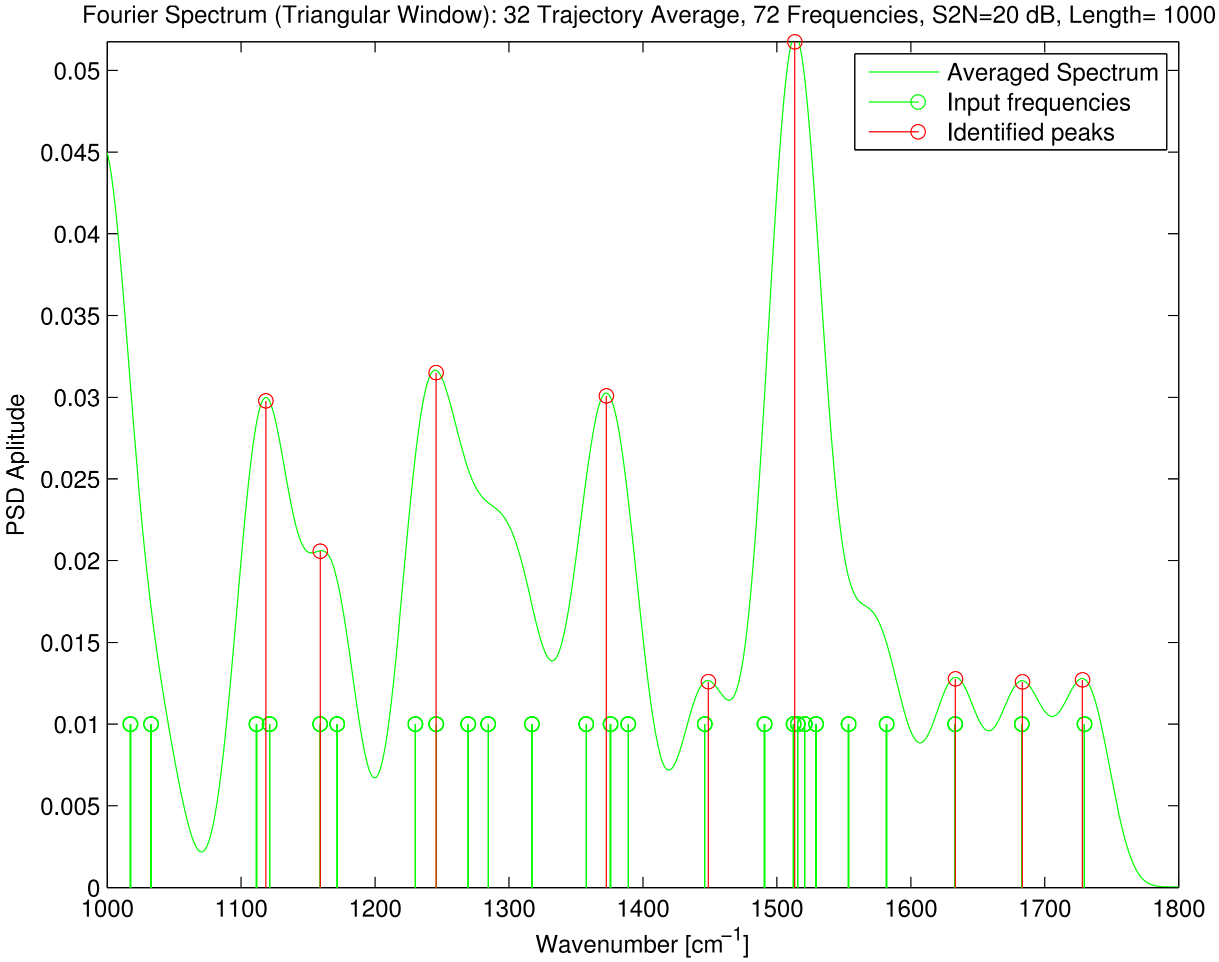

Figure 7.7: Portion of Fourier PYP like 1000 fs spectrum with 72 NMA frequencies (green); Fourier peaks (red); Fourier spectrum (green line).

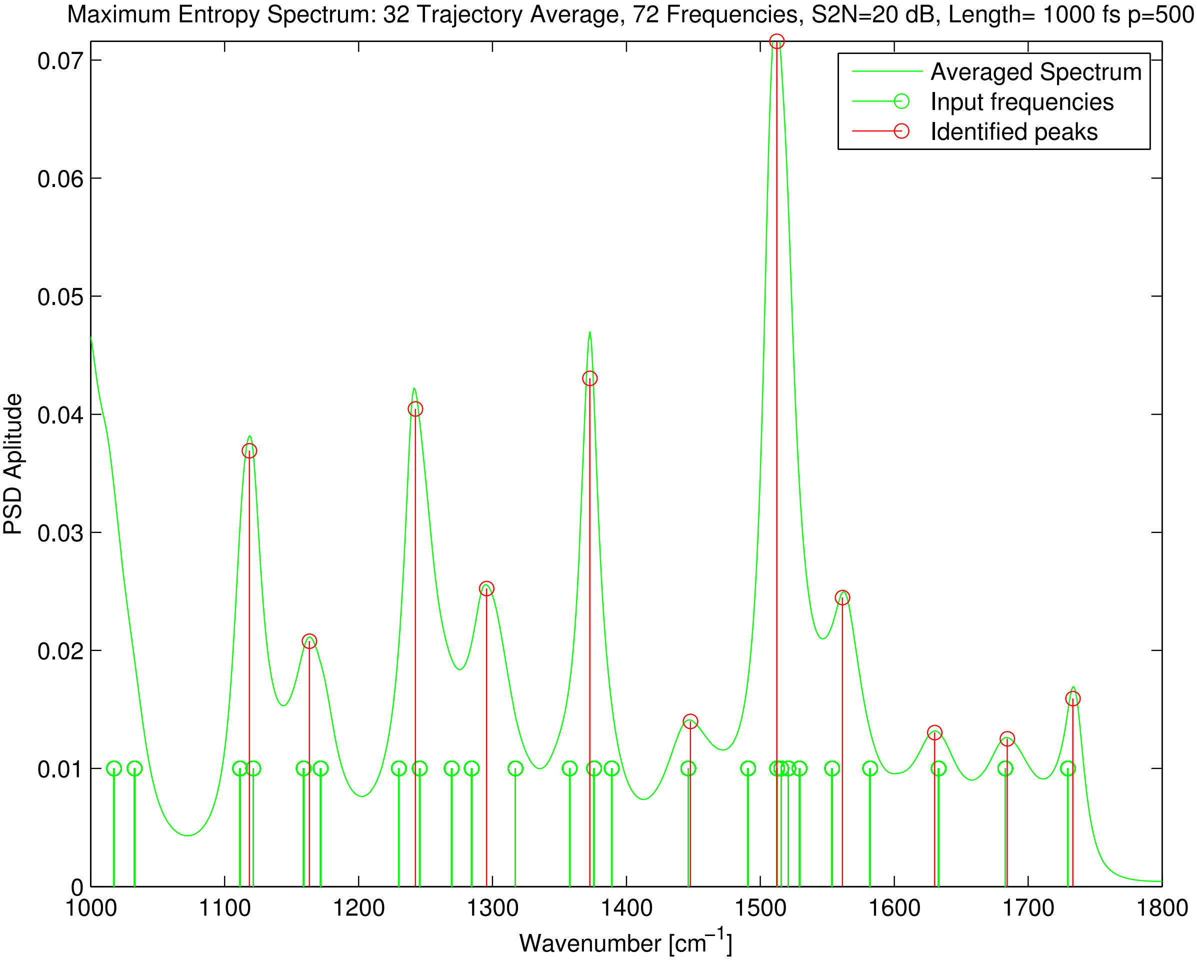

Figure 7.8: Portion of maximum entropy PYP like 1000 fs spectrum with 72 NMA frequencies (green); maximum entropy peaks (red); maximum entropy spectrum (green line).

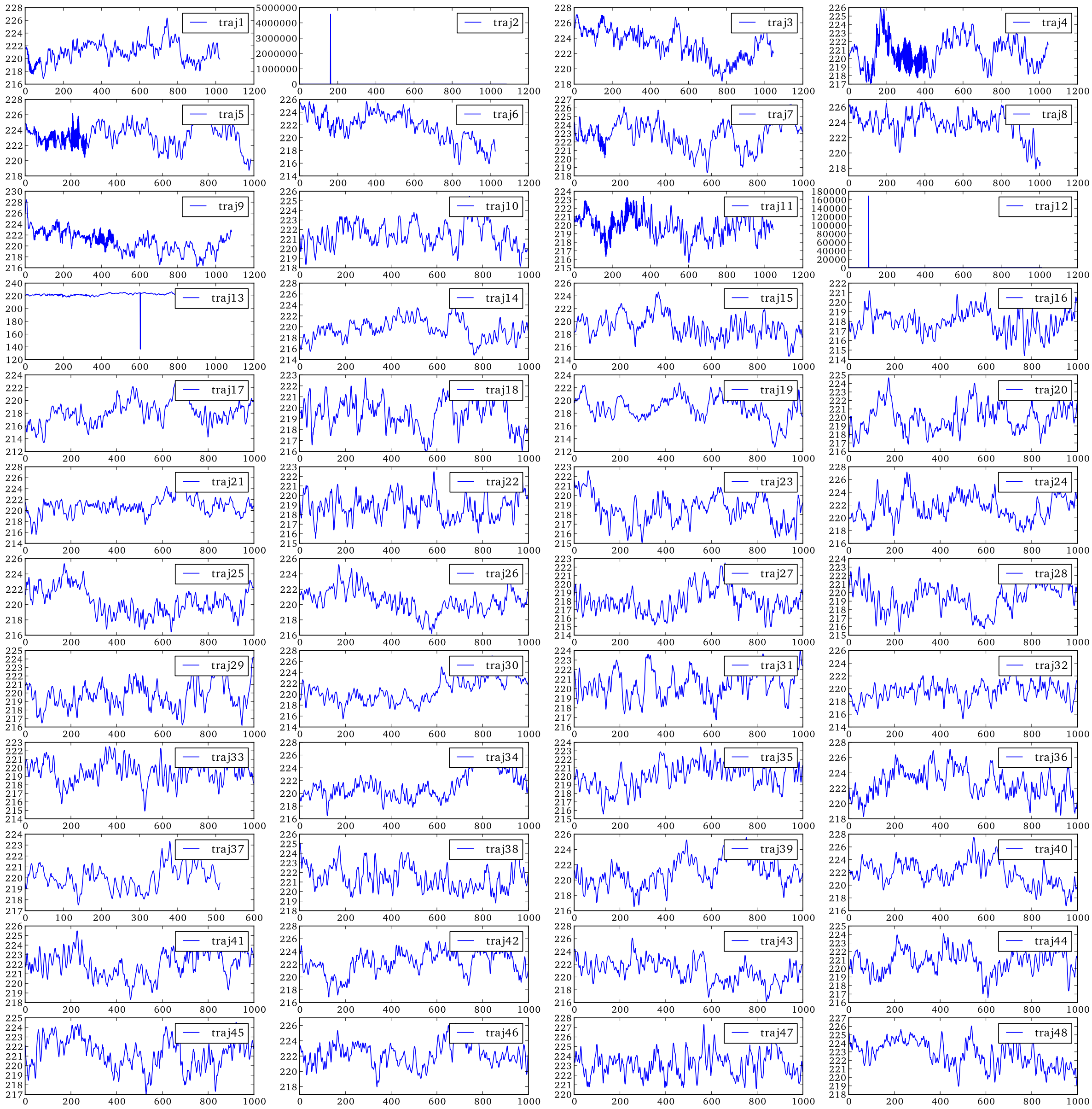

Figure 7.9: GFP ground state dipole moment time series from 48 1000fs QM/MM simulations B3LYP/6-31g*; data excluded from spectral analysis: trajectories 1-13 due to failure of remote gwdg storage system causing multiple uncontrolled job restarts; trajectory 30 due to breaking of H-bond network; trajectory 37 due to slow performance of cluster node.

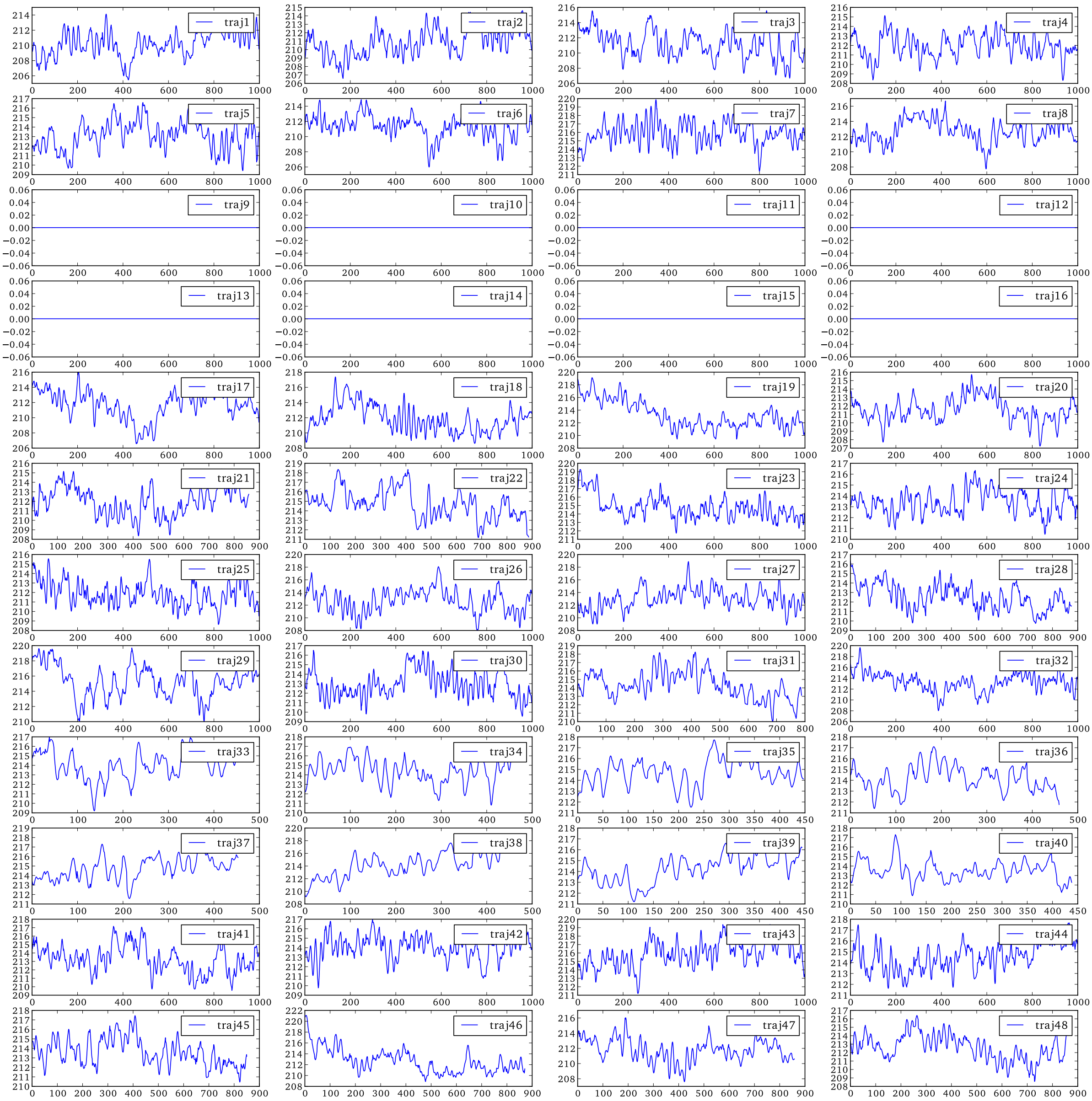

Figure 7.10: GFP I_1 state dipole moment time series from 48 1000fs QM/MM simulations B3LYP/6-31g*; data excluded from spectral analysis: trajectories 9-16 due to machine failure; trajectories 33-40 due to slow performance of cluster nodes L«1000fs; trajectories 7,19,22,23,29 due to H-bond breaking or bad conformations; 41,42,46 starting structures showed twist in GLU222

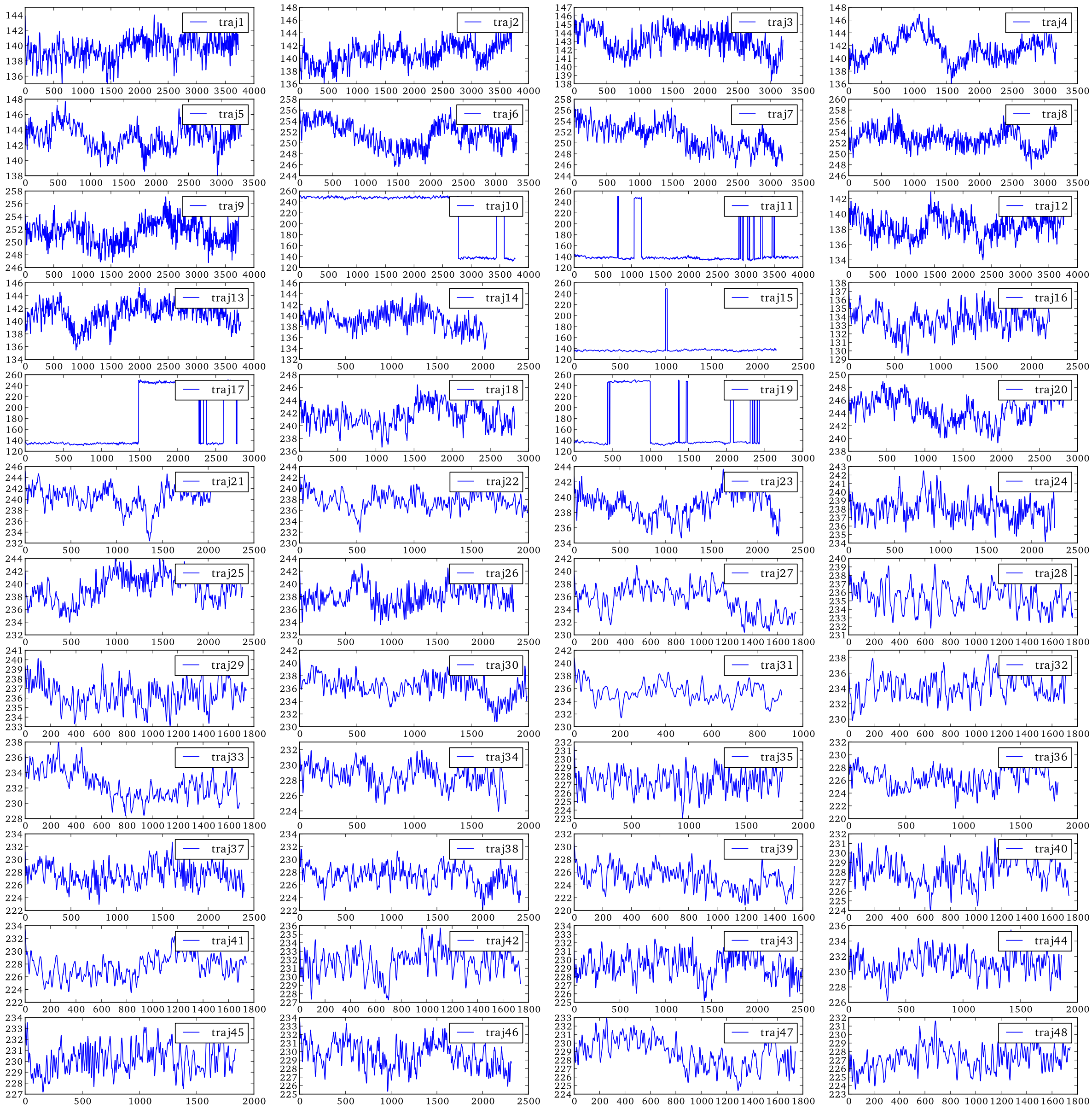

Figure 7.11: PYP WT ground state dipole moment time series from 48 QM/MM simulations B3LYP/6-31g*; data used for spectral analysis: trajectories 1, 2, 3, 5, 6, 7, 8, 9, 12, 13; only the longest trajectories were considered to improve spectral resolution.

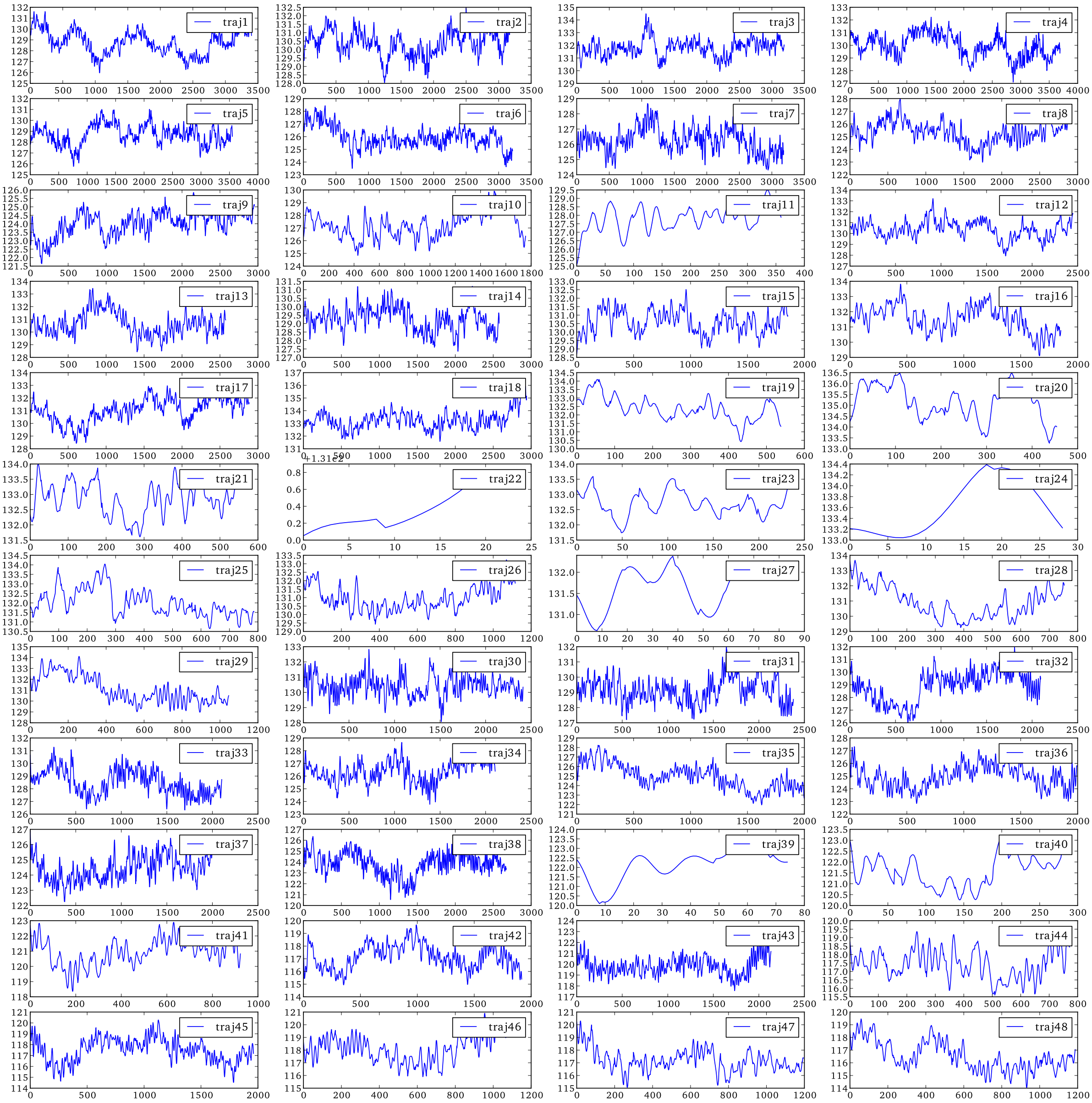

Figure 7.12: PYP WTL ground state dipole moment time series from 48 QM/MM simulations B3LYP/6-31g*; data used for spectral analysis: trajectories 1, 2, 3, 4, 5, 6, 7, 8, 9, 17, 18; only the longest trajectories were considered to improve spectral resolution.

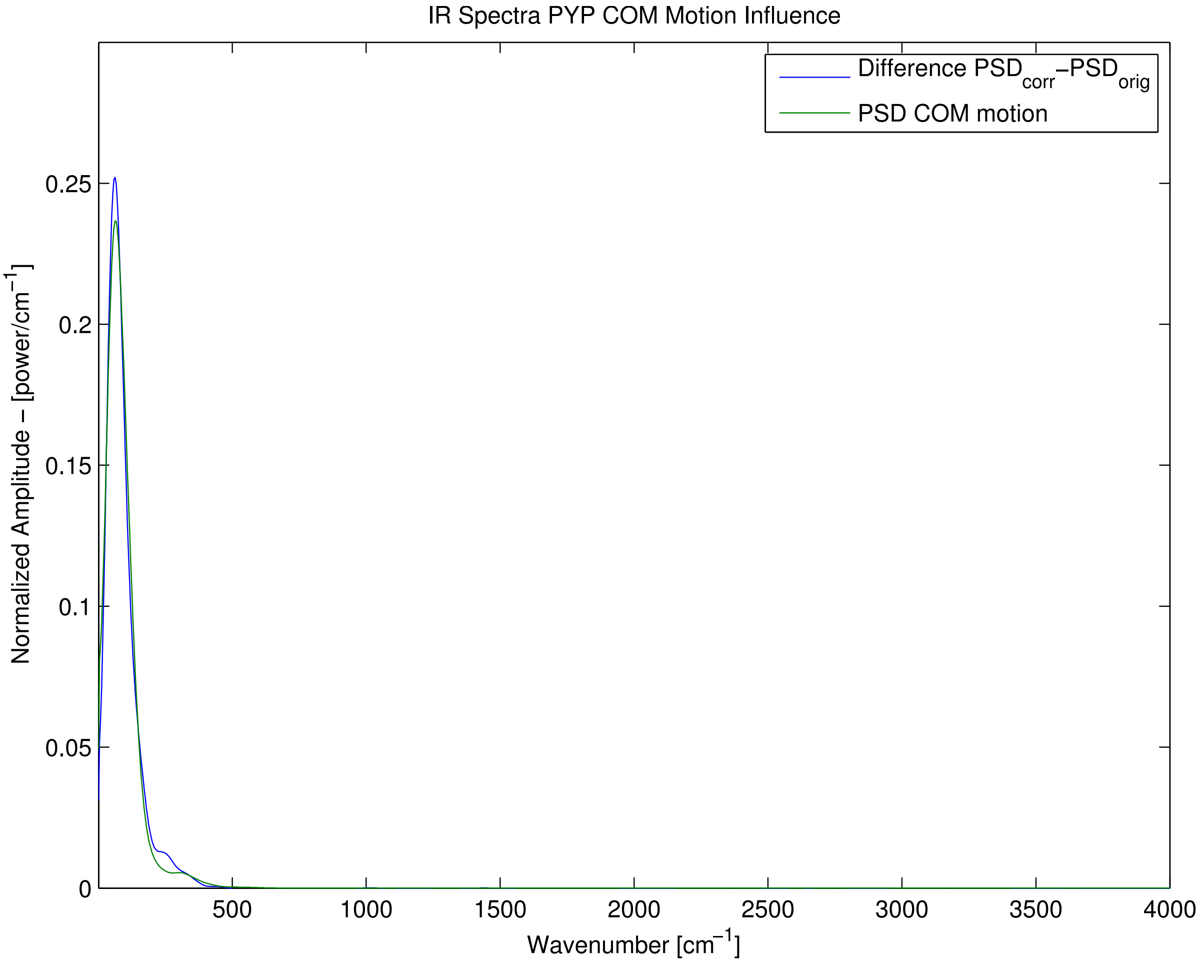

Figure 7.13: Effect of removing the center of mass motion (COM) from the dipole moment time series; COM spectral contribution localized below 1000 cm^{-1}

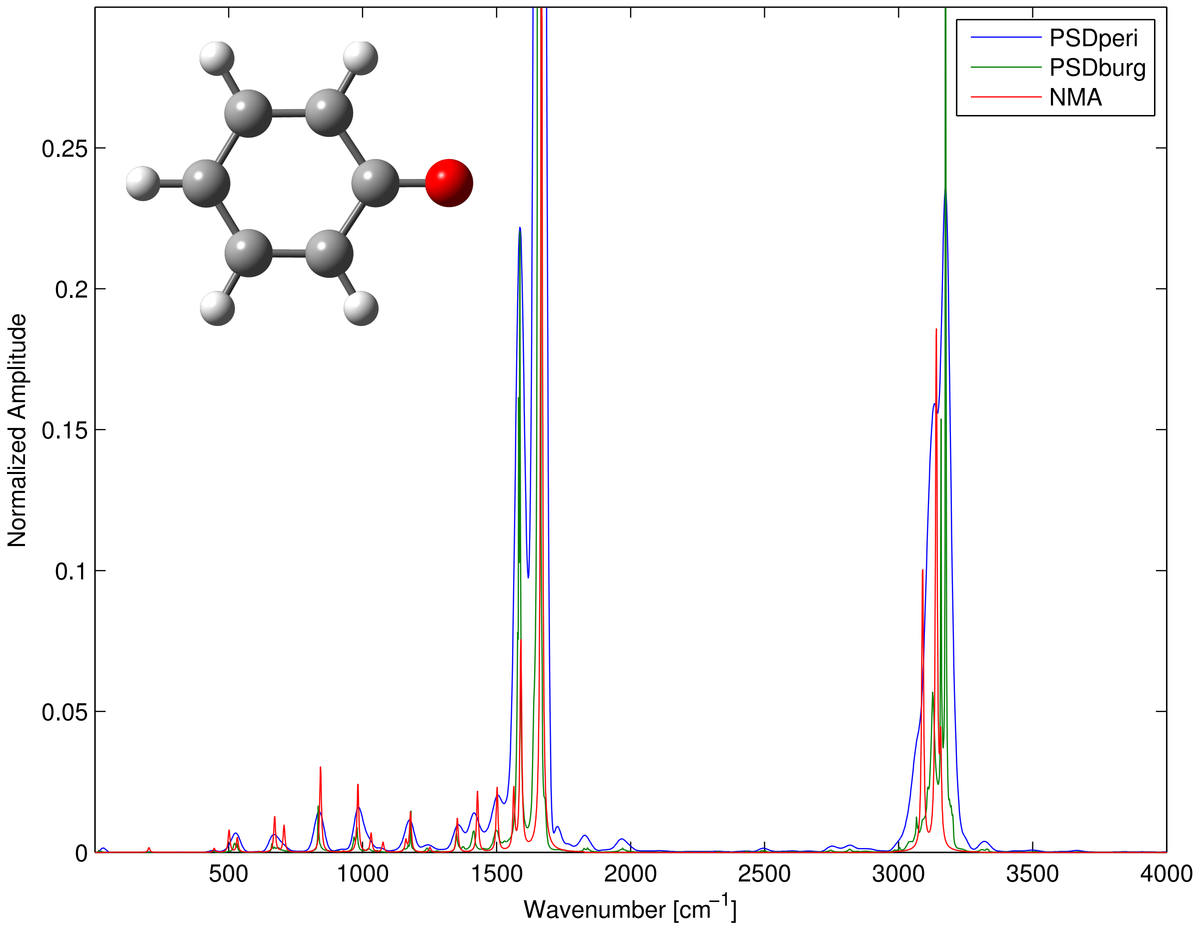

Figure 7.14: Reference calculation for the depicted molecule in vacuum using both the Normal Mode Analysis and dipole time series analysis, no protein environment was present; blue line: Burg spectrum; green line: Fourier spectrum; red line: normal mode spectrum. Small frequency shifts are visible which are not environment related. Spectrum calculated using 2 cm^{-1} resolution which might explain the shifts due to numerical uncertainty.

Acknowledgment

Financial Acknowledgments: Marloes Groot for supporting my scholarship applications and Helmut Grubmüller for partially financing my studies in Amsterdam, thus keeping me from flipping burgers at McDonalds. MPIBPC for affordable housing.

Scientific Acknowledgments: Gerrit Groenhof, Marloes Groot and Helmut Grubmüller for fruitful discussions and critical input throughout my project; Carsten Kutzner and Sören Wacker for many helpful discussions on all aspects of my work; Daan P. Geerke for introducing me to polarizable force fields; Bela Voß for quoting me in his thesis and discussions; Maarten Wolf and Martin Hoefling for highly valued discussions on the multipole expansion and charged dipole moments; David Köpfer, Ulrich Zachariae and Jan Peters for listening; Maike Clemens, Mehdi Davari, Carl Burmeister and Ludger Inhester for discussions on various quantum chemistry problems; Ulrike Gerischer helping me with by poster designs; Ansgar Esztermann and Martin Fechner for administrating the department's supreme computing facilities; Eveline Heinemann for keeping everyone organized.

Many thanks again to Marloes Groot and Gerrit Groenhof for letting me become part of their groups.

Declaration

This thesis is mine and I wrote it myself. It presents my research work in the specified time period. Contributions by others were indicated to my best knowledge by reference to the literature. Research input and discussions were acknowledged. All figures in this thesis were generated by me but you may use them under the CC0 1.0 Universal license (http://creativecommons.org/publicdomain/zero/1.0/). All other thesis related elements may be used under Attribution 3.0 Unported license (CCBY3.0) (http://creativecommons.org/licenses/by/3.0/). I also declare that this thesis, or parts of it, were only submitted to the VU University as part of the thesis requirement of the international masters program in physics.

Göttingen, August 1, 2011

(Timo Marcel Daniel Graen)