Efficient Production of Hot Molecular Line Lists

Ahmed Faris Al-Refaie

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Physics and Astronomy

University College London

August 3, 2016

I, Ahmed Faris Al-Refaie, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Associated Publications

1. **ExoMol line lists VII: The rotation-vibration spectrum of phosphine up to 1500 K**

   C. Sousa-Silva, Ahmed F. Al-Refaie, S. N. Yurchenko and J. Tennyson

   Monthly Notices of the Royal Astronomical Society, **446**, 2337-2347, 2015;

2. **ExoMol line lists VIII: A variationally computed line-list for hot formaldehyde**

   A. F. Al-Refaie and S. N. Yurchenko and A. Yachmenev and J. Tennyson

   Monthly Notices of the Royal Astronomical Society, **448**, 1704-1714, 2015;

3. **A variationally calculated room temperature line-list for H_2O_2**

   A. F. Al-Refaie and R. I. Ovsyannikov and O. L. Polyansky and S. N. Yurchenko and J. Tennyson

   Journal of Molecular Spectroscopy, **318**, 84-90, 2015;

4. **ExoMol line lists XV: A hot line-list for hydrogen peroxide**

   A. F. Al-Refaie and O. L. Polyansky and R. I. Ovsyannikov and J. Tennyson and S. N. Yurchenko

   Monthly Notices of the Royal Astronomical Society, **461**, 1012-1022, 2016;

5. **GPU Acelerated INtensities MPI (GAIN-MPI): a new method of computing Einstein A coefficients**

   A. F. Al-Refaie, S. N. Yurchenko and J. Tennyson

   Computational Physics Communications (Submitted)

Abstract

Molecular line lists are of utmost importance in understanding and characterising the molecular composition of atmospheres from their spectra. Cool stars and exoplanets such as Hot Jupiters have temperature ranges that allow for a significant composition of molecules in their atmospheres with extremely complex and rich spectral structures. Building a comprehensive line-list to model such phenomena is a non-trivial task. Therefore efficient production is a necessity. This thesis presents three molecular line lists produced using the theoretical methodologies of the TROVE program suite.

GPU Accelerated Intensities (GAIN) is a new addition to TROVE and allows for the rapid calculation of billions of transitions by exploiting graphics processing units (GPUs) to speed up the evaluation of the line strength by almost 1000x compared to previous codes. The program's extensive usage in computing the 17 billion transitions for the hot phosphine line list SAlTY is briefly discussed.

A hot H_2CO line-list applicable to 1500 K is computed using TROVE and GAIN from a refined potential energy surface (PES) and *ab initio* dipole moment surface (DMS). Results are compared to experimental data and problems encountered from the PES refinement are discussed.

A preliminary room temperature line list for H_2O_2 is produced from a purely *ab initio* PES and DMS and compared to experimental result. The *ab initio* PES is then refined to spectroscopic accuracy and a final hot line list is produced applicable up to 1250 K.

Acknowledgements

First and foremost, I wish to thank my supervisors Prof. Jonathan Tennyson and Dr. Sergey Yurchenko for their wisdom, knowledge, guidance, encouragement and unwaivering patience for the duration of my Ph.D. Their expertise has formed the foundation of my acquired skills and knowledge and have made this thesis possible.

I wish to thank my father: Dr. Faris N Al-Refaie, my mother: Lamya Ali, my brother: Yassir Al-Refaie and my sisters: Nadia, Sara and Reem Al-Refaie for giving me unending love, support and joy in my life. No thanks will be given to the family cats Kimchi and Clooney.

The friends I have made in UCL during my time have been invaluable to me: Duncan Little and Clara Sousa-Silva have been with me since he start of my Ph.D and have given me kindness, support and happiness. My thanks go to the ExoMol team: Phillip Coles, Emil Zak, Andrey Yachmenev, Katy Chubb, Daniel Underwood, Christian Hill, Emma Barton, Barry Mant, Laura McKemmish, Oleg Polyansky, Ala'a Azzam, Renia Diamantopoulou, Alec Owens, Marie Gorman and the late Anatoly Pavlyuchko for giving me the pleasure of working with such an incredible group of people and for putting up with my terrible journal articles during our tea sessions!

To those outside of UCL: I thank Asya Yurchenko for her kindness and support and for making me laugh when I had no laughter left and to Edmund Lovell for our enlightening discussions on how good Deus Ex is. To my two best friends Sarfraz Ahmed Aziz and Rory Gleeson; Thank you being there for me for all these years.

This work was supported by the ERC under the Advanced Investigator Project 267219.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Studies of the nuclear motion problem are essential for the deeper understanding of a molecule's physical properties. Molecules that interact with electromagnetic radiation in the infrared and microwave regions will undergo a change in how the molecule rotates and/or the motion of its atoms. This interaction gives rise to a spectrum with unique features that characterise the molecule and can be exploited in probing the physical properties and conditions of remote environments. In such fields where *in situ* experimental data is unavailable and/or impractical, remote sensing and spectroscopy provides the only means of studying such environments. In particular, the study and characterisation of astrophysical atmospheres is reliant on obtaining and interpreting spectra. The atmospheres of cool stars and exoplanets have temperature ranges that allow for a significant abundance of molecules in their atmospheres. However, studying the composition and chemistry of these atmospheres requires spectroscopic data for molecules over a large range of temperatures and pressures in order to accurately characterise the spectra we obtain from them. Molecular line lists provide this spectroscopic data and can be produced either experimentally or theoretically. In principle, experimental data is the best but they are often limited in temperature applicability and spectral range. This can stem from a range of issues that include: difficulty in obtaining samples, safety of molecule, spectral limitations of equipment, resolution, available equipment time, confusing spectra, contamination, thermal volatility and other such factors. Theoretical data therefore provides a means to overcome these limitations. However, producing the-

oretical data for molecules with three or more atoms requires a huge amount of computational resources in order to achieve experimental accuracy and high temperature applicability. As such, techniques must be developed in order to overcome the computational burden required in order to compute such line lists for bigger molecules efficiently, accurately and completely.

This thesis explores the techniques and methodologies used in the ExoMol project to compute the billions of transitions needed to model spectra of polyatomics at high temperatures. The remainder of this introductory chapter will describe the application of these line lists in modelling exoplanetary atmospheres. In particular, it will focus on the concept and necessity on 'completeness' of the line lists at different temperatures and their importance in accurate modelling and characterisation. As such, this will lead into the motivation of this thesis and the ExoMol project [8].

Chapter 2 describes the theories and key methods used in production of polyatomic line lists. Presented is a derivation of the underlying approximations and the TROVE (**T**heoretical **ROV**ibrational **E**nergies) nuclear motion code [30] used in first half of the line list production pipeline. Chapter 3 describes the computation of the second half of the line list and includes the introduction of the new code dubbed GAIN (**G**PU **A**ccelerated **IN**tensities) that was developed to rapidly compute the billions of transitions required for completeness. Focus will be given on the utilization of newer accelerator hardware that was exploited to achieve this and includes a case study application of the code on the phosphine molecule to produce the SAlTY (**S**ousa-Silva, **Al**- Refaie, **T**ennyson, **Y**urchenko) line list.

Chapters 4 and 5 focus on the three molecular line lists that were produced during the course of this PhD through the application of the methods described in Chapter 2. Chapter 4 concerns the production of the hot formaldehyde (H_2CO) line list and describes the importance of the molecule, the difficulties encountered in production which include a failed initial line list that demonstrates interesting effects and the final published hot line list AYTY (**Al**- Refaie, **Y**achmenev, **T**ennyson, **Y**urchenko) compared to existing experimental data. The next chapter involves the production of two molecular line lists for the hydrogen peroxide (H_2O_2) molecule.

Similarly to H_2CO, a description of its importance to the ExoMol project is presented and general application of theory before being partitioned into two distinct parts. The first half describes the published *ab initio* room temperature line list and compares and contrasts it with the currently available experimental data. The second half involves the computation of the high temperature line list APTY (**Al**-Refaie, **P**olyansky, **T**ennyson, **Y**urchenko) for H_2O_2 and the additional work required to improve the *ab initio* line list to experimental accuracy before concluding with a comparison with both the room temperature line list and experimental data.

Chapter 6 is a short discussion on the process of computing opactities and molecular cross sections and introduces the code GEXS (**G**PU **E**xoMol **X**ross **S**ections) being developed to improve the efficiency of atmospheric modelling. The final chapter concludes the thesis with an overview of the achievements described in the thesis and of the ExoMol Project as a whole. It briefly presents further oppurtunites presented by this thesis that include applications of the computed line lists and future work on larger molecules.

1.1 Transit Spectroscopy

The methodology of extracting spectra of exoplanet atmospheres came about from the seminal paper by Seager and Sasselov [31] as well as additional papers by Brown [1] and Hubbard et al. [32] using the stellar flux during the transit of the Hot Jupiter HD 209458b as a model. As an exoplanet transits its host star, its radius can be observed. If a gaseous atmosphere is present, it will cause the apparent radius of the planet to become a function of wavelength. The change of opacity over wavelength comes from the molecular absorption of stellar flux and as a result, an absorption spectrum is superimposed in the measured stellar spectra[31]. The retrieval of the exoplanet's spectra therefore requires extensive knowledge of the host star's contribution to the overall observed flux. Great care must be taken in removing the stellar influence in the combined star-exoplanet spectra as its contribution is influenced by other processes such as stellar activity, sunspots and flares. Additionally, clouds in the exoplanet's atmosphere must also be taken into account

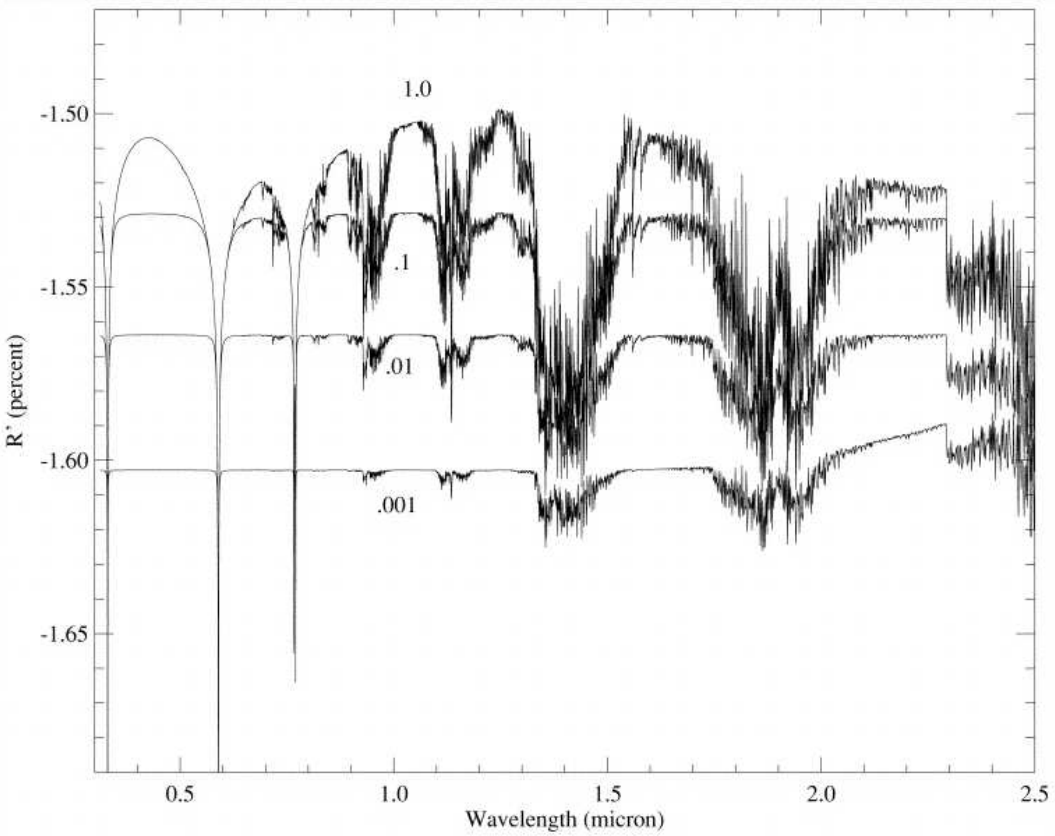

Figure 1.1: Taken from Figure 8 of Brown [1], described is the effect of a simulated transmission spectrum from an exoplanet transiting HD 209458b with cloud depths at varying pressures labelled (in bar).

as they can affect the apparent radius at different pressures and mask absorption features as seen in Figure 1.1. Hubbard et al. [32] expanded on this by considering additional effects such as Rayleigh scattering, refraction and cloud scattering. As such, detection of spectral features not only describes the atmosphere's chemistry but also cloud depth, pressure and temperature. Seager and Sasselov [31] suggested a means of detecting atmospheres by observing the wavelength region at 590 nm for a strong Na I doublet feature and Hubbard et al. [32] further suggested additional observations in the infrared for strong molecular absorption features such as H_2O. The prediction of the sodium feature was confirmed by the very first definitive observation of an exoplanet atmosphere by Charbonneau et al. [33]. Later observations in the infrared (IR) of the Hot Jupiter HD 189733b confirmed the presence of water vapour [34] and carbon monoxide[35], with a tentative observation of methane by

Swain et al. [36].

1.1.1 Atmospheric Retrieval

Transit spectroscopy provides the means with which to retrieve spectra from an exoplanet but interpreting and characterising it is a complex affair. The basis of exoplanet atmosphere analysis is through radiative transfer modelling of the observed features. This type of atmospheric analysis is not exclusive to exoplanet characterization as it is well established in extracting physical parameters and the chemistry of terrestrial and solar system atmospheres [37]. All radiative transfer modelling is fundamentally based on the radiative transfer equation that describes how radiation intensity changes as it propagates through a medium; radiation loses energy to absorption, gains energy by emission and distributes it through scattering. As radiation passes through an atmosphere, the source of its energy loss comes from absorption by molecules. The effect of absorption on radiation has two parts: the strength and the shape.

The strength of an absorption is a characteristic of the species of molecules involved, their abundance, their mixing ratios and of the temperature of their enviroment. Each molecule in an isolated system produces a unique absorption spectrum that can be described as their 'fingerprint'. This behaviour arises from the molecule gaining energy and transitioning into a different state. In particular, the IR region of the electromagnetic spectrum deals mostly in the change in a molecule's rotational and vibrational motion. These changes are known as rovibrational transitions and their intensity at an arbitrary temperature can be computed if their spectroscopic data is provided.

Transitions are discrete and are therefore represented as a single sharp line with a specific frequency and intensity. However there are various mechanisms that can cause these lines to broaden and must be taken into account in order to properly represent a molecule's spectrum. An absorber is not stationary, it moves with a velocity due to thermal energy. The velocity of molecules in a volume follows the Maxwell–Boltzmann distribution, with higher velocities becoming more likely at elevated temperatures. The veloctiy of the absorber Doppler shifts the absorption

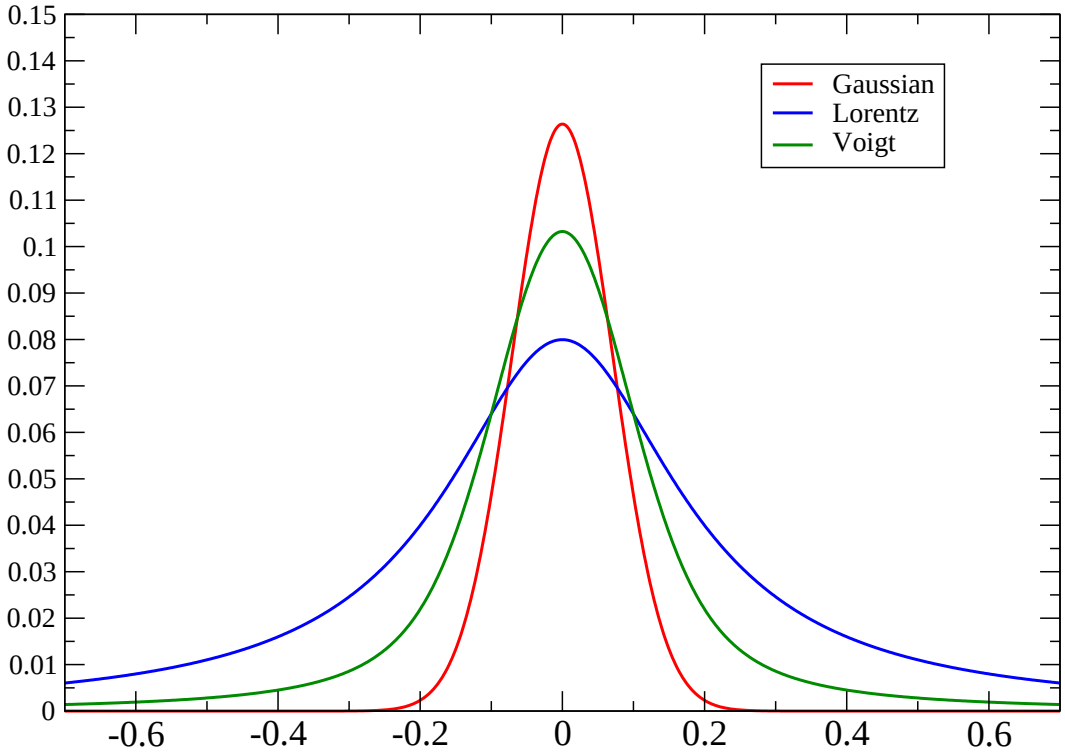

Figure 1.2: A plot of the Doppler, Lorentzian and Voigt line profiles.

frequency slightly, this appears as a broadening in the lines profile with a greater effect appearing at higher temperatures. This is known as Doppler broadening and has the form of a Gaussian line profile with associated Doppler width. Molecules in a volume are subject to collisions with surrounding molecules with the rate of collisions influenced by the pressure of their environment. An absorbing molecule that has collided causes an interruption of the absorption process shortening its characteristic time. There is therefore an uncertainty in the energy absorbed causing a broadening effect to occur. This broadening is refered to as pressure broadening and takes the form of a Lorentzian profile with associated Lorentzian width. The convolution of the two profiles is known as the Voigt profile and represents a more complete model of line broadening. A comparison of all three is seen in Figure 1.2. Therefore a knowledge of the temperature-pressure profile must be taken into account when interpreting the shape of lines.

Environments rarely contain a single species of molecule and are instead composed of many molecules in a mixture. The affect of this is that their spectra are

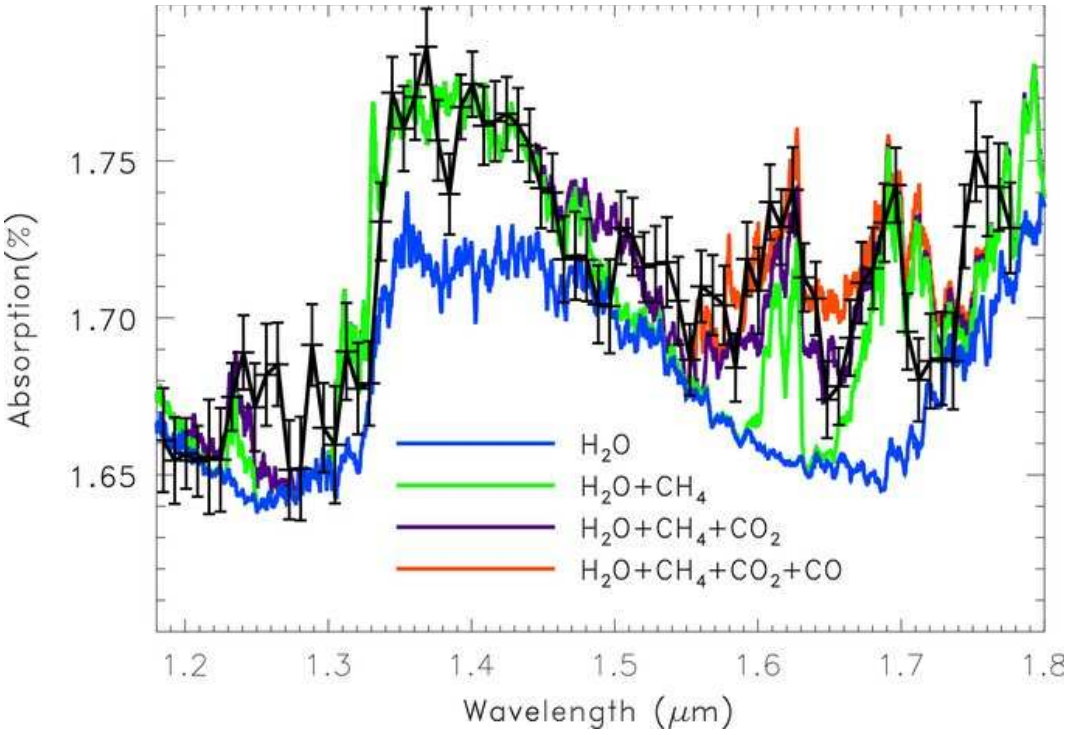

Figure 1.3: Taken from Figure 2 of Tinetti et al. [2]. Primary transit spectrum of XO-1b obtained with the Hubble–NICMOS instrument. The observed data is black and the coloured lines represent various atmospheric models consisting of H_2O, CH_4, CO and CO_2.

convoluted. The nature of this convolution is based on their abundance and mixing ratio, with molecules that are abundant and/or with strong absorption features having a greater expression in the resultant absorption spectra. It is worth noting that the inverse of this, the removal of spectra in a convolution, is also a major topic in literature. In particular, water has very strong IR features and is extremely abundant in the Earth's atmosphere, this makes ground based observation in the IR difficult as its transmission spectrum heavily masks observations. Figure 1.3 illustrates the process of characterising the molecular species in the atmosphere of the Hot Jupiter XO-1b. Here the effect of mixing molecules causes changes in the features of the overall spectra;

Radiative transfer is a forward model in the sense that it predicts data if the parameters are already known. The inverse problem is estimating or retrieving these parameters from the data. The basis of solving the inverse problem of atmospheric analysis is by replicating as closely as possible the observed data through fitting

of parameters. Older atmospheric retrieval methods used an estimated model of the temperature-pressure (T-P) profiles and abundance ratios. Madhusudhan and Seager [38] presented a method of retrieving T-P profiles and abundance ratios by modelling millions of spectra to cover a significant portion of the parameter space and selecting the best fit. This method was later refined by Benneke and Seager [39] by constraining the parameter space through Bayesian analysis of the spectra, this allows for a more iterative approach of retrieving these state variables and has now become 'industry standard'. Identifying the absorbers involved requires identifying their spectroscopic signature in the observed spectrum. Generally absorbers such as water are easily identified due to their recognisable spectrum. However, emerging codes such as τ-Rex[3] utilise automated pattern recognition (see Figure 1.4) to identify absorbers in a more robust fashion.

The exact process of modelling requires looping over the frequencies of interest and calculating and integrating the contributions of all transitions within the region for all molecular species involved. This line-by-line calculation therefore requires knowledge of the spectroscopic characteristics of these transitions. This is commonly computed from spectroscopic data supplied by molecular line lists. Molecular line lists supply transitions between different energy states for a particular spectral range. Each transition consists of a frequency, lower state energy and Einstein-A coefficients from which the line strength of the observed line at an arbitrary temperature can be computed. More importantly, this spectroscopic data must have a high degree of completeness. The concept of 'completeness' of a line list lies in how well represented the spectra is for a certain temperature and frequency coverage. This aspect cannot be understated; the absorption spectrum of a molecule can vary significantly at different temperatures. Line lists that are incomplete will have missing features and/or experience a sudden drop in opacity at high temperatures. This can cause a mis-identification of atmospheric parameters in the observation. Previous atmospheric studies have been afflicted by this. An example is the inconclusive observation of methane in HD 189733b by Swain et al. [35] due to lack of spectroscopic data. Radiative modelling requires a robust identification

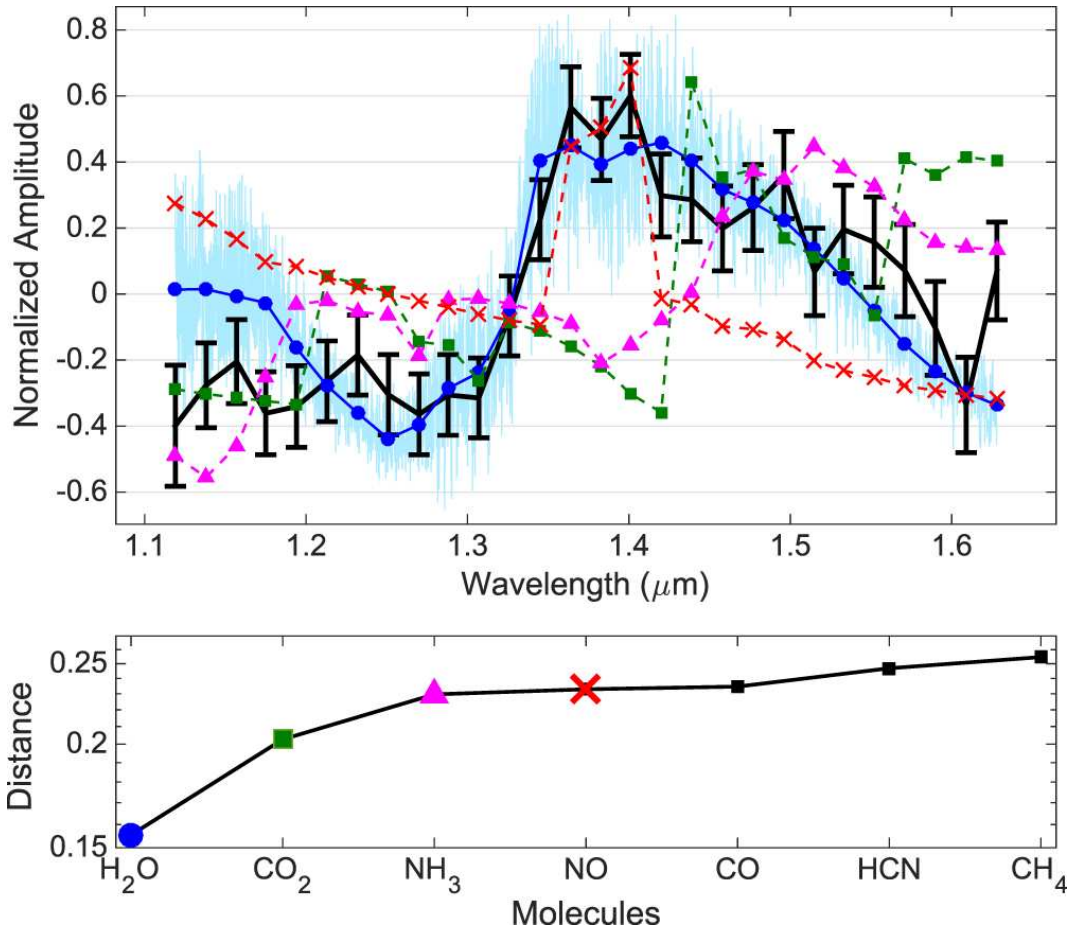

Figure 1.4: Taken from Figure 5 of Waldmann et al. [3]. Hubble/WFC3 transmission spectrum of HD 209458b [4] (black). Here the spectra of H_2O (blue dots), CO_2 (green squares), NH_3 (magenta triangles), and NO (red squares) are computed at the resolution of the observation. A high resolution spectrum of H_2O is also shown. The bottom plot describes the overall normalized distance between the observed data and the spectra of each absorber. Distance relates to how well a particular absorber's spectrum matches with the observation with lower values indicating a better fit. The conclusion is that H_2O represents the principal contributer and best fit to the observation.

of the main absorbers involved to match features in the observation. Modelling using incomplete spectroscopic data makes interpreting the spectra extremely difficult and inconclusive. High quality complete line lists are thus a necessity in the effectiveness of atmospheric retrieval codes.

Figure 1.5 describes the atmospheres of ten Hot Jupiters; highlighted here is the spectral coverage, observational sensitivity and frequency that is being attained by the current exoplanet community. As such, molecular line lists must satisfy a high degree of completeness in the near to far-infrared region for Hot Jupiters whose temperatures can reach up to 3,000 K. Newer ground based telescopes such as the European-Extremely Large Telescope (E-ELT) and space-based James Webb Space Telescope (JWST) will further increase both coverage and sensitivity of transit spectroscopy. The coming years will also expect to include definitive observations of currently tentative or undetected molecules including ammonia, hydrogen sulphide, phosphine, hydrogen cyanide [40], acetylene and ethylene. A huge burden is thus imposed on the availability and applicability of molecular line lists in order to make unambiguous interpretations of exoplanet atmospheres.

1.2 Molecular Spectroscopic Databases

Multiple databases exist that aggregate spectroscopic data of molecules. Examples of these include the HITRAN [41, 42, 6], CDMS [43], JPL [44] and PNNL-IR [10]. These databases provide experimental spectroscopic data on a range of molecular species for specific applications. The High Resolution Transmission (HITRAN) database originally concerned itself with molecules relevant to studying transmission spectra in Earth's atmosphere but has since seen numerous upgrades to cater to solar system atmospheres. It is one of the biggest databases available with over 4.2 million transitions across 47 molecular species and their isotopologues. The Cologne Database for Molecular Spectroscopy (CDMS) is a smaller database that expands on the aims of HITRAN by dealing with molecules that are relevant to cool interstellar enviroments and provides only rotational spectra. It has been partially integrated into later editions of the HITRAN database[6]. The Jet Propulsion

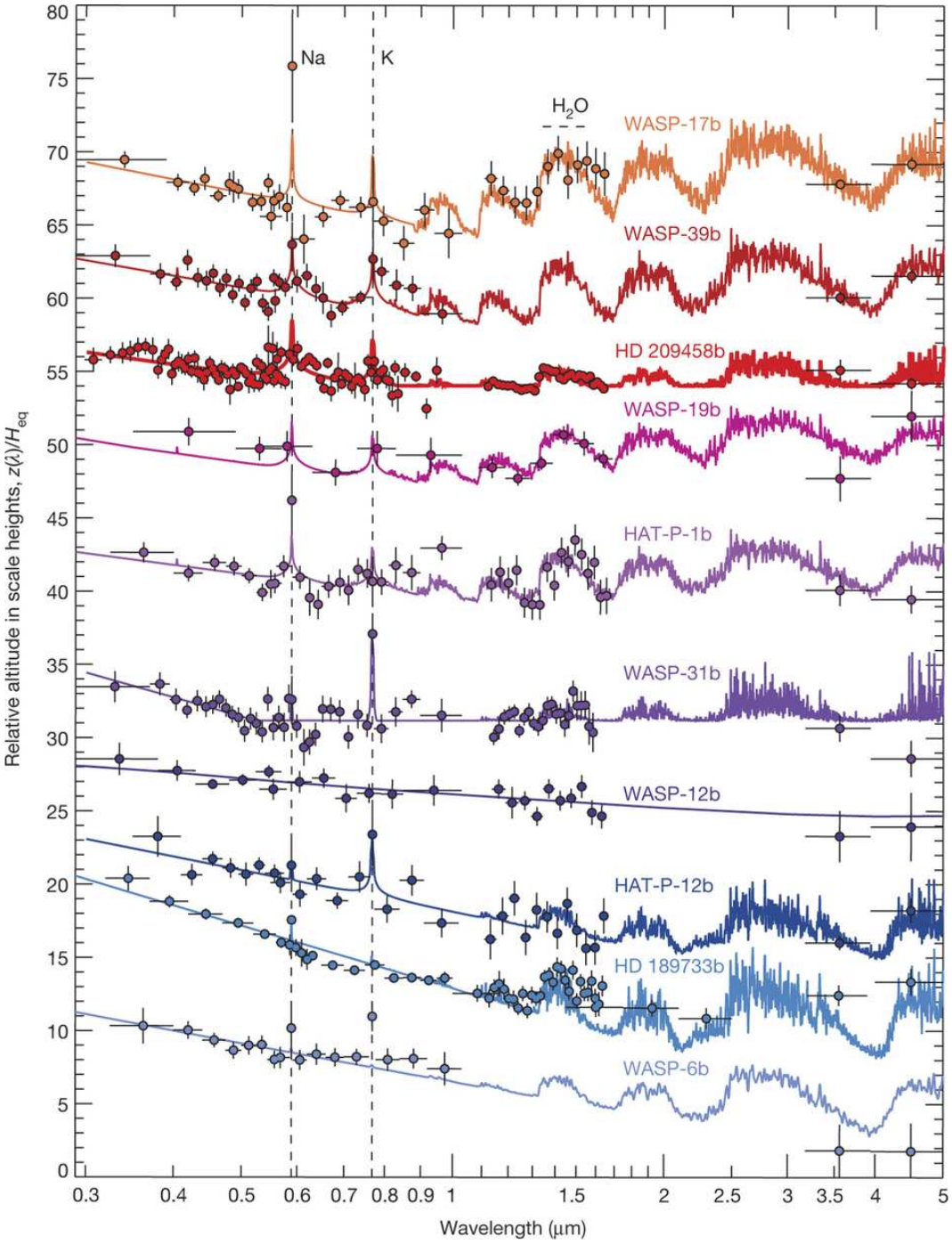

Figure 1.5: Taken from Figure 1 of Sing et al. [5] HST/Spitzer transmission spectra of 10 Hot-Jupiters. Solid circles indicate observed data with error bars and solid lines are modelled spectra.

Laboratory (JPL) spectroscopic database catalogs over 1.5 million spectral lines for atoms and molecules at wavelengths greater than 30 μm, the goal of which is to aid in observation and analysis of spectra of the interstellar medium and atmospheres. The Pacific Northwest National Laboratory IR (PNNL-IR) database provides measured room-temperature opacities of a range of molecules related to enviromental, energy and hazardous remote sensing. The CDMS, JPL and HITRAN databases are primarily composed of reliable high accuracy experimental or semi-empirical sources and provide parameters such as transition linestrengths or Einstein-A coefficients to simulate spectra at any temperature. As discussed at the beginning of the thesis, there are many obstacles to obtaining experimental data and they are often very difficult to overcome. Because of this, these databases suffer from a lack of spectral coverage and temperature applicability. Figure 1.6 highlights the issue of spectral coverage in the HITRAN database. The spectral range chosen coincides with that seen in Figure 1.5 and displays the common problem of gaps and unavailable data. Whilst there is some coverage in these regions by acetylene and hydrogen sulphide, phosphine is the most affected by this as there is no data below 2.7 μm. Overall this lack of data makes use of them unsuitable in radiative transfer modelling in these spectral ranges. The second issue presented refers to the temperature applicability of the data. The HITRAN database specifies its spectroscopic parameters for room temperature modelling only. Whilst it is possible to extrapolate these parameters to higher temperatures, doing so makes the data less complete. Temperature completeness arises from the Boltzmann distribution of energy states in a population. The population N of a state i with energy E_i is proportional to the Boltzmann distribution function:

$$N(i) \propto e^{-\frac{E_i}{kT}} \tag{1.1}$$

where k is the Boltzmann constant and T is the temperature in Kelvin. The transition intensity from a state Ψ^i to any other state is also proportional to $N(i)$. Therefore the opacity of a molecule becomes more evenly distributed across states making highly excited states stronger, lower energy states weaker and overall smoothing spectral

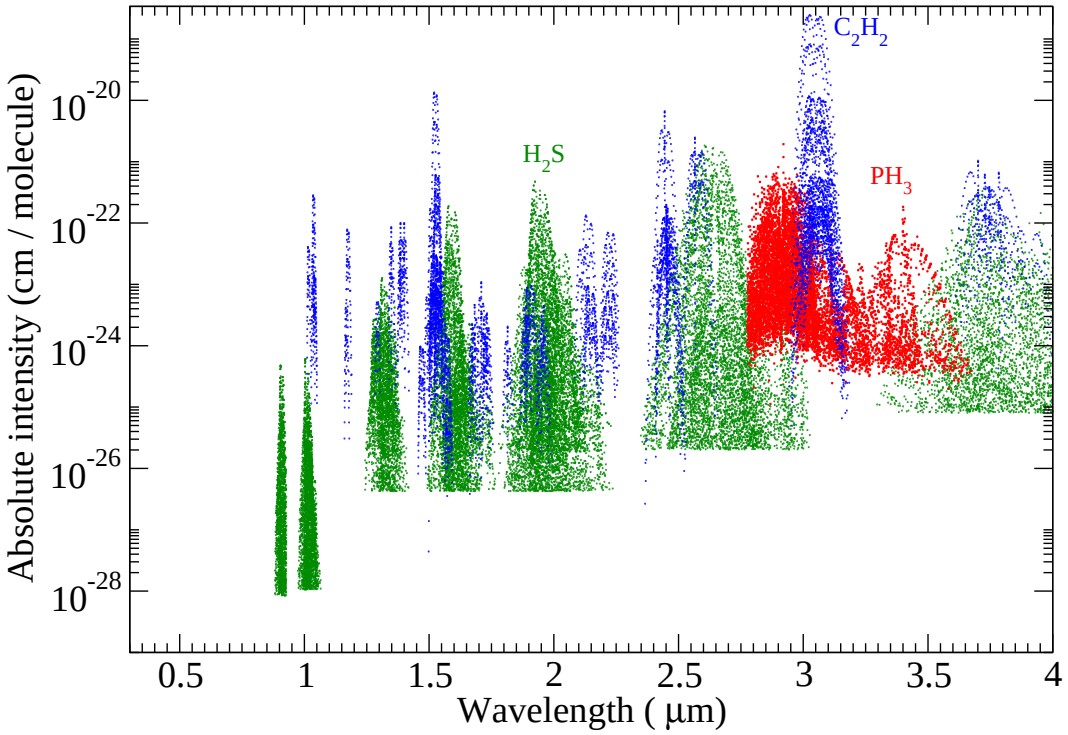

Figure 1.6: An overview of HITRAN data[6] in the far infrared region at room temperature (296 K). The wavelength range is the most common in astrophysical observation and the molecules chosen are acetylene (C_2H_2), hydrogen sulphide (H_2S) and phosphine (PH_3)

features. Therefore highly excited states may not be observable at room temperature as their transition intensity may lie below laboratory equipment sensitivity. Additionally, observations at elevated temperatures will produce more complicated spectra with billions of transitions that make them increasingly difficult to analyse and assign. As such, the room temperature data provided by these databases is usually inadequate for the high temperature modelling required for Hot Jupiters. The HITEMP database[45] is a high temperature version of HITRAN that provides appropriate data for modelling hot environments but is limited to only five molecular species. Theoretical methods therefore present a solution to this impasse.

1.3 Theoretical Line Lists

For polyatomic molecules, theoretical line lists present a way of providing the spectral range and temperature completeness that experimental studies struggle to achieve. Computed line lists can also provide other useful parameters such as

robust assignments, partition functions [46, 47], cooling functions and radiative lifetimes[12]. The usefulness of a high quality line list cannot be understated: H_3^+ is one of the most common cations in the universe yet it is almost impossible to measure its transition intensities experimentally; this has made it one of the most well studied molecules through theoretical means[48, 49, 50] and such methods were paramount in its detection in supernovae[51]. A calculated H_3^+ line list by Neale et al. [52] has allowed for the improved modelling of white dwarfs [53] and storage-ring experiments[54]. In addition, its usage in the hydrodynamic models of Hot Jupiter atmospheres[55] have indicated that H_3^+ is important to the atmosphere's thermal stability as its IR emissions act as a significant cooling mechanism. The benefit of theoretical line lists also extends to molecules that are experimentally well-studied, in particular H_2O. Whilst there are only around 80,000 experimentally known transition intensities [6] for H_2O, the computed hot line list BT2 by Barber et al. [29] extends this coverage to over half a billion allowing for a better spectral range and temperature dependance in its opacities. These features provided by BT2 has attracted extensive usage in a wide range of fields. BT2 was utilized in the atmospheric retrieval of HD 189733b which resulted in the first detection of water in an exoplanetary atmosphere[34]. It has improved the modelling of brown dwarfs and very cool stars[56]. Atmospheric retrieval using BT2 allowed for the the detection of the very cool brown dwarf ULAS J003402.77[57]. Interestingly, Warren et al. [57] discuss how the lack of hot line list for NH_3 and CH_4 at the time made it extremely difficult to acquire adequate opacities compared to H_2O. Interstellar, cometary[58] and protoplanetary[59] modelling was also improved through usage of BT2 and allowed for detection of water in cometary spectra of 8P/Tuttle[60] and of its isotopologues[61] in astrophysical media. Outside of astronomical settings, BT2 is also used in benchmarking the resolution of new spectroscopic methods[62] and was also used in analysing the refractive index of humid air[63]. This list is not exhaustive and many more uses of the BT2 can be found in various fields of literature.

There are two distinct theoretical methodologies used to construct these line

lists: Perturbation and variational theory. Perturbation theory has formed the foundation of high resolution spectroscopy. The majority of experimental spectra are analysed and assigned by these methods using *effective Hamiltonians* and spectroscopic constants. Spectroscopic Hamiltonians exclusively utilize experimental observations in order to construct synthetic spectra and are useful in characterising the nature of rovibrational transitions. Such spectra boast a high degree of agreement with the line positions against experiment. However they are entirely reliant on both the quality and quantity of experimental data and generally fail to predict or replicate weak transitions and unobserved regions in the spectrum. In short, a high temperature synthetic spectrum requires high temperature experimental data. Additionally, there are many systems and problems were perturbative treatments break down; The standard assumption with perturbation is that the molecule will undergo small amplitude motion, and this is poorly represented when the molecule contains hydrogen or approaches dissociation. Variational methods address these limitations as it attempts towards a more complete solution. They are seen as a more robust method of replicating spectroscopic data as they are based on using *ab initio* ("First principles") theories and methodologies. The main disadvantage is the computational and memory cost required for high accuracy that rapidly scales with increasing complexity of the molecule. With the advent of increased computing power, robust codes and new methodologies, they are now not only a viable option, but, for some systems, can even compete with experimental data in terms of accuracy.

1.4 The ExoMol Project

The ExoMol project [8] aims to build upon these previous endeavours by providing a comprehensive database of molecular line lists for usage in characterising and modelling astrophysical phenomena, with particular interest in exoplanets and cool stars. The goal is not only to improve on currently available line lists for high tempreature modelling but to also provide line lists for molecules that are expected to be of interest in the near future. The ExoMol project has already produced hot line

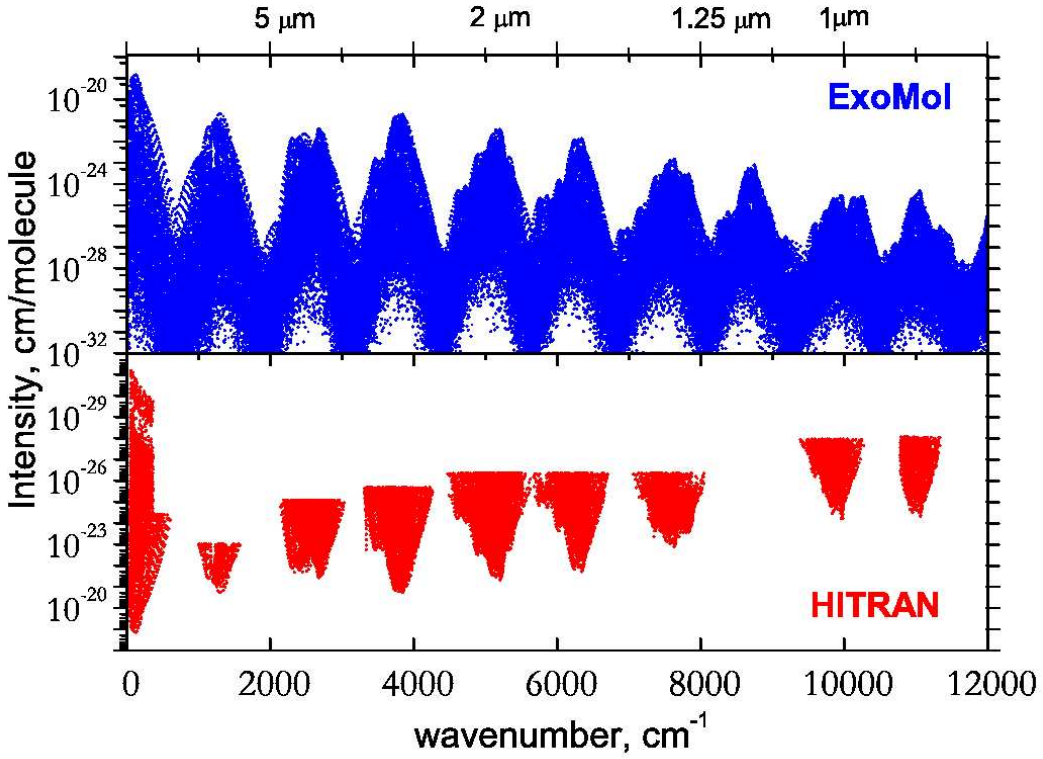

Figure 1.7: Taken from Figure 2 of Azzam et al. [7]. A 296 K simulated spectrum from the AYT2 line list for hot H_2S compared against HITRAN.

lists for important molecules including PH_3 [9] (Chapter 3.8.1), SO_2 [21] and CH_4 [64] which are being used extensively. The '10to10' hot CH_4 line list [64] in particular has seen use in analysis of late-T dwarf spectra [65], atmospheric modelling of the super-Earth 55 Cancri e[40] and the mass photometry of the HD8799 system [66] . Such line lists can contain tens of billions of transitions necessary for hot temperature modelling. Figure 1.7 demonstrates the high coverage and completeness a typical ExoMol line list provides. For this reason, the τ-Rex[3] code has heavily adopted ExoMol line lists. Figure 1.8 describes the production outline for a line list in the ExoMol project as well as the problems that are required to be solved. The main codes used across the line list production pipeline are:

- *Solve the electronic structure problem*: This is done via standard quantum chemistry calculation codes such as MOLPRO [67] and CFOUR [68].

- *Solve the nuclear motion problem*: Here, the ExoMol project utilizes the LEVEL [69] and Duo code [70] for diatomics, DVR3D [71] for triatomics

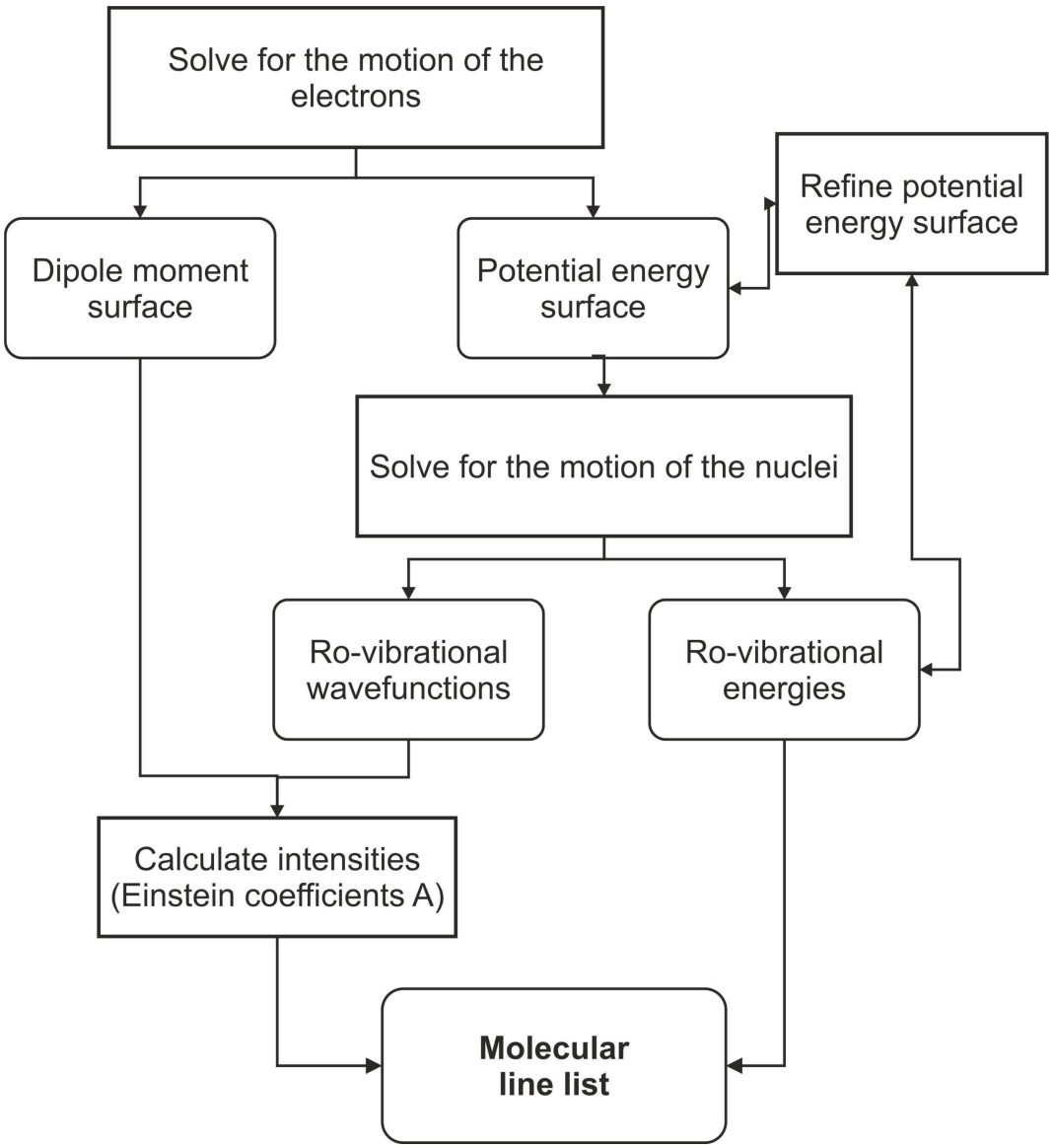

Figure 1.8: Taken from Figure 3 of Tennyson and Yurchenko [8]. The general production outline of a molecular line list in the ExoMol project.

and TROVE [30] for polyatomics to calculate energies and wavefunctions.

- *Compute Einstein-A coefficients*: All codes for the nuclear motion problem compute Einstein-A coefficients as standard but an additional new code GAIN [72] (Chapter 3) can also produce these coefficients from TROVE wavefunctions.

Experimental accuracy in most cases is not possible using purely *ab initio* methods. The ExoMol pipeline therefore allows for the introduction of experimental data through refinement (see Chapter 2). Even with state-of-the-art codes and the usage of multiple high performance computing facilities, the task of computing a line list is still computationally expensive and time consuming. Each ExoMol team member is therefore usually assigned a single molecule to complete but with the improvement of methods and codes within the group, it is becoming increasingly common to work with multiple species. This work in this thesis presents two contributions to the ExoMol project. The first is the production and publication of a hot molecular line list for the formaldehyde molecule and a room temperature and hot line list for hydrogen peroxide. The second is the contribution of a new method and code for the rapid calculation of billions of line strengths. The aim for this code is to help the ExoMol project to continue to push towards bigger, hotter and more difficult molecular systems.

Chapter 2

Solving the Nuclear Motion Problem

Consider a molecule with P particles, of which N are nuclei and $N_e = P - N$ are electrons. We will not consider the effect of spin. The exact non-relativistic molecular translational(t)-rovibrational(rv)-electronic(e) Hamiltonian in a Cartesian co-ordinate system (X,Y,Z) is of the form:

$$H_{trve} = T_e + T_N + V(R_N, r_e) \tag{2.1}$$

where:

$$T_e = -\frac{\hbar^2}{2m_e}\sum_{i=N+1}^{P}\nabla_i^2 \qquad \text{total electron kinetic energy (KE)} \tag{2.2}$$

$$T_N = -\frac{\hbar^2}{2}\sum_{i=1}^{N}\frac{\nabla_i^2}{m_i} \qquad \text{total nuclear KE} \tag{2.3}$$

$$V(R_N, r_e) = \sum_{r<s=1}^{P}\frac{C_r C_s e^2}{4\pi\varepsilon_0 R_{rs}} \qquad \text{Electrostatic potential energy} \tag{2.4}$$

with $R_N = (X_1, Y_1, Z_1, X_2, Y_2, Z_2, ..., X_N, Y_N, Z_N)$ as the nuclear co-ordinates, $R_e = (X_{N+1}, Y_{N+1}, Z_{N+1}, X_{N+2}, Y_{N+2}, Z_{N+2}, ..., X_P, Y_P, Z_P)$ as the electronic co-ordinates and the gradient and Laplace operators defined as

$$\nabla_i = \frac{\partial}{\partial X_i} + \frac{\partial}{\partial Y_i} + \frac{\partial}{\partial Z_i} \quad ; \nabla_i^2 = \nabla_i \cdot \nabla_i \tag{2.5}$$

Additionally m_i is the mass of each nucleus, m_e is the electron mass, M_N is the total nuclear mass, $C_r e$ is the charge of a particle and R_{rs} is the separation between

particles r and s. The full derivation of these equations is given in Bunker and Jensen [20]. Therefore one must solve:

$$H_{trve}\Psi_{trve,n} = E_{trve,n}\Psi_{trve,n} \tag{2.6}$$

Unfortunately Eq. (2.6) is practically impossible to solve for even the smallest of molecules. Therefore a physically appropriate approximation must be made. The most important of which is the Born-Oppenheimer Approximation.

2.1 The Born-Oppenheimer Approximation

The basis of this approximation relies on the idea that the motion of the nuclei is significantly slower compared to the fast motion of the electrons. Therefore the kinetic energy contribution of the nuclei to the motion of the electron is minimal. Taking the variational approach, we can therefore separate the rovibronic wavefunction as such:

$$\Psi_{rve,m} = \sum_{n'} \Phi^{m}_{trv,n'}(R_N)\Phi_{e,n'}(R_N, r_e) \tag{2.7}$$

Putting this into Eq. (2.6) and integrating with $\Phi_{e,n}(R_N, r_e)$ over all r_e we get:

$$[T_N + E_e(R_N)]\Phi^{m}_{trv,n} + \sum_{n'} \langle \Phi_{e,n}|T_N|\Phi_{e,n'}\rangle \Phi^{m}_{trv,n'} = E_{trve,m}\Phi^{m}_{trv,n} \tag{2.8}$$

the second term is the *vibronic coupling* which the Born-Oppenheimer approximation neglects, we arrive at the ro-vibrational form:

$$[T_N + E_{e,n}(R_N)]\Phi^{m}_{trv,n} = E_{trve,m}\Phi^{m}_{trv,n} \tag{2.9}$$

Where solving the electronic wave-equation for $E_{e,n}(R_N)$:

$$[T_e + V(R_N, r_e)]\Phi_{e,n}(R_N, r_e) = E_{e,n}(R_N)\Phi_{e,n}(R_N, r_e) \tag{2.10}$$

provides the potential function for Eq. (2.9). As there is no nuclei mass term in Eq.(2.10), the potential function is isotopically invariant. Solving Eq. (2.10) re-

quires fixing R_N for each solution effectively 'clamping' the nuclei. For polyatomic molecules the potential function has the form of an N-dimensional hypersurface and is therefore referred to as a *Potential Energy Surface* (PES). The polyatomics discussed in this thesis only deal with the ground electronic state ($n = 0$) as higher excited electronic states only occur at frequencies significantly greater than infrared. With this we can simplify Eq. (2.7) as:

$$\Psi_{trve,m} = \Phi^m_{trv}(R_N)\Phi_e(R_N, r_e) \tag{2.11}$$

and modify the form of ro-vibrational equation from Eq. (2.9) to:

$$[T_N + V_B(R_N)]\Phi_{trv} = E_{trv}\Phi_{trv} \tag{2.12}$$

Where $V_B(R_N) = E_{e,0}(R_N)$.

2.2 Potential Energy Surface

As discussed previously, the solution of the clamped nuclear Hamiltonian given by Eq. (2.10) takes the form of a hypersurface referred to as a PES. Despite the Born-Oppenheimer approximation, the clamped nuclear or 'electronic structure' problem is still analytically impossible to solve for most molecular systems and must instead rely on numerical means. However, solving through numerical integration methods is extremely difficult as the problem and computational cost scales exponentially with number of particles[73]. Therefore the N-body problem must be treated with specially developed methods that introduce significant approximations. These methods form the foundation of the well established field of theoretical quantum chemistry.

2.2.1 Hartree-Fock Method

The simpliest model in *ab initio* electronic structure theory is the Hartree-Fock (HF) mean field theory[74]. It makes the assumption that the electron is interacting with a mean field potential caused by other electrons in the system and further postulates that the N electron wavefunction can be approximated using a single Slater

Determinant:

$$\Phi_e = \frac{1}{\sqrt{N_e!}} \begin{vmatrix} \chi_1(r_1) & \chi_2(r_1) & \cdots & \chi_{N_e}(r_1) \\ \chi_1(r_2) & \chi_2(r_2) & \cdots & \chi_{N_e}(r_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(r_{N_e}) & \chi_2(r_{N_e}) & \cdots & \chi_{N_e}(r_{N_e}) \end{vmatrix} \tag{2.13}$$

where χ_i are the one electron functions or 'orbitals' as a function of the position r_i of electron i. This determinant has the property of antisymmetry with respect to the interchange of two identical electrons, satisfying the Pauli principle. The variational principle can be used to approximate the wavefunction on the principle that the energy of a trial wavefunction E_v and the exact energy E are related by:

$$E \leq E_v \tag{2.14}$$

Therefore the electron orbitals χ of the trial wavefunction can optimized in order to minimize its energy. An iterative approach is utilized whereby the optimized wavefunctions are reused as the trial wavefunction and then the optimization is repeated until convergence is achieved. This is known as the *self-consistent* method[75, 76]. The choice of χ orbitals is based on chosing a model function or *basis set* that best represents the electrons distribution in the system. Larger basis sets approximate orbitals more accurately as they can represent more nuances in the electron distribution at the cost of computational requirements. The need for a large number of basis sets can be offset with more sophisticated functions that represent aspects such as polarization (6-31G basis sets), diffuse (6-311++G) and splitting (G3Large) with decreasing integration efficiency. Therefore, the choice of basis set is always a trade-off between accuracy and computational cost. Energies that are calculated with the most ideal basis set (infinite or exact orbital representation) are known as the *Hartree-Fock limit* energies. However, the approximations used by the HF method means that even the Hartree-Fock limit energy will always be higher than the exact energy. This difference is formulated as:

$$\varepsilon = E - E_{HF} \tag{2.15}$$

Where E_{HF} is the Hartree-Fock energy and ε is known as the *electron correlation* energy. This residual correlation term comes from the single Slater determinant that assumes the electronic motion is independant of every other electron. In absolute terms, the HF method recovers $\approx$99% of the exact total energy[77] near equilibrium which is impressive for a simple wavefunction. However, this accuracy is still insufficient as important chemical properties such as bond angles and lengths are extremely sensitive to the 'lost' energy from electron correlation[77]. Further away from equilibrium, the electron correlation term grows and HF breaks down further making it unsuitable for usage in accurate PES calculations[77]. Therefore, *Post-Hartree-Fock* methods have been developed to in order to retrieve this residual energy.

2.2.2 Coupled Cluster

The development of Post-Hartree-Fock methods is well established and has brought about many new methodologies that aim to improve the accuracy of HF methods. Discussions and comparisons between methods are common in the quantum chemistry literature[73, 78, 79, 77] with the most well known Post-Hartree-Fock methods including configuration interaction (CI), Møller-Plesset perturbation theory (MP), Density Functional Theory (DFT) and Coupled Cluster (CC) [80, 81]. For brevity, only CC will be discussed as it is the theoretical basis of all PES calculations in this thesis (Chapters 4 and 5). The form of the CC wavefunction is:

$$\Psi = e^{T}\Psi_0 \tag{2.16}$$

Where Ψ is the exact wavefunction, Ψ_0 is a reference wavefunction (usually a HF wavefunction) and T is the coupled cluster operator which when acted on the reference wavefunction, produces a linear combination of excited determinants. The form of T is written as:

$$T = T_1 + T_2 + T_3 + \cdots \tag{2.17}$$

where T_i represents the operator for excitations with $i = 1$ for single, $i = 2$ for double, $i = 3$ for triple and so on. Higher excitations incur a greater computational

demand and are therefore truncated. This gives rise to a family of CC methods with increasing accuracy and computation expense and are named on the truncation imposed: CCS is truncated up to T_1, CCSD up to T_2, CCSDT up to T_3 etc. Computational effort can be saved on higher excitations by introducing approximate treatments and is reflected in the naming convention by surrounding parenthesis. The one most commonly used in this thesis is CCSD(T) where the single and double excitations are fully treated and the triple excitations utilize a perturbative approximation. CCSD(T) presents the best balance in computational effort and accuracy[77] and is sometimes considered the 'gold standard in quantum chemistry' [73].

The calculation of the PES provides the potential term in Eq. (2.12) and allows us to begin solving the rovibrational Hamiltonian.

2.3 TROVE

Solving Eq. (2.12) in its current form does not take into account that the translation can be separated from the internal motion (rotational and vibrational) of the molecule. A commonly used system is the molecule-fixed axis system with origin at the nuclear center of mass. This can be described as the axis following and rotating with the molecule such that the molecule can be viewed as 'stationary'. The internal vibrational motion therefore only depends on the displacement of the nuclei via some internal coordinate system. Therefore a transformation from $R_N = (X_1, Y_1, Z_1, ..., X_N, Y_N, Z_N)$ co-ordinates to new $3N$ co-ordinates $\mathfrak{R} = (X^0, Y^0, Z^0, \theta, \phi, \chi, \xi_1, \xi_2, ..., \xi_{3N-6})$ is required where (X^0, Y^0, Z^0) is the nuclear center of mass coordinate given by:

$$X = \frac{\sum_{i=1}^{N} m_i X_i}{\sum_{j=1}^{N} m_j} \quad Y = \frac{\sum_{i=1}^{N} m_i Y_i}{\sum_{j=1}^{N} m_j} \quad Z = \frac{\sum_{i=1}^{N} m_i Z_i}{\sum_{j=1}^{N} m_j} \tag{2.18}$$

(θ, ϕ, χ) are the Euler angles that define the orientation of the molecule-fixed axis in relation to the laboratory-fixed system and $\xi = (\xi_1, \xi_2, ..., \xi_{3N-6})$ are the internal co-ordinates. The momentum operators also become: $\wp = (P_X, P_Y, P_Z, J_x, J_y, J_z, p_1, p_2, ..., p_{3N-6})$ where (P_X, P_Y, P_Z) are the momeda relating to the translation (X, Y, Z), (J_x, J_y, J_z) are the total rotational angular momenta op-

erators and $(p_1, p_2, ..., p_{3N-6})$ are the momenta relating to the internal co-ordinates. This co-ordinate change to the kinetic energy operator can be facilitated by using the Podolsky trick (See Chapter 7 of Bunker and Jensen [20]) which gives the general form for the Hamiltonian:

$$H_{trv} = \frac{1}{2}\sum_{\lambda,\mu}^{3N} \wp_\lambda G_{\lambda\mu}(\xi)\wp_\mu + U(\xi) + V(\xi) \tag{2.19}$$

Explicity this is written as:

$$H_{trv} = \frac{1}{2}\sum_{F=X,Y,Z} P_F G_{FF} P_F \tag{2.20}$$

$$+\frac{1}{2}\sum_{\alpha=x,y,z}\sum_{\alpha'=x,y,z} J_\alpha G_{\alpha\alpha'}(\xi) J_{\alpha'} \tag{2.21}$$

$$-\frac{i\hbar}{2}\sum_{\alpha=x,y,z}\sum_{n=1}^{3N-6}\left[J_\alpha G_{\alpha,n}(\xi)\frac{\partial}{\partial\xi_n} + \frac{\partial}{\partial\xi_n}G_{\alpha,n}(\xi)J_\alpha\right] \tag{2.22}$$

$$-\frac{\hbar^2}{2}\sum_{n=1}^{3N-6}\sum_{n'=1}^{3N-6}\frac{\partial}{\partial\xi_n}G_{nn'}(\xi)\frac{\partial}{\partial\xi_{n'}} + U^(\xi) + V^(\xi) \tag{2.23}$$

where $U(\xi)$ is the psuedo-potential and G is the kinetic energy co-efficient matrix that is block diagonal with 3 blocks. Two blocks have size 3×3 and relate to the translation and rotational parts which are completely diagonal, a $(3N-6)\times 3$ block which represents the coriolis coupling of the vibrational motion with rotation and a $(3N-6)\times(3N-6)$ block that represents the vibrational terms and coupling. We can simplify the operator to:

$$H_{trv} = H_{tran} + H_{ro} + H_{cor} + H_{vib} \tag{2.24}$$

where H_{tran} is the translational part, H_{ro} is the rotational, H_{cor} is the coriolis coupling part and H_{vib} is the vibrational part of the Hamiltonian which includes the potential V and psuedo-potential U. A problem appears in the definition of the momenta conjugate relating to the ξ. Here they each require explicit definition for molecules that differ by number of nuclei and/or structure due to the different internal co-ordinates to describe the vibrational motion. This can be overcome by de-

riving the kinetic energy numerically which is the procedure performed by the code **T**heoretical **Rov**ibrational **E**nergies (TROVE)[30]. TROVE is a variational nuclear motion solver that can be used for general polyatomic molecules with arbitrary structure. It works by deriving an approximate kinetic energy operator (AKEO) in terms of a Taylor expansion of the internal co-ordinates. To do this we must be able to express $G_{\lambda\mu}(\xi)$ and $U(\xi)$ as a series expansions:

$$G_{\lambda\mu} = \sum_{l_1,l_2,\ldots} G^{\lambda\mu}_{l_1,l_2,\ldots} g_1^{l_1} g_2^{l_2} \ldots \tag{2.25}$$

$$U = \sum\nolimits_{l_1,l_2,\ldots} U_{l_1,l_2,\ldots} g_1^{l_1} g_2^{l_2} \ldots \tag{2.26}$$

where $G^{ij}_{l_1,l_2,\ldots}$ and $U_{l_1,l_2,\ldots}$ are expansion co-efficients and $g_1, g_2, \ldots$ are functions of our choosing that depend on a single internal co-ordinate ($g_n = g_n(\xi_n)$). The approach follows the procedure given by Sørensen [82] where we start by expressing the Cartesian kinetic energy operator as:

$$T = \sum_{i=1}^{N} \frac{\mathbf{P_i}^2}{2m_i} \tag{2.27}$$

where P_i is a 3-dimensional momentum vector with $\mathbf{P_i} = (P_{iX},,P_{iY},P_{iZ})$ and:

$$P_{iR} = -i\hbar \frac{\partial}{\partial R_i} \tag{2.28}$$

for nuclei i and co-ordinate $R = X,Y,Z$. To transform $\wp \rightarrow \mathbf{P}$ requires applying the chain-rule transformation of $\mathfrak{R} \rightarrow R_N$:

$$\mathbf{P_i} = \frac{1}{2} \sum_{\lambda=1}^{3N} \sum_{R=X,Y,Z} (s_{\lambda,iR} \wp_\lambda + \wp_\lambda s_{\lambda,iR}) \tag{2.29}$$

where:

$$s_{\lambda,iR} = \frac{\partial \mathfrak{R}_\lambda}{\partial R_i} / \tag{2.30}$$

is the Jacobian matrix of the chain-rule transformation. Therefore inserting Eq. (2.29) into Eq. (2.27) and comparing with Eq. (2.19) we can define $G_{\lambda,\mu}$ as:

$$G_{\lambda\mu} = \sum_{i}^{N} \frac{1}{m_i} \sum_{R=X,Y,Z} s_{\lambda,iR} s_{\mu,iR} \tag{2.31}$$

and U as:

$$\sum_{i}^{N} \frac{1}{m_i} \sum_{\lambda}^{3N} \sum_{\mu}^{2N} \sum_{R=X,Y,Z} s_{\lambda,iR} \wp_\lambda s_{\mu,iR} \wp_\mu + s_{\lambda,iR} \wp_\lambda \wp_\mu s_{\mu,iR} + \wp_\lambda s_{\lambda,iR} \wp_\mu s_{\mu,iR} \tag{2.32}$$

The inverse Jacobian matrix $t_{iR,\lambda} = \frac{\partial R_i}{\partial \Re_\lambda}$ yields the relation:

$$\sum_{i}^{N} \sum_{R=X,Y,Z} s_{\lambda,iR} t_{iR,\mu} = \delta_{\lambda,\mu} \tag{2.33}$$

with explicit expressions [82] for t given by:

$$t_{iR,\alpha} = \begin{cases} \delta_{R,\alpha} & \text{if } \alpha = (X,Y,Z) \text{ (translation)} \\ \sum_{\beta=X,Y,Z} \varepsilon_{R\alpha\beta} \beta_i & \text{if } \alpha = (X,Y,Z) \text{ (rotation)} \\ \frac{\partial R_i}{\partial \xi_\alpha} & \text{if } \alpha = (1,2..,3N-6) \text{ (vibration)} \end{cases} \tag{2.34}$$

2.3.1 Approximate Kinetic Energy Operator

Eq. (2.33) is a linear system of $3N \times 3N$ equations. These can be approximated by expanding s and t as the power series:

$$\begin{aligned} s_{\lambda,iR}(\xi) &= \sum_{\mathbf{L}} s^{\lambda,iR}_{l_1,l_2,\ldots} g_1^{l_1} g_2^{l_2} \cdots \\ t_{iR,\mu}(\xi) &= \sum_{\mathbf{L}} t^{iR,\mu}_{l_1,l_2,\ldots} g_1^{l_1} g_2^{l_2} \cdots \end{aligned} \tag{2.35}$$

Where $\mathbf{L}$ represents an N-dimensional vector space defined as $\mathbf{L} \in \{\mathbb{N}^N || l_1 + l_2 + \cdots + l_N| \leq L\}$ where L is the maximal expansion order. Inserting Eq. (2.35) into

Eq. (2.33) gives the recursive expression:

$$\sum_{i}^{N}\sum_{R=X,Y,Z} s_{\mathbf{L}}^{\lambda,iR} t_{\mathbf{0}}^{iR,\mu} = \delta_{\lambda,\mu}\delta_{\mathbf{L0}} - \sum_{i}^{N}\sum_{R=X,Y,Z}\sum_{\mathbf{K}} s_{\mathbf{K}}^{\lambda,iR} t_{\mathbf{L}-\mathbf{K}}^{iR,\mu} \tag{2.36}$$

where $\mathbf{K}$ is a subspace of $\mathbf{L}$ with the condition that $|\mathbf{K}| \leq |\mathbf{L}|$, $\mathbf{0}$ is the zero-set subspace of $\mathbf{L}$. From this we see that each $s_{\mathbf{L}}^{\lambda,iR}$ gives a system of linear equations that depends on coefficients of a lower order. Hence it is only required that $t_{\mathbf{0}}^{iR,\mu}$ be defined from Eq. (2.34) with respect to a co-ordinate system of our choosing and its derivatives in order to begin the procedure. Inserting Eq. (2.35) into Eq. (2.31) gives us the expression:

$$G_{\lambda\mu} = \sum_{\mathbf{L}}\sum_{\mathbf{K}}\sum_{i}^{N}\frac{1}{m_i}\sum_{R=X,Y,Z} s_{\mathbf{L}}^{\lambda,iR} s_{\mathbf{L}-\mathbf{K}}^{\mu,iR} g^{\mathbf{L}} \tag{2.37}$$

where:

$$g^{\mathbf{L}} = \prod_{i}^{3N-6} g_i^{\mathbf{l_i}} \tag{2.38}$$

is a re-expression of the internal co-ordinate function product series in compact form. The kinetic energy expression is in the series expansion form presented in Eq. (2.25). Therefore an expression for the expansion co-efficient is:

$$G_{\mathbf{L}}^{\lambda\mu} = \sum_{\mathbf{L}}\sum_{\mathbf{K}}\sum_{i}^{N}\frac{1}{m_i}\sum_{R=X,Y,Z} s_{\mathbf{L}}^{\lambda,iR} s_{\mathbf{L}-\mathbf{K}}^{\mu,iR} \tag{2.39}$$

The same methodology can be applied to the psuedo-potential and potential to arrive at a similar expressions. This approach is general and has allowed TROVE to handle a wide range of molecules and more complicated coordinate systems such as curvilinear coordinates[83]. This method has a few drawbacks, first is the numerical error that can accumulate from the finite difference method used in computing derivatives of higher order. This can be solved through using quadruple precision or automatic differentiation[83]. Secondly, an exact kinetic energy operator (EKEO) allows for the coupling of rotational and vibrational motion. This coupling allows molecules to smoothly transition from bent to linear geometries which is essential

for molecules such as C_2H_2 and floppy molecules[84]. In an AKEO, molecules that exhibit linear geometries will have coordinates (such as the torsional) that become undefined, the consequence of this is that the t matrix becomes singular[85] and the s matrix is unobtainable by inversion. Codes such as WAVR4[86] utilize an EKEO and are able to resolve such motion at the cost of significantly longer computation time. However in TROVE, this limitation with the AKEO has recently been overcome for the C_2H_2 molecule[87].

2.3.2 Symmetry

Before we move on to solving the ro-vibrational Hamiltonian we must briefly discuss symmetry. Molecules that contain a number of identical species of atoms can be classified under a Molecular Symmetry Group (MS)[20]. A symmetry group consists of a set of symmetry operations such as rotational, reflection and inversion that leave the energy of a system of particles unchanged, or to put it mathematically:

$$[R, H_{rovib}] = 0 \tag{2.40}$$

The symmetry operation R commutes with the Hamiltonian. Symmetry plays an important role in understanding the dynamics and spectra of molecules. One concept involves labelling molecules based on their symmetry or their *irreducible representation*:

$$H_{rovib} R\psi = ER\psi = E\psi^R = Ec^R\psi \tag{2.41}$$

The operation R generates a new wavefunction ψ^R with the same energy as ψ. This can be represented as a coefficient or 'character' c^R to the original wavefunction. A *character table* can be constructed and a symmetry label Γ can be assigned to the wavefunction and energy. Additionally the product rule states:

$$\psi = \phi_i\phi_j\phi_k \tag{2.42}$$

$$\Gamma(\psi) = \Gamma(\phi_i) \otimes \Gamma(\phi_j) \otimes \Gamma(\phi_k) \tag{2.43}$$

The major molecules dealt in this thesis involve Abelian symmetry groups that generally have character tables consisting solely of values ± 1. They also have the property of commutivity:

$$\Gamma(\phi_i) \otimes \Gamma(\phi_j) = \Gamma(\phi_j) \otimes \Gamma(\phi_i) \tag{2.44}$$

One of the most important and most powerful features of symmetry involves the vanishing integral rule[88]:

$$\langle \psi' | \hat{O} | \psi'' \rangle = 0 \quad \text{If} \quad \Gamma(\psi') \otimes \Gamma(\hat{O}) \otimes \Gamma(\psi'') \not\supset \Gamma^{(s)} \tag{2.45}$$

Here $\Gamma(\psi')$, $\Gamma(\psi'')$ and $\Gamma(\hat{O})$ are the symmetries of the wavefunctions and operator $\hat{O}$ respectively and $\Gamma^{(s)}$ refers to a totally symmetric symmetry species. This rule dictates that if the product of all symmetries does not contain a totally symmetric species then the integral will be zero. Therefore, with prior knowledge of symmetries, we can avoid unnecessary computation. With this we can now begin computing rovibrational energies.

2.3.3 Formulating the Ro-vibrational Hamiltonian

Since H_{tran} is separable mathematically from the rotation and vibrational parts in Eq. (2.20), we will not consider it and only deal with the ro-vibrational terms. Since TROVE is variational, the form of wavefunction for a particular state i is:

$$\Psi_i = \sum_{v,J,K,\tau} c^i_{v,K,\tau} |v\rangle |J,K,\tau\rangle \tag{2.46}$$

where $|J,K,\tau\rangle$ is the rotational basis-set, $|v\rangle$ is the vibrational basis-set and $c^i_{v,K,\tau}$ are the variational coefficients to be solved for.

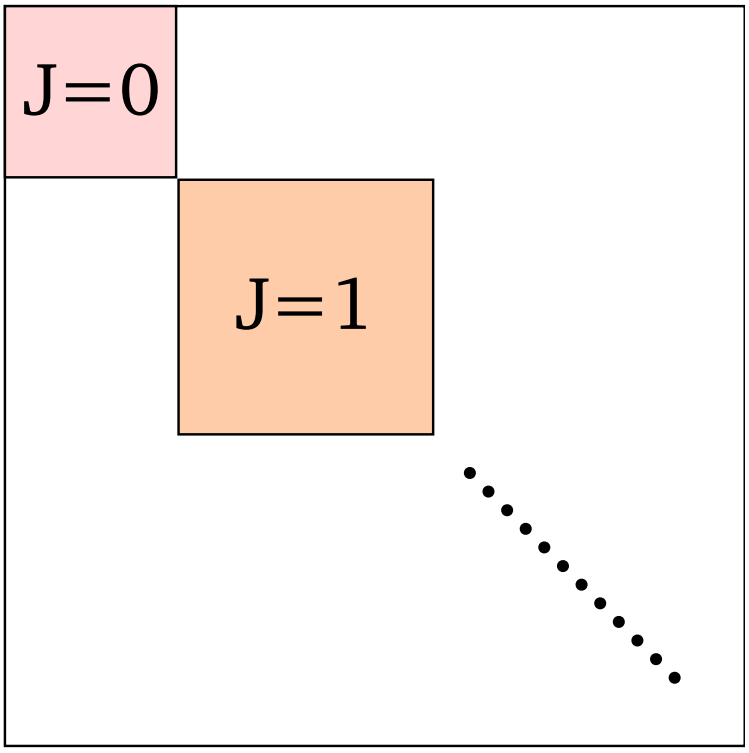

Figure 2.1: Block matrix structure of the Hamiltonian for different J

2.3.3.1 Rotational basis-set

The rotational basis-set used in TROVE is defined as:

$$|J,K,\tau\rangle = \begin{cases} |J,0\rangle & K=0 \\ \frac{1}{\sqrt{2}}(|J,K\rangle + (-1)^{J+K}|J,-K\rangle) & K>0\ \tau=0 \\ \frac{(-1)^{\Sigma}}{\sqrt{2}}(i|J,K\rangle - (-1)^{J+K}i|J,-K\rangle) & K>0\ \tau=1 \end{cases} \tag{2.47}$$

where J is the total rotational angular momentum, K is the projection of the rotational angular momentum on the body-fixed axis z, $|J,K\rangle$ are the rigid-rotor functions[20] and τ can be considered the rotational parity [30, 89]. The rotational symmetry properties of J means that Eq. (2.45) is zero if $J' \neq J''$. This gives rise to a block diagonal form given in Fig. 2.1 and means that each J matrix can be separately constructed and diagonalized. Additionally, the rotational symmetry species Γ_{rot} for the rotational basis can be determined analytically (for non-cubic-group symmetries) from K and τ and the integral $\langle J,K'|J_\alpha|J,K''\rangle$ in Eq. (2.20) can also be be determined analytically.

2.3.3.2 Vibrational basis-set

The vibrational basis set function $|v\rangle$ in TROVE is constructed as product of 1-D basis-functions:

$$|v\rangle = |v_i^1\rangle|v_j^2\rangle \ldots |v_n^{3N-6}\rangle \tag{2.48}$$

Where the $|v_j^i\rangle$ is a basis-function representing a vibrational mode i with quantum number j. These are obtained from solving the 1-dimensional schrodinger equation:

$$H_n^1 = -\frac{\hbar^2}{2}\frac{\partial}{\partial \xi_n} G_{nn}^1(\xi_n)\frac{\partial}{\partial \xi_n} + U^1(\xi_n) + V^1(\xi_n) \tag{2.49}$$

where the superscript 1 represents the one dimensional reduction of the Hamiltonian by restricting all co-ordinates except for ξ_n to their equilibrium values. This is solved via Numerov-Cooley method [90, 91] on a set of grid-points. This basis-set is truncated via a polyad scheme up to a maximum defined polyad number P_{max}:

$$P = \sum_n a_n v_n \leq P_{max} \tag{2.50}$$

where a_n is a polyad coefficient. This makes it simple to control the size of the basis-set. However it can be seen that certain vibrations or *class* of vibration exhibit the same motion with the same species of atoms through symmetry (and by this, the same energy). We can combine them into a single contracted basis-set which exhibit better symmetry properties and simplifies the process of determining their irreducible representations. An example would be combining primitive stretches into a new contracted form:

$$|\phi^{s_1,s_2,\ldots,s_n}\rangle = \sum_{i,j,k,\ldots,n} A_{i,j,k,\ldots,n}^{v,s_1,s_2,\ldots,s_n} |v_i^{s_1}\rangle|v_j^{s_2}\rangle \ldots |v_j^{s_n}\rangle \tag{2.51}$$

Where $v_i^{s_i}$ are primitives of the same class s_i and $A_{i,j,k,\ldots,n}$ is solved via a reduced Hamiltonian:

$$H_{s_1,s_2,\ldots,s_n}^{red} = \langle v_0^{b_n}| \ldots \langle v_0^{b_2}|\langle v_0^{b_1}|H_{vib}|v_0^{b_1}\rangle|v_0^{b_2}\rangle \ldots |v_0^{b_n}\rangle \tag{2.52}$$

Where $v_0^{b_i}$ refer to primitive vibrational basis in their ground state that are not a basis for Eq. (2.51) (in this example b_i represents bends). The matrix is diagonalized and eigenvalues and irreducible representations are determined via Eq. (2.43) and stored for each class for later use. Our new vibrational basis set now has the form:

$$|\Phi\rangle^{\Gamma} = \sum_{i,j,k,\ldots,n} B^{v}_{i,j,k,\ldots,n} |v_i^{c_1}\rangle |v_j^{c_2}\rangle \ldots |v_n^{c_n}\rangle \tag{2.53}$$

where c_i referes to a particular class i and Γ is the total symmetry of the basis set from a product of all symmetries of each contracted basis function. Here $B^{v}_{i,j,k,\ldots,n}$ can be solved for by constructing the pure vibrational $J = 0$ hamiltonian matrix H_{vib} and diagonalizing for eigenvalues and eigenvectors:

$$\langle \Phi_i^{\Gamma_i} | H_{vib} | \Phi_j^{\Gamma_j} \rangle = E_{i,j}^{J=0,\Gamma_i \otimes \Gamma_j} \delta_{i,j} \tag{2.54}$$

where the integral is zero if Γ_{ij} satisfies Eq. (2.45) which for Abelian symmetry groups implies $\Gamma_i = \Gamma_j$. The eigenvectors are also calculated and stored. One last contraction step can be still be performed. When constructing the $J > 0$ Hamiltonian, the contracted vibrational basis-set can instead be replaced by the new $\Psi_v^{J=0}$ solutions. This is refered to as the $J = 0$ representation [92]. In this contraction step, a symmetry adapted basis-set is first constructed. This can be done via a standard projection operator but for Abelian symmetries this is easily accomplished by only selecting contracted basis with a specific symmetry species Γ. This has the advantage of making the J matrices block diagonal along Γ as seen in Fig. 2.2 which again this arises from Eq. (2.45). Doing this we can construct and diagonalize different Γ blocks separately for $J = 0$ and $J > 0$. We are now in possession of a new $J = 0$ wavefunction: $\Psi_v^{J=0,\Gamma}$. This wavefunction is used as the vibrational basis-set and we can truncate this through an eigenvalue threshold $E_{\max}^{J=0}$. Our ro-vibrational wavefunction now has the form:

$$\Psi_i^{J,\Gamma_{\rm tot}} = \sum_{v,K,\tau} c_{v,K,\tau}^{J,\Gamma_{\rm tot},i} \Psi_v^{J=0,\Gamma_{\rm vib}} |J,K,\tau\rangle \tag{2.55}$$

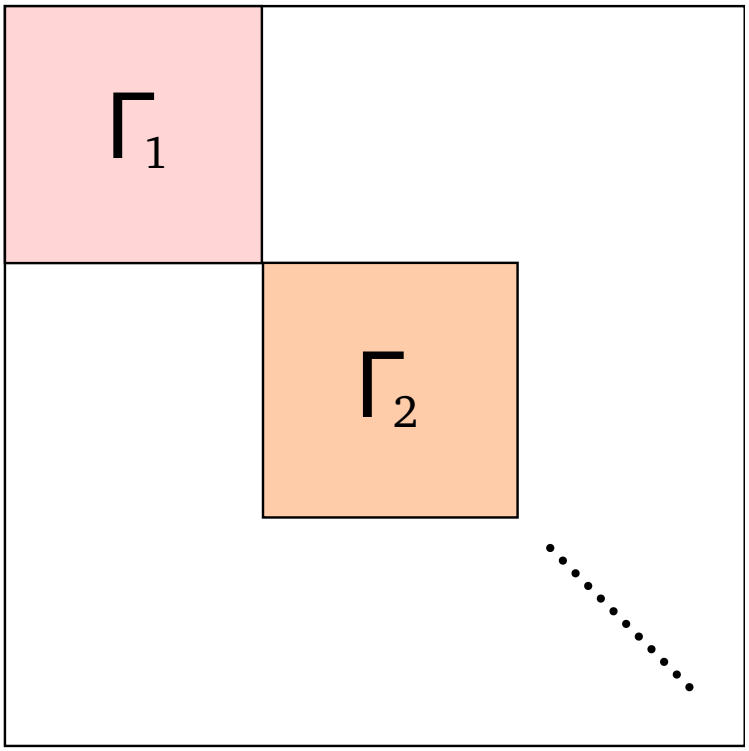

Figure 2.2: Block matrix structure of the symmetry adapted Hamiltonian for different Γ

Using Eq. (2.43) the $\Gamma_{\rm tot}$ for Abelian symmetries is determined as:

$$\Gamma_{tot} = \Gamma_{rot} \otimes \Gamma_{vib} \tag{2.56}$$

A symmetry adapted basis-set can therefore be constructed for $J > 0$ and Γ by selecting only $\Gamma_{tot} = \Gamma$. Substituting Eq. (2.55) into the Hamiltonian (Eq. (2.24)) for a particular J and Γ gives the matrix expression:

$$\begin{aligned} H^{J,\Gamma}_{i,i'} =& \langle J, K_i, \tau_i | H_{ro} | J, K_{i'}, \tau_{i'} \rangle \\ &+ \langle J, K_i, \tau_i | \Psi_i^{J=0,\Gamma_{\rm vib}} | H_{cor} | \Psi_{i'}^{J=0,\Gamma'_{\rm vib}} | J, K_{i'}, \tau_{i'} \rangle \\ &+ \langle \Psi_i^{J=0,\Gamma_{\rm vib}} | H_{vib} | \Psi_{i'}^{J=0,\Gamma'_{\rm vib}} \rangle \end{aligned} \tag{2.57}$$

This simplifies the vibrational contribution as:

$$\langle \Psi_i^{J=0,\Gamma} | H_{vib} | \Psi_{i'}^{J=0,\Gamma} \rangle = E^{J=0,\Gamma}_{i,i'} \delta_{i,i'} \tag{2.58}$$

which is diagonal and are simply the $J = 0$ energies calculated previously. This has the additional benefit where experimental band centres can be substituted instead of the eigenvalues. This has the effect of shifting the band centre onto the experimental value and allows further rotational excitations to be computed around the experimental band centres[92].

2.3.4 Diagonalization

The variational method of solving the Schrodinger equation takes the form of a standard eigenvalue problem:

$$A\mathbf{x} = \lambda\mathbf{x} \tag{2.59}$$

The A matrix is the Hamiltonian whose elements are given in Eq. (2.57), λ are the eigenvalues corresponding to energy of a state and $\mathbf{x}$ are the eigenvalues and are the variational coefficients in Eq. (2.55). TROVE implements several methods of diagonalizing a matrix. The most commonly used is the dense double precision symmetric eigenvalue solver DSYEV from the standard library LAPACK [93]. The matrices are generated and solved in seperate J,Γ blocks. The size of the Hamiltonian at $J > 0$ in Eq. (2.57) is related by J and the $J = 0$ energy truncation $E_{\max}^{J=0}$. The relationship is:

$$d_J^\Gamma = (2J+1)d_{J=0}^\Gamma \tag{2.60}$$

where $d_{J=0}^\Gamma$ is the size of the $J = 0,\Gamma$ basis-set determined by $E_{\max}^{J=0}$ and d_J^Γ is the size of the Hamiltonian at rotational excitation J and symmetry Γ. The memory in gigabytes (GB) required to hold a matrix of this size is:

$$\text{Matrix memory} = \frac{8(d_J^\Gamma)^2}{10^9} \tag{2.61}$$

The main high performance computing centres used in computing line lists come from the DiRAC supercomputing cluster, specifically the Cambridge Darwin and COSMOS computing clusters. The Darwin cluster provides about 64 GB of memory on each node which translates to a maximum matrix size of d_J^Γ =89,000. A typical molecule will have a $d_{J=0}^\Gamma$ in the order of 1,000 giving a maximum com-

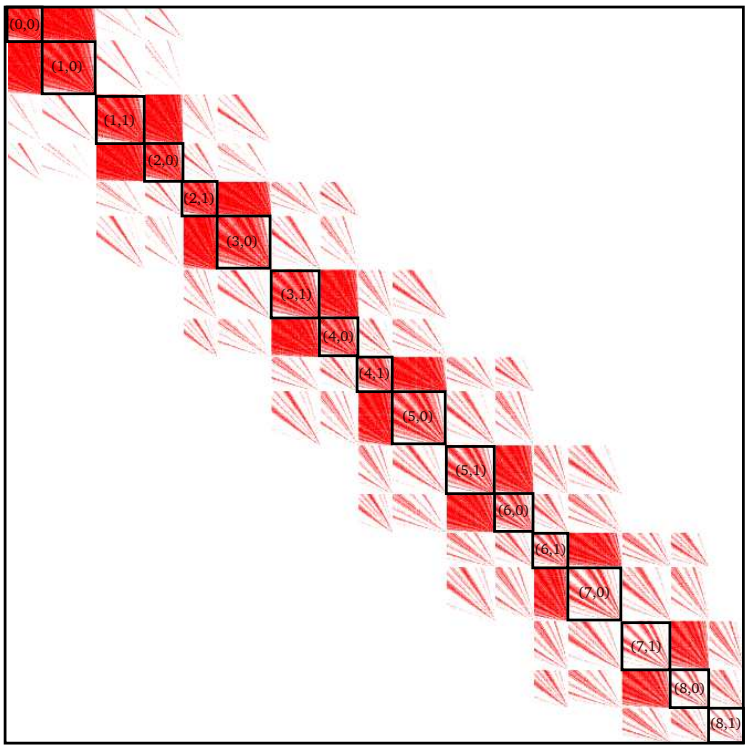

Figure 2.3: An image generated from the $J = 8, \Gamma = A_1$ matrix for H_2CO. The boxes highlight blocks belonging to specific (K, τ) quantum numbers. Colours approaching white represent values at each matrix element approching zero.

putable J around $J = 44$. However molecules such as H_2CO and H_2O_2 require $J > 60$ energy states in order to qualify for high temperature usage. One way is to exploit the sparsity of the matrix. A consequence of the rotational basis set is that the matrices are block diagonal along K. Fig. 2.3 shows that the only non-zero values in the matrix are around the diagonal. Therefore, by only storing non-zero elements, the memory requirements are significantly reduced. An iterative sparse diagonalizer like ARPACK[94] can instead be used to solve for a set number of eigenvalues. Typically, iterative diagonalizers are used to solve for a ≈1–10% of the eigenvalues. For high J, this is advantageous as the number of eigenvalues within the energy limit $E_{\max}^{J=0}$ is significantly smaller compared to the huge size of the matrix. Figure 2.4 highlights the relationship between J, number of eigenvalues and non-zero elements for the PH_3 SAlTY linelist[9]. However for J matrices that are generally large and require 50–90% of their eigenvalues, these iterative methods are extremely inefficient to use.

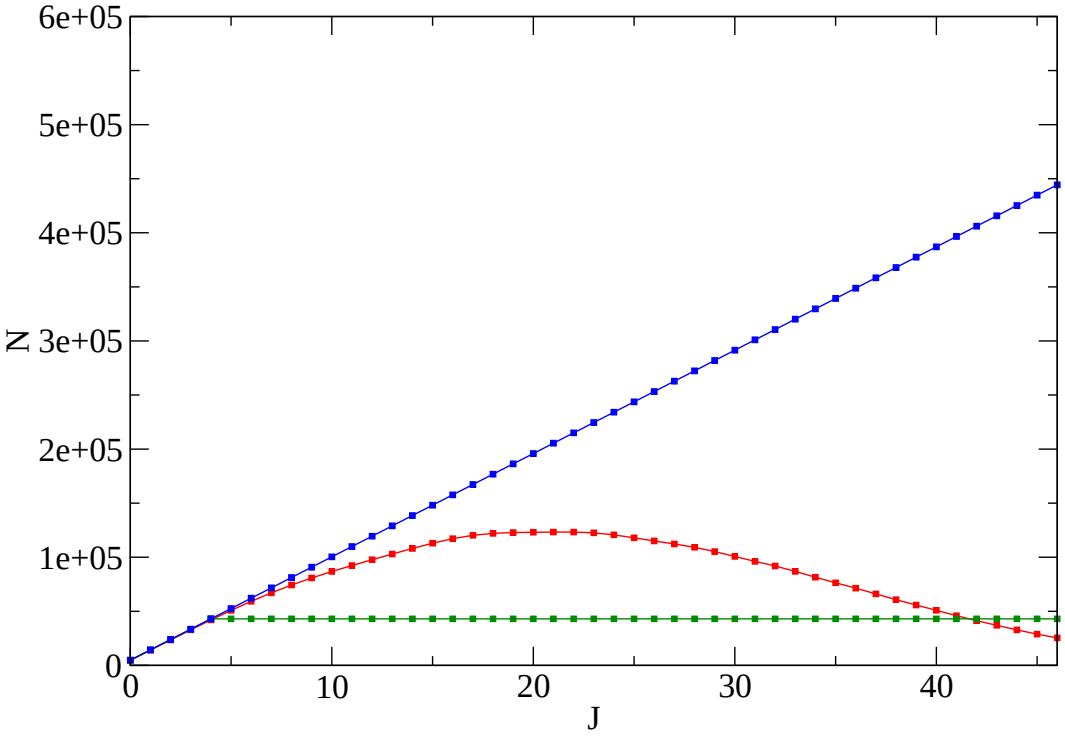

Figure 2.4: Dimensions of the SAlTY[9] matrices with J (blue), the corresponding number of eigenvalues below a threshold (red) and number of non-zero elements on each row (green).

For matrices that are too big and dense for Darwin, diagonalization moves to the COSMOS cluster. COSMOS is a Non-Uniform Memory Access (NUMA) shared memory system and can link multiple nodes into a single memory area providing terabytes of usable memory. This provides us the ability to diagonalize large dense matrices but at the cost of poor performance where diagonalization can take days to complete. The reason for this comes from the *NUMA* architecture itself. Generally when allocating memory in Linux, the memory isn't truly allocated to hardware until a process initializes or 'touches' it (assigned a value, set to zero etc.). For NUMA, this memory is allocated to the node that first touched this memory. The problem comes when another process from a different node wants to access it, it must travel through a slower interlink in order to retrieve the value. LAPACK was not designed for these shared memory systems and therefore does not take into account the communication overhead that is occurring in the background to maintain the single memory space.

2.3.4.1 MPI diagonalizer

Message Passing Interface (MPI) is a system that allows for different computers or nodes to communicate with each other to complete a single given task. Programs designed using this interface are typically more efficient when using multiple nodes as they are known to be *communication-aware*. Therefore a linear algebra library that is communication-aware is required. SCALAPACK [95] is a distributed parallel linear algebra library based on LAPACK. It is built around MPI and can make full use of nodes for diagonalization via the parallel double precision symmetric eigenvalue solver *PSDYEV*. The matrix is distributed across all processors with more processors providing more memory. An external program was written that is able to read matrices exported from TROVE, diagonalize over hundreds of cores and outputs in TROVE's eigenvalue and eigenvector format directly. A matrix of size $\approx$300,000 diagonalized over 512 cores only takes around 2 hours to complete for all eigenvalues and eigenvectors.

2.3.4.2 Quantum number assignments

State quantum numbers are assigned using the largest coefficient in the eigenvector. The ro-vibrational state (of specific J, Γ) with wavefunction in Eq. (2.55) is assigned the quantum numbers v, K and τ from the basis set with the largest contribution i.e the basis function with the largest variational coefficient $(c^{J,\Gamma,i}_{v,K,\tau})^2$. The criterion for unique and well-defined assignment is $(c^{J,\Gamma,i}_{v,K,\tau})^2 > 0.5$ as given by Hose and Taylor [96]. Eigenstates that do not meet this criterion are likely to exhibit pseudo-degeneracies but these usually occur in highly vibrationally excited states and are generally rarely observed experimentally. Additionally the vibrational quantum numbers are local mode and refer to the basis-set arrangement defined in a TROVE input file[97], these usually do not correspond to the normal mode quantum numbers that is commonly used in literature. TROVE provides a facility to reassign these quantum numbers for each contracted class based on their energies. This facilitates matching states to literature line positions and is useful for less well-behaved molecules such as H_2O_2. Details of this can be seen in Chapter 5.3.

2.3.5 Refinement of the Potential Energy Surface

Generally, the current level of *ab initio* electronic structure theory does not give experimental accuracy for ro-vibrational energies within a resonable timeframe. The empirical shifting technique given in Eq. (2.58) is therefore a means to achieve this degree of accuracy with minimal computational effort. Whilst band-centre shifting can reproduce experimental energies very easily, its accuracy and predictive ability is limited to vibrational bands whose band centre positions are already experimentally characterised. A more robust method of correcting the PES is through fitting or refining to experimental energies. The procedure [98] implemented in TROVE describes a correction surface ΔV to the initial (*ab initio*) surface V. The new refined surface V' can therefore be written as $V' = V + \Delta V$ where ΔV is expanded in Taylor series with expansion coefficients $\Delta f_{ijkl\ldots}$:

$$\Delta V = \sum_{i,j,k\cdots} \Delta f_{ijk\ldots} \xi_1^i, \xi_2^j, \xi_3^k \cdots \tag{2.62}$$

Our new Hamiltonian has the form $H' = H + \Delta V$, where H is the Hamiltonian of the starting point for the refinement. Therefore we can write a new corrected energy E' with respect to the *ab initio* energies E as:

$$\begin{aligned} \langle \Psi_n^{J,\Gamma} | H' | \Psi_m^{J,\Gamma} \rangle &= E_{n,m}^{J,\Gamma} \delta_{n,m} + \langle \Psi_n^{J,\Gamma} | \Delta V | \Psi_m^{J,\Gamma} \rangle \\ E_{n,m}^{\prime J,\Gamma} &= E_{n,m}^{J,\Gamma} + \sum_{i,j,k\cdots} \Delta f_{ijk\ldots} \langle \Psi_n^{J,\Gamma} | \xi_1^i, \xi_2^j, \xi_3^k \cdots | \Psi_m^{J,\Gamma} \rangle \end{aligned} \tag{2.63}$$

Interestingly, the differential of the energy with respect to adjustible parameters can be found using the Hellmann–Feynman theorem [99]:

$$\frac{\partial E_{n,m}^{J,\Gamma}}{\partial \Delta f_{ijkl\ldots}} = \langle \Psi_n^{J,\Gamma} | \xi_1^i, \xi_2^j, \xi_3^k \cdots | \Psi_m^{J,\Gamma} \rangle \tag{2.64}$$

Which has the same form as the second term in Eq. (2.63), therefore we can simplify the corrected energy as:

$$E'^{J,\Gamma}_{n,m} = E^{J,\Gamma}_{n,m} + \sum_{i,j,k\cdots} \Delta f_{ijk\cdots} \frac{\partial E^{J,\Gamma}_{n,m}}{\partial \Delta f_{ijkl\cdots}} \tag{2.65}$$

For simplicity we will refer to a set of $i, j, k\cdots$ parameters under a single label n. We therefore require a set of Δf_n that provides the best overall modification in the calculated energies towards experimental observation. An objective function F can be constructed that represents this:

$$F = \sum_{i}^{N} w_i (E_i^{obs} - E_i)^2, \tag{2.66}$$

where N is the number of observed energies E_i^{obs}, E'_i are the corresponding calculated energies and w_n are the weights which dictate the importance of experimental energies in the refinement. The task is therefore to find a set of coefficients Δf_n that minimize Eq. (2.66). The requirements for the minimum is for $\frac{\partial F}{\partial \Delta f_n} = 0$ which in conjunction with Eq. (2.65) gives:

$$\sum_{i}^{N} w_i \frac{\partial E_i}{\partial \Delta f_n} \frac{\partial E_i}{\partial \Delta f_m} \Delta f_m - \sum_{i}^{N} w_i (E_i^{obs} - E_i) \frac{\partial E_i}{\partial \Delta f_n} = 0 \tag{2.67}$$

this is a linear equation of the form:

$$A\mathbf{x} - \mathbf{b} = 0 \tag{2.68}$$

which can be minimized using linear least squares fitting from linear algebra libraries such as LAPACK [93]. The refinement procedure can cause the PES to assume unphysical shapes so the fitting is constrained around the energies at each *ab initio* geometries [100, 98, 97, 9]. Great care must still be taken in ensuring that the PES is not over-fit as it can subtly destroy the behaviour of coupled vibrational states, a practical example of this is discussed in Chapter 4.5. This PES refinement methodology used in TROVE has been successful in producing accurate PES for

molecules such as NH_3 [101], SO_3 [102], PH_3 [9] and CH_4 [64].

Chapter 3

Transitions and GAIN

With the procedures outlined in the previous chapter we are now in possession of eigenvalues and eigenvectors for a range of excitations. However, this only represents the first half of a line list calculation. These states give us great insight into the behaviour of the molecule in isolation, however, the observation of molecules requires them to be subjected to an electric field of some kind (such as a photon) and their interaction with this field must be understood in order to model systems such as atmospheres.

3.1 The Dipole Moment

A molecule in the most general sense is a system of charged particles. Charge particles also exhibit an electric field. When the molecule's electric field interacts with an electric field ε the associated interaction energy E is defined as:

$$E = -\sum_{A=X,Y,Z} \mu_A \cdot \varepsilon \tag{3.1}$$

Where μ_A is the dipole moment vector and is defined with Cartesian co-ordinate A:

$$\mu_A = \sum_r C_r e A_r \tag{3.2}$$

The dipole moment vector therefore describes the charge distribution or the electric field in a molecule. A molecule that exhibits a non-zero dipole moment in its equilibrium configuration is said to have a permanent dipole moment. This is com-

monly defined quantity in literature and is usually given in units of Debye (D). Like the potential energy surface, the dipole moment for each nuclear geometry of the molecule can be computed and an N dimensional hypersurface can be constructed known as the Dipole Moment Surface (DMS). This surface can also be computed *ab initio* in a similar fashion to the electronic structure problem. However, unlike the potential energy, the dipole moment is a vector and not a scalar quantity. Therefore a DMS must be constructed for each Cartesian co-ordinate in order to fully describe both the direction and magnitude at each geometry. The strength or 'intensity' of an electric field interaction is defined by the expectation value of E:

$$\langle E \rangle = - \sum_{A=X,Y,Z} \langle \Psi_n | \mu_A | \Psi_m \rangle \cdot \varepsilon \tag{3.3}$$

If the matrix element given by the integral is non-zero in the off-diagonal, it implies that a transition is occuring from state Ψ_m to state Ψ_n. Transitions are therefore driven by the interaction of the external and molecules electric field. The probability or 'strength' of a transition is square of this integral and is written as:

$$S(m \leftarrow n) = \sum_{A=X,Y,Z} |\langle \Psi_m | \bar{\mu_A} | \Psi_n \rangle|^2 \tag{3.4}$$

this is commonly known as the line-strength. The symmetry of the dipole moment operator gives insight into the manner of these transitions. Applying symmetry operations to Eq. (3.2), a permutation symmetry operation will leave the dipole moment unchanged while an inversion ($A_r = -A_r$) permutation operation will change the sign of the dipole. Therefore the symmetry of the dipole is the symmetry species of a Molecular Symmetry group that has symmetric permutation operations and anti-symmetric inversion-permutation operations. Table 3.1 highlight a few of these symmetries for different symmetry groups. This property allow us to apply the vanishing integral rule given in Eq. (2.45) to Eq. (3.4). The rules for when this integral is zero are:

$$\Gamma(\Psi_n) \otimes \Gamma(\Psi_m) \not\supset \Gamma(\mu_A) \tag{3.5}$$

Table 3.1: Symmetry of the dipole moment operator for various Molecular Symmetry (MS) groups from Bunker and Jensen [20]

MS Group	$\Gamma(\mu_A)$
C_2	A
C_{2v}	A_2
C_{2h}	A_u
D_{2h}	A_u
D_{3h}	A_1''

star and forms the basis of the symmetry selection rules. If the linestrength evaluates to zero, these transitions are considered *forbidden*. Such a term is a misnomer as these transitions aren't truly forbidden but are considered weak under the Born-Oppenheimer approximation[88]. The main features of a spectrum can still be understood without considering them.

3.2 Simulating Spectra

By evaluating Eq. (3.4) important spectroscopic quantities can be derived, most importantly the Einstein-*A* coefficents:

$$A_{fi} = \frac{8\pi^4 \tilde{\nu}_{if}^3}{3h}(2J^f + 1)S(f \leftarrow i) \tag{3.6}$$

where $\tilde{\nu}_{if} = E_f - E_i$ is the transition frequency. This quantity is temperature-independent and with this the temperature dependent absolute intensity can be computed:

$$I(f \leftarrow i) = \frac{A_{fi}}{8\pi c} g_{\text{ns}}(2J_f + 1)\frac{\exp\left(-c_2\frac{E_i}{T}\right)}{Q\ \tilde{\nu}_{if}^2}\left[1 - \exp\left(\frac{-c_2\tilde{\nu}_{if}}{T}\right)\right], \tag{3.7}$$

where E_i is the computed energy for state i, $\tilde{\nu}_{if} = E_f - E_i$ is the transition frequency, T the absolute temperature, g_{ns} is the nuclear spin statistical weight factor and c_2 is the second radiation constant defined as:

$$c_2 = \frac{hc}{k} \tag{3.8}$$

The partition function Q can be computed by the explicit summation of states [46]. This is given by:

$$Q = \sum_i g_i \exp\left(\frac{-c_2 E_i}{T}\right), \tag{3.9}$$

where g_i is the degeneracy of a particular state i with energy E_i. By providing only Einstein-A coefficents, the spectra of a molecule at any temperature can be easily modelled provided it is adequately populated with enough rotational excitations.

3.3 Evaluating the Linestrength

Our dipole is represented in space-fixed Cartesian co-ordinates (μ_s). It is often more practical to evaluate the linestrength using spherical tensor representations instead of the Cartesian representation of the dipole moment expressed in terms of the molecular-fixed (μ_m) components as given by Bunker and Jensen [20]:

$$\mu_s^{1,\sigma} = \sum_{\sigma'=-1}^{1} D_{\sigma,\sigma'}(\phi,\theta,\chi)\mu_m^{1,\sigma'} \tag{3.10}$$

were

$$\mu_m^{1,\pm 1} = \frac{\mp\mu_X - i\mu_Y}{\sqrt{2}}, \quad \mu_m^{1,0} = \mu_Z \tag{3.11}$$

Now Eq. (3.4) becomes:

$$S(f \leftarrow i) = \sum_{\sigma=-1}^{1} \Big| \sum_{\sigma'=-1}^{1} \langle \Psi^f | D^1_{\sigma,\sigma'} \bar{\mu_m}^{1,\sigma'} | \Psi^i \rangle \Big|^2 \tag{3.12}$$

Between a lower state with energy E^i and upper state with energy E^f. The dipole is dependent only on vibrational coordinates and the $D^1_{\sigma,\sigma'}$ is dependent on rotational (Euler) coordinates, therefore substituing the wavefunction from Eq. (2.55) into Eq. (3.12) and employing Clebsch-Gordan algebra to evaluate the rotational part

gives [20]:

$$S(f \leftarrow i) = (2J'+1)(2J+1)\times|\sum_{\nu',K',\tau'}\sum_{\nu,K,\tau}\sum_{\sigma=-1}^{1}(-1)^{K}c_{\nu',K',\tau'}^{J',\Gamma',f}c_{\nu,K,\tau}^{J,\Gamma,i}\begin{pmatrix}J & 1 & J'\\ K & \sigma & -K'\end{pmatrix}$$
$$\langle\Psi_{\nu'}^{J=0,\Gamma'_{\rm vib}}|\bar{\mu_m}^{1,\sigma}|\Psi_{\nu}^{J=0,\Gamma_{\rm vib}}\rangle|^2 \tag{3.13}$$

From the 3-j symbols, the J selection rules for transitions are:

$$\Delta J = 0,\pm 1 \quad J+J' \geq 1 \tag{3.14}$$

For individual matrix elements, the non-zero elements are determined by the K quantum number:

$$\Delta K = \sigma = 0,\pm 1 \tag{3.15}$$

which effectively eliminates the summation over all σ. There are two properties we can exploit. Firstly the symmetry-adapted wavefunction can inflated back into non-symmetry-adapted form $\Psi_{\nu}^{J=0}$ with coefficients $c_{\nu,K,\tau}^{J}$. In this process, any $c_{\nu,K,\tau}^{J}$ that do not possess the same symmetry as the wavefunction are set to zero. Therefore we will retain the label $c_{\nu,K,\tau}^{J,\Gamma}$ for the coefficients that are non-zero. This appears to be counterproductive at first as the size of the wavefunction increases significantly but by doing this, we can precompute all elements of the integral $\langle\Psi_{\nu'}^{J=0}|\bar{\mu_m}^{1,\sigma'}|\Psi_{\nu}^{J=0}\rangle$ and form the three dipole matrices: $\mu_{\nu,\nu'}^{\Delta K}$ saving significant computation time. Secondly, the 3-j symbols can also be precomputed as well producing another matrix: $F_{\Delta J,\Delta k}$. With this, the linestrength simplifies to:

$$S(f \leftarrow i) = (2J'+1)(2J+1)|\sum_{\nu',K',\tau'}\sum_{\nu,K,\tau}(-1)^{K}c_{\nu',K',\tau'}^{J',\Gamma',f}c_{\nu,K,\tau}^{J,\Gamma,i}F_{\Delta J,\Delta K}\mu_{\nu',\nu}^{\Delta K}|^2 \tag{3.16}$$

therefore the only varying quantity is the variational coefficients. However this is an $O(N^3)$ operation where N is the size of the basis set, and evaluating it for each of the possibly billions of transitions with basis sets sizes that could reach in the millions is cumbersome and inefficient. Therefore a two-step strategy is developed.

First the computation is split along the lower state wavefunction $\Psi^{J,i}$:

$$s^{J',\Gamma'}_{v',K',\tau'}(\leftarrow i) = \sum_{v,K,\tau} c^{J,\Gamma,i}_{v,K,\tau}(-1)^K F_{\Delta J,\Delta K}\mu^{\Delta K}_{v',v} \tag{3.17}$$

where $s^{J',\Gamma}_{v',k'}(\leftarrow i)$ is a vector that represents a 'half' transition from a lower state i to an undetermined state with J',Γ'. This is referred to as the *half linestrength* and is an $O(N^2)$ operation. A transition to any state $\Psi^{J',\Gamma',f}$ that satisfies Eq. (3.14) can then be completed by performing a dot product:

$$S(f \leftarrow i) = (2J'+1)(2J+1)|\sum_{v',K',\tau'} c^{J',\Gamma',f}_{v',K',\tau'} s^{J',\Gamma'}_{v',K',\tau'}(\leftarrow i)|^2 \tag{3.18}$$

which is an $O(N)$ operation. This two step procedure has many advantages. When computing transitions up to a maximum frequency $\nu_{\max}$, the maximum energy of lower states $E^i_{\max}$ is determined by the relation:

$$E^i_{\max} = E^f_{\max} - \nu_{\max} \tag{3.19}$$

where $E^f_{\max}$ is the maximum upper state energy. Because of this, $N^i < N^f$ where N^i and N^f are the number of lower and upper states respectively. Additionally this has the consequence $N^t >> N^i$ where N^t is the number of transitions. From this, majority of work is performed by the cheaper and faster dot product in Eq. (3.18) instead of the more expensive Eq. (3.17). This two step method is the methodology used.

3.4 Computing linestrengths in TROVE

TROVE's input is controlled by keywords and makes use of Stone's input parser [103]. Computing a transition requires modifying a TROVE input file to include an intensity block with the keywords given in Table 3.2. An example of the intensity part of the TROVE input file is given in Figure 3.4. The dipole matrix elements $\mu^{\Delta K}_{v',v}$ are precomputed and stored in a checkpoint file. This is only done once in the entire TROVE pipeline and is read into memory for every transition run. At each run

Table 3.2: Keywords used in a TROVE input file.

Keyword	Comment
mem	total memory in gb
symgroup	Molecular symmetry group
intensity	Beginning of intensity block
absorption	Required keyword
thresh_line	Smallest linestrength to output
thresh_intens	Smallest absolute intensity to output
thresh_coeff	$c^J_{v,K}$ below this will be skipped in Eq. (3.17)
temperature	Temperature of calculation
qstat	Partition function Q
gns	nuclear statistical weight
selection	symmetry selection rules
J	J range of calculation
zpe	zero point energy
Freq-window	Frequency range of calcuation
energy	*low*: E^i min,max — *upper*: E^f min,max
end	end of intensity block
( )	comments

the matrix $F_{\Delta J,\Delta K}$ for a specified J range is computed and stored in memory, while the eigenvalues and quantum number assignments of states for all J in the range required are loaded and sorted by energy. The transition calculation occurs by looping through each lower state eigenvector $\Psi^{J,\Gamma,i}$ within the corresponding lower state energy range and computing all possible half linestrengths. For each lower state J,Γ,i all half linestengths (see Eq. (3.16)) are computed for each J' and Γ within the upper energy range that satisfy the selection rules given by Eqs. (3.14,3.5). OpenMP is utilized during these half linestrength computations independently working on K' and v' amongst available cores. The sum in Eq. (3.16) is restricted to $|K-K'| \leq 1$ and is also subject to the the expansion coefficient threshold condition

$$|c^{J,\Gamma,i}_{v,K}| \leq C_{\text{thresh}},$$

with C_{thresh} of about $1\times10^{-12} - 1\times10^{-16}$. After this stage, a nested loop for upper state eigenvectors $\Psi^{J',\Gamma',f}$ is executed for for energies and frequencies within the ranges requested, where the appropriate matrix $s^{J',\Gamma',i}_{v',k'}(\leftarrow i)$ is selected for the dot

```
MEM 64 gb
SYMGROUP C2V(M)
(Other inputs)
INTENSITY
   absorption
  THRESH_INTES  1e-40
  THRESH_LINE   1e-40
  THRESH_COEFF  1e-40
  temperature   1000.0
  QSTAT        33314.25
  GNS          1.0  1.0  3.0  3.0
  ZPE  5773.228049563373
  selection  1 1 2 2
  J,  8,9
  freq-window  -0.001,  10000.0
  energy low   -0.001, 8000.00, upper   -0.00, 18000.0
END
(Other inputs)
```

Figure 3.1: An example TROVE input with only relevant keywords for computing intensities for the H_2CO molecule

product in Eq. (3.18). The dot-product is evaluated using the vendor specific BLAS [104] sub-routine. In order to reduce the input/output (I/O) the needed eigenvectors are batched into RAM. Since transitions from different lower states are independent, the intensity calculations can be split into independent sub-ranges for $E_n^{(i)} \leq E_i \leq E_{n+1}^{(i)}$ and run in parallel over different nodes. After this stage, a nested loop of upper states $\Psi^{J',f}$ are filtered for energy and frequency range and the appropriate $s_{\nu',K',\tau'}^{J',\Gamma,i}(\leftarrow i)$ is selected for the dot product. The dot-product for Eq. (3.18) comes from a vendor specific BLAS[104].

The main two factors that dictate the completion time of intensity calculations are (i) the total number of transitions (controlled via the keywords in Table 3.2) and (ii) the size of the basis set.

The size of a particular basis set d_J is given by:

$$d_J = (2J+1)\, d_{\text{vib}}, \tag{3.20}$$

where d_{vib} is the size of the corresponding vibrational basis set. Therefore as J

Table 3.3: Times in seconds for computing the half linestrength for molecules. H_2CO, PH_3 and SO_3 are from the AYTY (Chapter 4), SAlTY [9] and UYT2 [21] hot line lists, respectively. The last row gives the total time in hours to compute half linestrengths for the most dense $J \leftrightarrow J'$ with about 4000 lower states

J	H_2CO	PH_3	SO_3
10	3.46	13.84	13.84
20	7.43	29.72	131.87
30	11.30	45.20	263.01
40	16.32	65.28	381.05
50	21.25	85.01	512.19
60	25.89	103.57	625.15
70	30.07	120.27	828.25
Dense J Total Time (hours)	12.56	33.02	920.28

increases the problem size increases. The time scaling for the half linestrength is $O(d_J^2)$ and for the dot-product, $O(d_J)$. For large basis sets, the largest bottleneck comes from the half linestrength evaluation itself. For example, basis sets of around $\approx 10^6$ can take between 30 seconds to 10 minutes per lower state to complete with typically thousands of lower states per selected intensity run, especially for higher excitations. Table 3.3 shows typical half linestrength times for a number of molecules as well as the total time spent performing this preprocessing step in the most demanding cases. Heavier molecules incur a greater burden at this step with SO_3 requiring over a month of wallclock time.

Additionally there is an issue with the dot-product in terms of the load-balancing; whether it is more efficient to use all cores for a single dot product or a single core for multiple dot products. The choice made can cause conflicts with other steps such as eigenvector inflation and filtering thus affecting the throughput of intensity calculations. The problem lies with the extremely small number of cores in CPUs that prevent the more balanced approach of calculating multiple transitions as quickly as possible.

However accelerators such as graphics processing units (GPU) contain a larger number of cores that can facilitate both the half linestrength calculation as well as allowing for a more balanced core distribution between the parallel calculation of transitions and the dot product itself.

3.5 GPU architecture

This section describes the terminology related to GPUs and features exploited in the program. The terminology and devices are based on the Compute Unified Device Architechture (CUDA). GPUs contain multiple streaming multiprocessors (SM) each containing physical cores used in execution, a register space and a small user-managed cache (32KB) called *shared memory*. There is also a large on-board memory (1-12 GB) accessible by all SMs called *global memory*. A CPU issues commands to be executed by GPUs by calling routines known as *kernals*.

The hierarchy of memory is that registers are extremely fast (a few cycles) with the scope of a single thread, the shared memory is moderately fast (10-20 cycles) with the scope of the thread block and the global memory is slow (400 cycles) with the scope of the entire GPU. Each thread can read and write to the global memory and these persist across multiple kernals whilst shared and register memory are lost. Global memory reads can be improved by ensuring ordered access (*memory coalescing*).

All performance characteristics shown below are with the I/O time removed and using the eigenvectors and dipole moments from the AYTY linelist (Chapter 4) calculations [105] unless otherwise stated. The AYTY line list has vibrational basis set size of $d_{\text{vib}} = 7\ 642$ and a maximum J at 70 giving a maximum basis set size of $d_J = 1\ 077\ 522$. The system used in measuring performance is the Emerald CfI cluster and comprises of 12 cores (two 2.50 GHz six-cores Intel Xeon E5-2640 processors) connected via NUMA with 8 nVidia Tesla M2090 GPUs attached (Fermi architecture).

3.6 Cache and Reduce Kernal

A naive *Baseline* kernal was produced that implements Eq. (3.17) almost exactly for TROVE. The basis-sets, 3-j symbols and dipole matrices are put into the GPUs global memory. Figure 3.2 shows the performance gain from the baseline kernal. Here we see that the GPU gives us a free speedup at smaller basis-sets but becomes much less efficient at higher dimensions with only a 10 % speedup at best. The

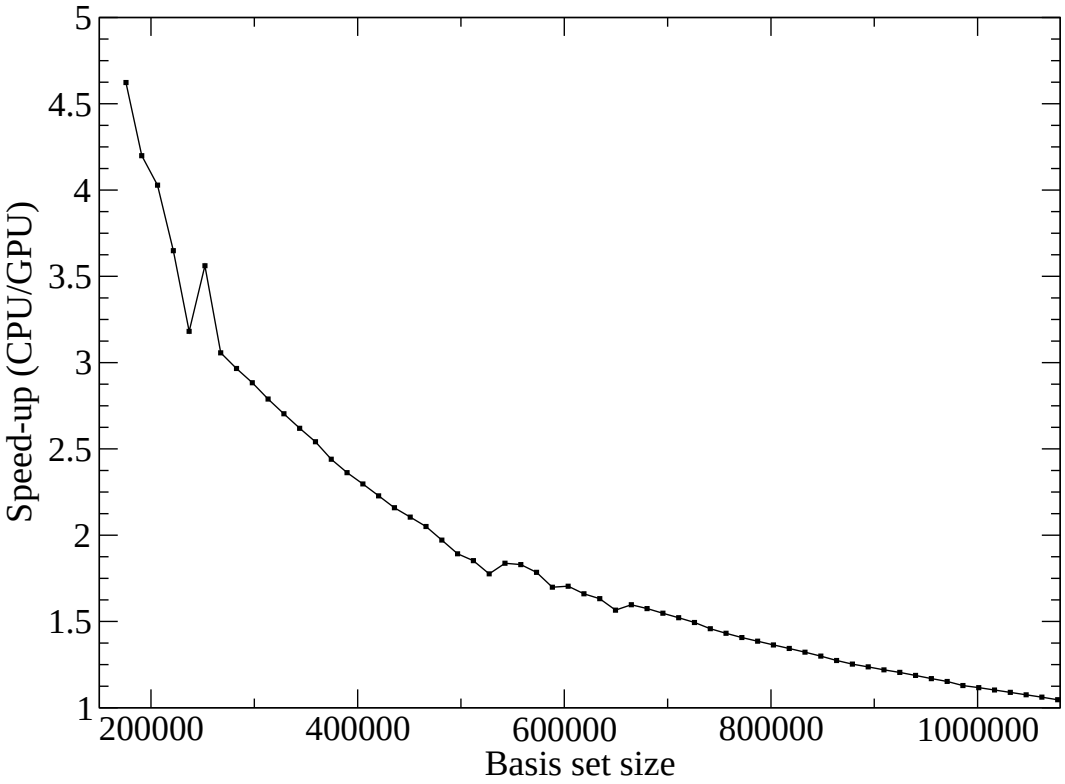

Figure 3.2: A plot showing the speed-up when directly implementing Eq. (3.17) to CUDA

reason for this is the sheer number of global reads for the quantum numbers and coefficients required in order to complete. A refactoring in the overall methodology is required. Firstly the basis set are arranged in increasing K effectively creating $J+1$ blocks. Eq. (3.17) can also be decomposed into two further steps. The first is a caching step:

$$C^{J',\Gamma',K,\Delta K,i}_{\nu',K',\tau'} = \sum_{\nu} c^{J,\Gamma,i}_{\nu,K,\tau'} F_{\Delta J,\Delta K} \mu^{\Delta K}_{\nu',\nu} \tag{3.21}$$

$C^{J',\Gamma',K,\Delta K,i}_{\nu',K',\tau'}$ is the half linestrength belonging to a specific K block that is completed with a specific Δk. The second is a 'reducing' step:

$$s^{J',\Gamma'}_{K',\nu',\tau'}(\leftarrow i) = C^{J',\Gamma',K,-1}_{K',\nu',\tau'} + C^{J',\Gamma',K,0}_{K',\nu',\tau'} + C^{J',\Gamma',K,1}_{K',\nu',\tau'} \tag{3.22}$$

In Eq. (3.21), all threads are guaranteed the same $c^{J,\Gamma}_{\nu,K,\tau}$ and $F_{\Delta J,\Delta K}$ making them easily cached into shared memory and reused in computation and the value for K is implicit based on which block is being executed. Therefore the only global

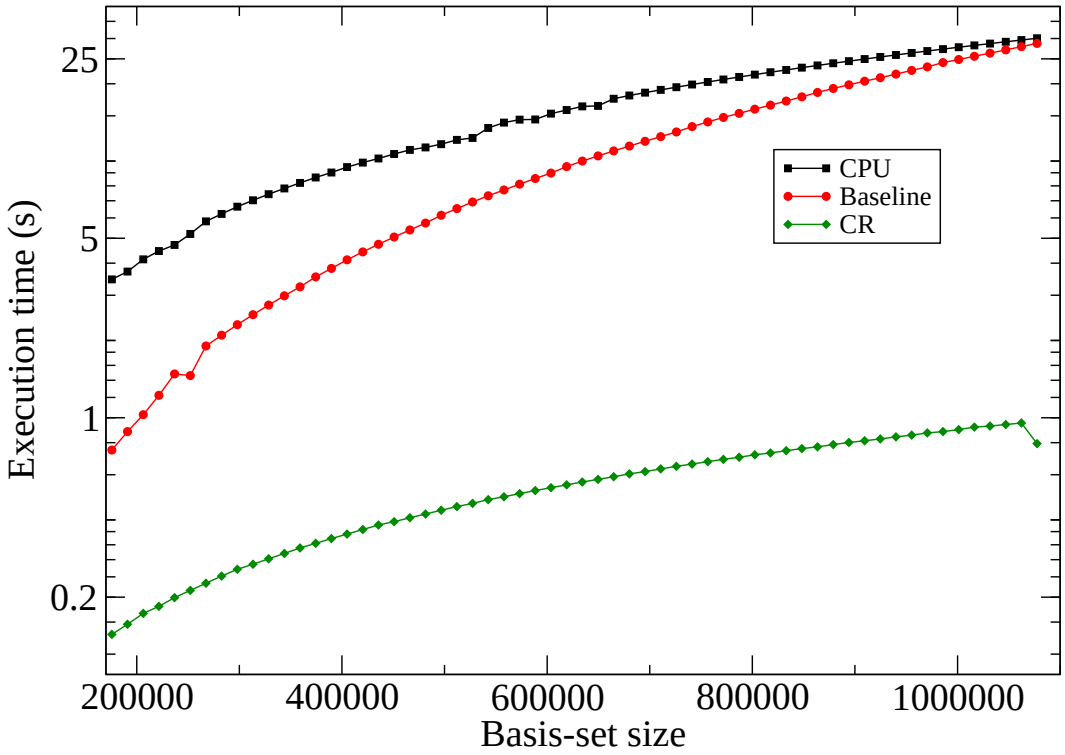

Figure 3.3: Half linestrength call time vs basis-set size; note time is given on a logarithmic scale.

memory read required are the ν indices and $\mu^{\Delta K}_{\nu',\nu}$ matrices. The dipole reading is further coalesced as all threads will read around the same area of memory. This kernal will be referred to as the *CR Kernal* (Cache and Reduce Kernal). The kernal is called $(2J'+1)$ times in order to complete and each call is able to be performed simultaneously, to a degree, by using multiple streams as they are each independent of each other. Figure 3.3 compares the efficiency of the *CR Kernal* to that of CPU and *Baseline* showing a substantial improvement with up to $30\times$ speed-up from the CPU call time. The *CR Kernal* can complete the half linestrength in less than a second for the largest basis set and speed-ups gained from this kernal increase with the growing basis set size. This comes entirely from the reduced global memory reads, utilization of the fast shared memory and data re-use. It should be noted that the threshold eigen-coefficient C_{thresh} is not applied here in contrast to the case above, i.e. there are no associated losses of accuracy.

3.6.1 Large Dipole Matrices

For the formaldehyde molecule in the AYTY hot line-list, $d_v = 7642$. This gives a dipole of dimension $7462 \times 7462 \times 3$ which is ≈ 1.2 GB using double precision. This can be fit easily into the M2090 or K20 GPU's global memory. However molecules such as PH_3 and SO_3 require d_v of 14386 and 15948 respectively. For the PH_3 case this gives a dipole of memory size 4.6 GB which only barely fits into the M2090. Unfortunately SO_3 requires ≈ 5.8 GB to store which is unfeasible for both the M2090 and K20 GPUs. The easiest solution would be to utilize the larger memory K40 and K80 GPUs to perform the computation but these are not well adopted in the HPC community. Another solution is to adopt a strategy of partitioning the dipole into blocks that can fit in GPU memory and calling a modified version of the CR kernal for each block of the dipole. The dipole matrix elements depend on v and v'. Each thread requires all of v' to complete a particular K' and v' and the matrix is therefore partitioned by v into p blocks. The relation between how the dipole matrix and Ψ' is shown in Fig. 3.4. A matrix block is transferred into the GPU and calls a block-CR kernal for each K. The block-CR kernal is almost exactly the same as the CR kernal only it spawns threads for v which exist in the matrix block. This is repeated for each matrix block in order to complete the half linestrength.

Figure 3.5 shows how the speedup gain varies over differing p values. Overall, there is a performance reduction with increasing block-size with a 2x reduction for $p = 2$ and 3x for $p = 3$. This makes sense as we effectively need to call the kernal p multiple times to complete the half linestrength as given by block-CR. Such blocking method is more beneficial for more difficult molecules such as SO_3. Figure 3.6 uses the basis-sets and vectors from the SO_3 hot line-list given by Underwood et al. [21]. As the dipole is too large to fit in the M2090 memory, only partitioned dipoles are shown. Here the speed-up given is significant and the change in speed-up between p values ranges by only $\approx 20-30\%$. This is because of the basis set that can easily saturate the GPU with work even after splitting.

Overall, the half linestrength is up to 70x faster than the CPU-only version and

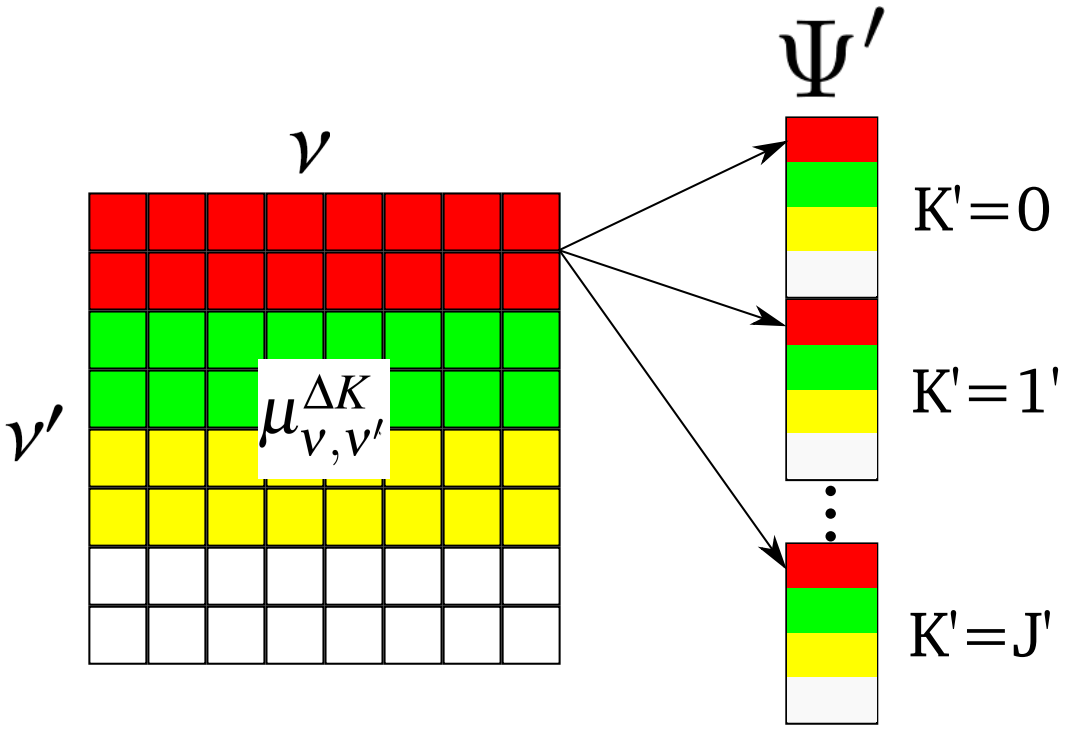

Figure 3.4: A visual representation of blocking the dipole matrix elements with $p = 4$, the colors and arrows show how each matrix block relates the K-blocks in state Ψ'

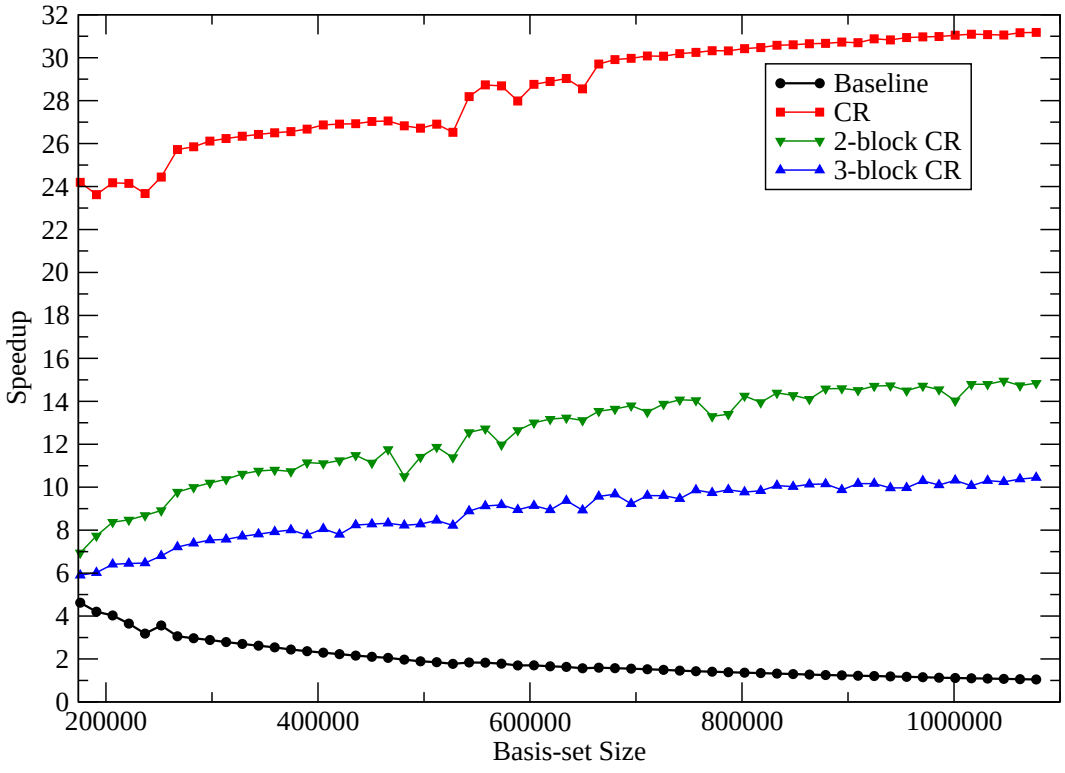

Figure 3.5: Performance characteristics of the CR kernal with varying values for p

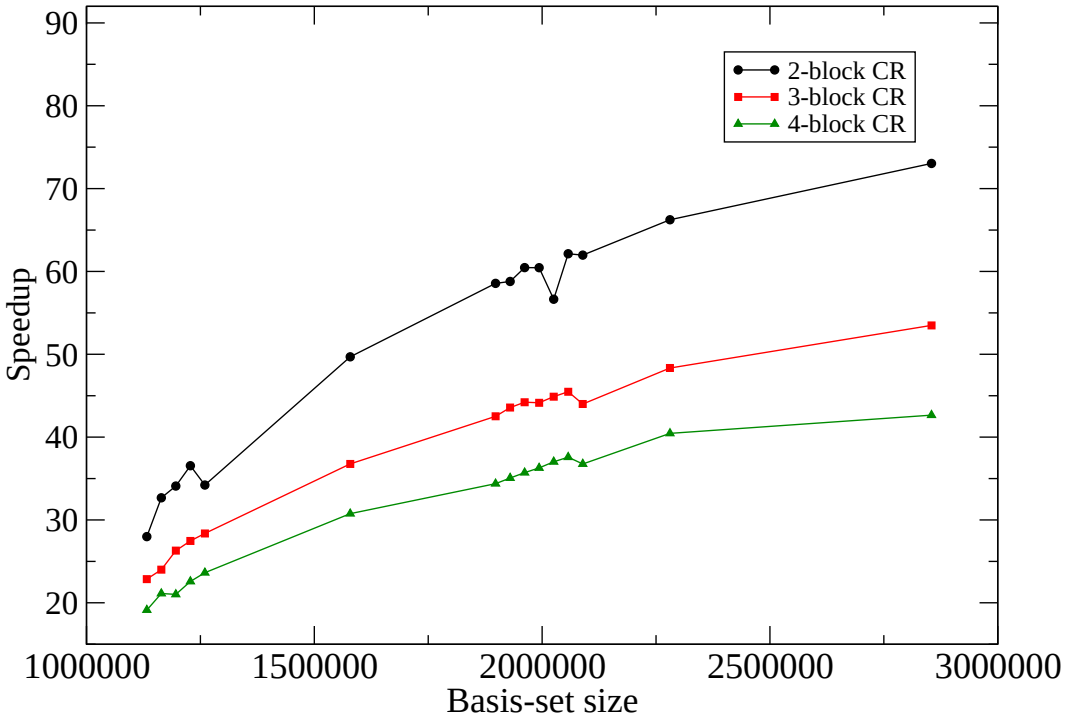

Figure 3.6: Performance characteristics of the CR kernal with varying values for p using basis-sets from SO3.

is especially suited to more difficult molecules. As the limitation of the vibrational basis-set on the line-list completion time is reduced. This may encourage the use of larger basis-sets that would improve aspects such as convergence as well as giving the line-list a larger frequency range.

It is worth comparing the ideal case where the entire dipole fits into memory for a large molecule like SO_3. Access to a K80 GPU was acquired with an identical setup to the M2090. The 12 GB of memory provided by the K80 gives us the ability to access how the non-blocking CR compares to the blocking CR. Figure 3.7 shows the speedup with the non-blocking CR kernal on the K80. Here we see that for the largest cases presented, we can achieve a 1000$\times$ speed-up compared to TROVE which is a substantial performance increase and a 10$\times$ increase from the 2-block CR. This is attributed to the reduced number of CR calls and the lack of stalling due to dipole matrix transfers. However the scaling of the algorithm is consistent for both with only a $\approx 50\%$ increase in execution time when the basis-set size is doubled compared to the CPU version which increases by $\approx 100\%$

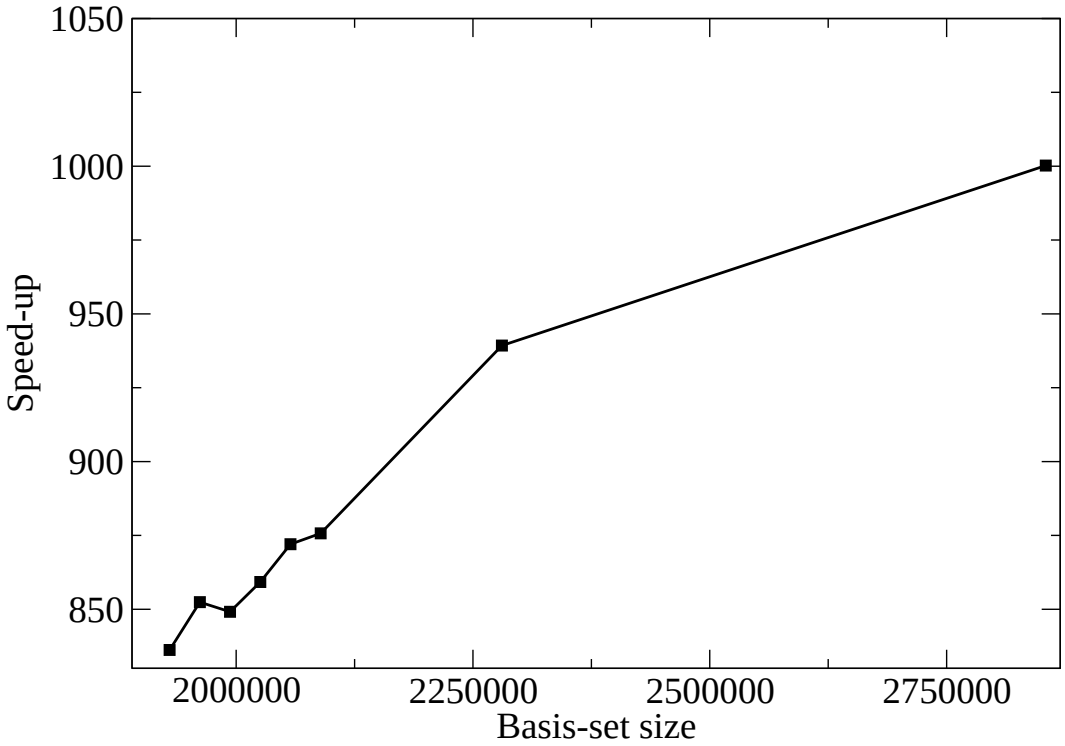

Figure 3.7: Speed-up achieved with the non-blocking CR on a K80 for SO_3

3.7 GAIN

GPU **A**ccelerated **IN**tensities (GAIN) is a set of functions that can compute transitions rapidly using the CR kernals and the cuBLAS[106] implementation of the dot product. The usage of the CR and non-blocking CR kernals are determined automatically by collecting GPU data and the dipole size at run-time. Additionally the code is asychronous and allows for the CPU to work whilst the GPU is computing. GAIN is also compatable with OpenMP, by passing in the total number of cores, it can perform up to 10 dot products in parallel on a single GPU. Multi-GPU configurations are also supported and will automatically detect distribute computation evenly across all cores reducing GPU core congestion in the dot-product step. Figure 3.8 shows the 'effective' speed-up of the dot product performance. As a single linestrength completion time doesn't change it is not a true 'algorithmic' speed-up. However with N GPUs we can compute $10N$ linestrengths simultaneously 'effectively' increasing throughput by $10N$.

The requirements to utilize GAIN are a dipole matrix $\mu^{\Delta K}_{\nu',\nu}$ and a K-block

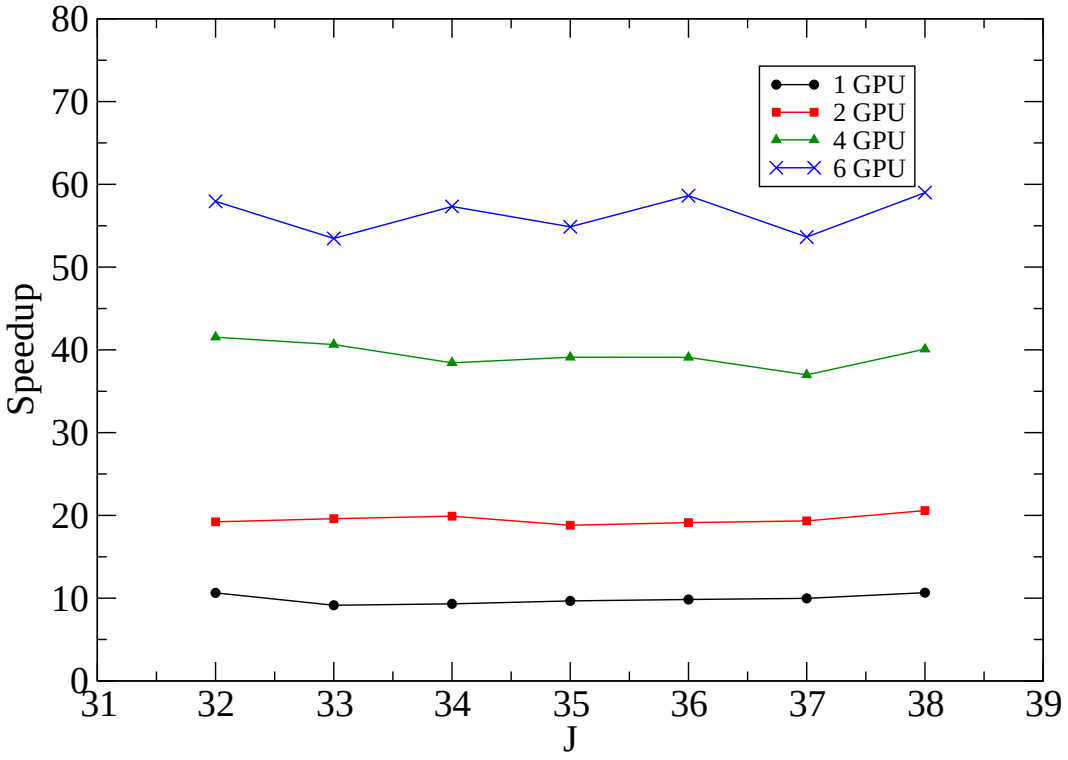

Figure 3.8: 'Effective' linestrength performance increase with varying GPU setups for GAIN.

basis set that can supply J, K and τ quantum numbers and ν indices. The `dipole_GPU.cu` source file contains the main kernals used to transfer basis sets, perform dipole blocking, utilize kernals and transfer results. Additionally there is code that can compute 3-j symbols by supplying J and K quantum numbers and transfer them to the GPU. The code is self-contained and automatically performs the required memory allocation on all available GPUs. The functions that call the kernals will automatically transfer the vector needed to complete the task and additional functions to return results. The functions and function naming are compatible with FORTRAN code and are seen as subroutines.

3.8 GAIN-MPI

GAIN-MPI is a hybrid OpenMP+MPI+CUDA C code that extends GAIN by increasing the total available memory and number of GPUs compared to single node setups. Its primary purpose is for the mass production of transitions. In its default form, it operates on TROVE's wavefunctions and utilizes the same input files

as TROVE, more specifically the intensity input blocks seen in Fig. 3.4 and keywords in Table 3.2. GAIN-MPI supports all symmetries that TROVE supports, this is because it utilizes TROVEs symmetry FORTRAN files. Therefore upgrading to support newer symmetries implemented in TROVE is a simple case of replacing the relevant FORTRAN files in GAIN.

Each process will read the input, load basis sets, states and perform the dipole splitting. After which the eigenvectors are distributed to a particular rank through the relation:

$$\text{Rank} = i \bmod N_{procs} \tag{3.23}$$

where i is a state counting number and N_{procs} is the total number of MPI processes. Eigenvectors are cached into RAM until the memory for each process is exhausted, after which all further eigenvectors access are from storage. Each process will read an initial state and determine whether is satisfies the filtering rules given by the input file and whether the state belongs to that particular rank. The process whom the state belongs to performs the necessary half linestrength calculations and broadcasts the results to all processes. Each process then loops through all states and performs the dot product on those that satisfy the filters and Eq. (3.23). Figure 3.9 describes this process. This 'striped' approach to distributing states and eigenvectors ensure that all ranks perform work within a given frequency range. Additionally, with enough MPI processes, I/O reads can be eliminated.

Originally GAIN-MPI was written to produce hot formaldehyde transitions as quickly as possible in order to make up for lost time from issues with the original hot line list discussed in Chapter 4.5. Due to the promising performance of the code, production of AYTY was temporarily halted and I moved to oversee the transitions of the higher priority phosphine hot line-list.

3.8.1 Phosphine

Phosphorus is the one of the most abundant chemically reactive volatile elements in a solar type system (with S, after H, C, N and O). Although phosphorus has considerably smaller cosmic abundances than H, O, C or N, it is predicted to have an

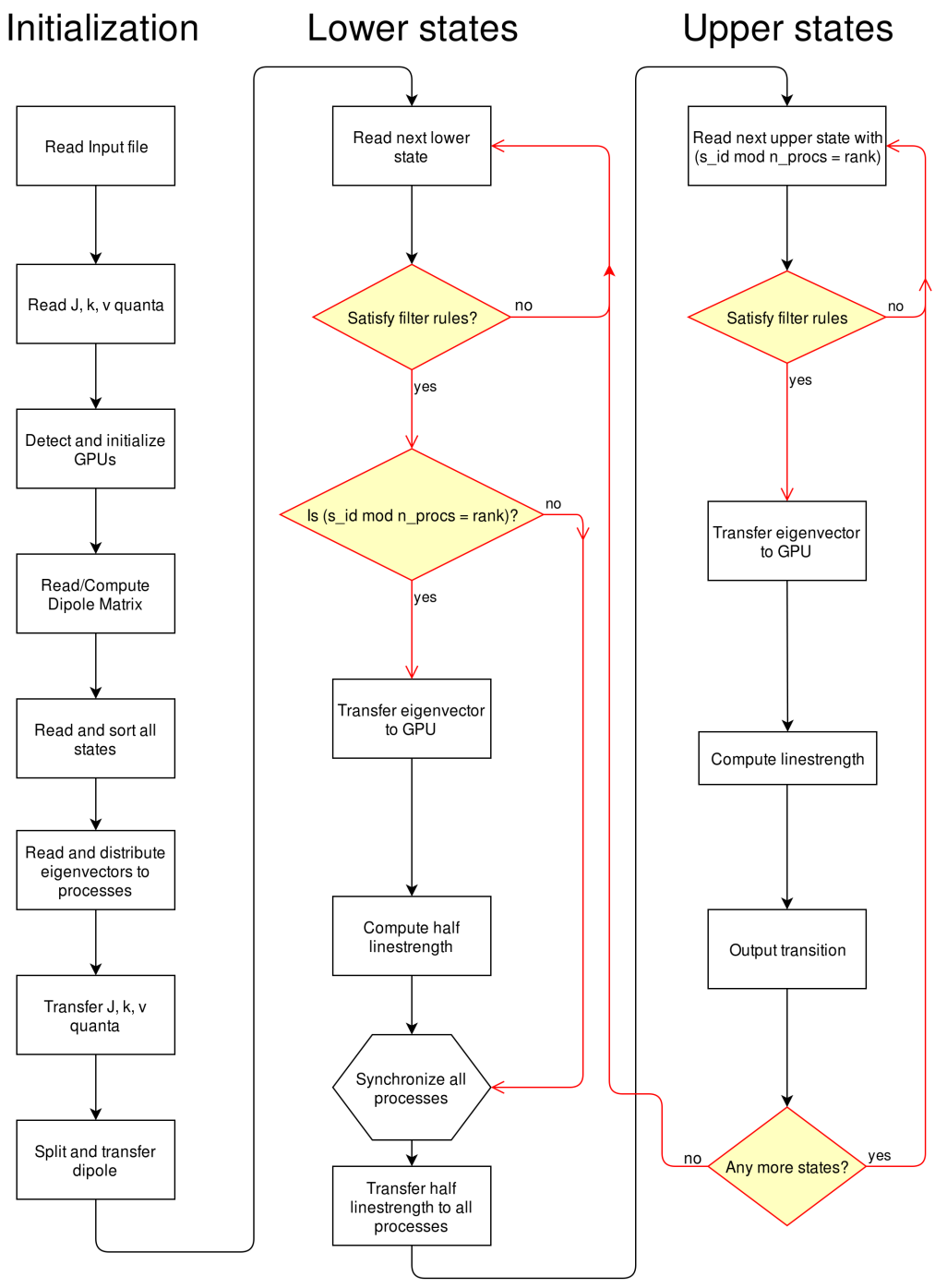

Figure 3.9: Flow chart depicting GAIN execution. *s_id* refers to the state running number, *n_procs* the total number of MPI processes and *rank* the current MPI processes rank

important role in atmospheric chemistry and dynamics. Phosphorus is not particularly common in the universe but it is ubiquitous and is important for most essential biochemical functions. Due to its role as biogenic particle, phosphorus could potentially be used in the search for extinct or extant life in other planets[107]. A large fraction of the existing phosphorus in various astronomical environments is expected to be found in the form of phosphine, or PH_3. A theoretical phosphine line list called SYT [108] was produced by the ExoMol project which was applicable to 300 K. This line list was unsuitable for hot atmosphere modelling so a high temperature line list was constructed which built upon the work done previously.

Phosphine belongs to the molecular symmetry group C_{3v}(M), spanning A_1, A_2 and E, where E is a two-dimensional irreducible representation. This means that the E symmetries has a matrix that is double the size and contains double the number of eigenvectors compared to the A symmetries. Whilst a relatively low maximum rotational excitation $J = 45$ is required for the line list, the extremely large $J = 0$ basis set, number of eigenvectors, wide frequency range at up to 10,000 cm^{-1} and high energy threshold of 18,000 cm^{-1} means that a staggering number of states and transitions are required to be computed and can be seen in Figure 3.10.

Even with the filters and symmetry reduction, J pairs such as $J = 20$ and $J = 21$ in E symmetry can take almost 1.5 months to complete the half a billion transitions required. To complete such transitions would require a budget of over 200,000 CPU hours which would constitute $\approx$20% of the ExoMol projects allowed computer time on the Darwin HPC center. The GAIN-MPI program was therefore a neccessity in quickly and cheaply producing these transitions.

The GAIN-MPI code performed spectacularly and Figure 3.11 shows completion time for PH_3 transitions at $J = 20$. Each process has 6 GB of memory and the required total to store all eigenvectors is $\approx$ 140 GB. When 30 processes are utilized, I/O is completely eliminated an completion time takes less than an hour. Overall, GAIN-MPI could do in less than a day what the CPU only version could complete in over a month. The GAIN code was used to compute the majority of the 17 billion transitions under time and under budget (Figure 3.12). This new line list is called

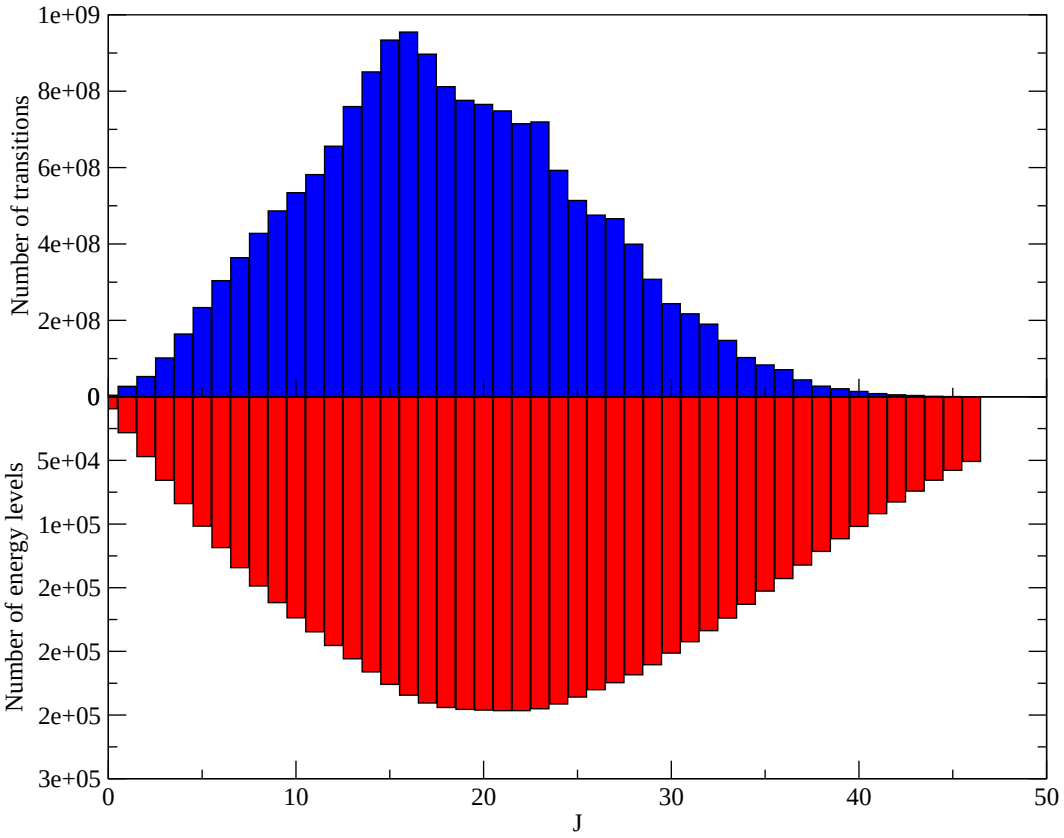

Figure 3.10: Number of energy levels in each rotational quantum number, J, and transitions between J and J + 1 in the SAlTY line list, summed over all symmetries.

SAlTY and has been successfully published [9]. GAIN has also been used to complete the transitions for AYTY H_2CO (Chapter 4), hot SO_3 [21], and the hot H_2O_2 (Chapter 5) line-list.

3.9 TROVE-GAIN

A new version of TROVE was also developed to implement a functional version of GAIN to encompass it as an overall nuclear motion solver suite. This code, dubbed TROVE-GAIN, gives TROVE the ability to directly utilise the highly efficient CR kernal and GPU dot products in order to evaluate transitions. Whilst the GPU kernals were relatively unchanged, the structure of the calculation had to be modified in order to be integrated into TROVE's computation pipeline. To overcome C++ and FORTRAN interoperability, wrapper functions were developed that facilitated the execution of kernals and the transfer of basis-sets, dipoles and vectors through stan-

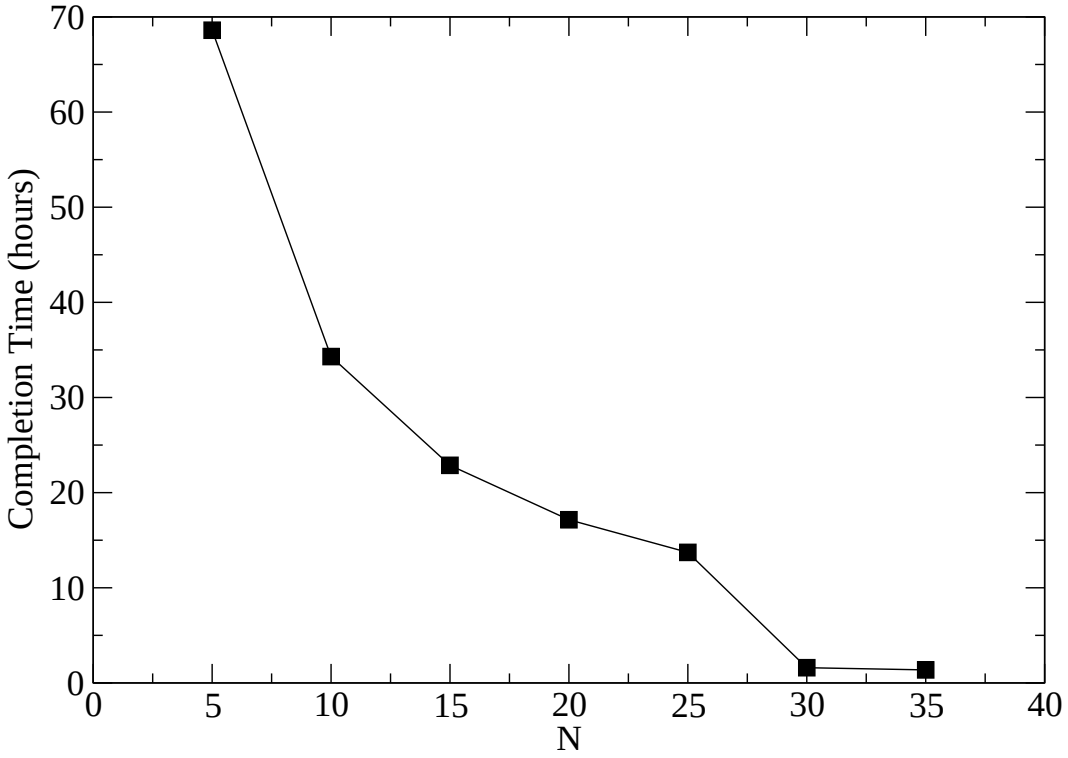

Figure 3.11: Completion time for $J = 21, 22$, E symmetry with $\approx$ 500,000,000 transitions for PH_3 against N MPI processes. At $N = 30$, I/O is effectively eliminated as all eigenvectors are stored in memory. The base time for the CPU-only completion is 1104.2 hours

dard data-types such as *doubles* and *integers* rather than C++ *structs* or FORTRAN *custom types*. The benefit of this is that the code is significantly more resistant to any change in TROVE's code. The only requirement is that it must interface with the simple wrapper functions in order to work. This code was extensively tested by computing the room temperature H_2O_2 linelist (Chapter 5.6) and was successful in producing the billions of transitions required. A paper has been submitted for publication [72]. Both the location of the source code for GAIN-MPI and the complete SAlTY line list can be found in Appendix A.

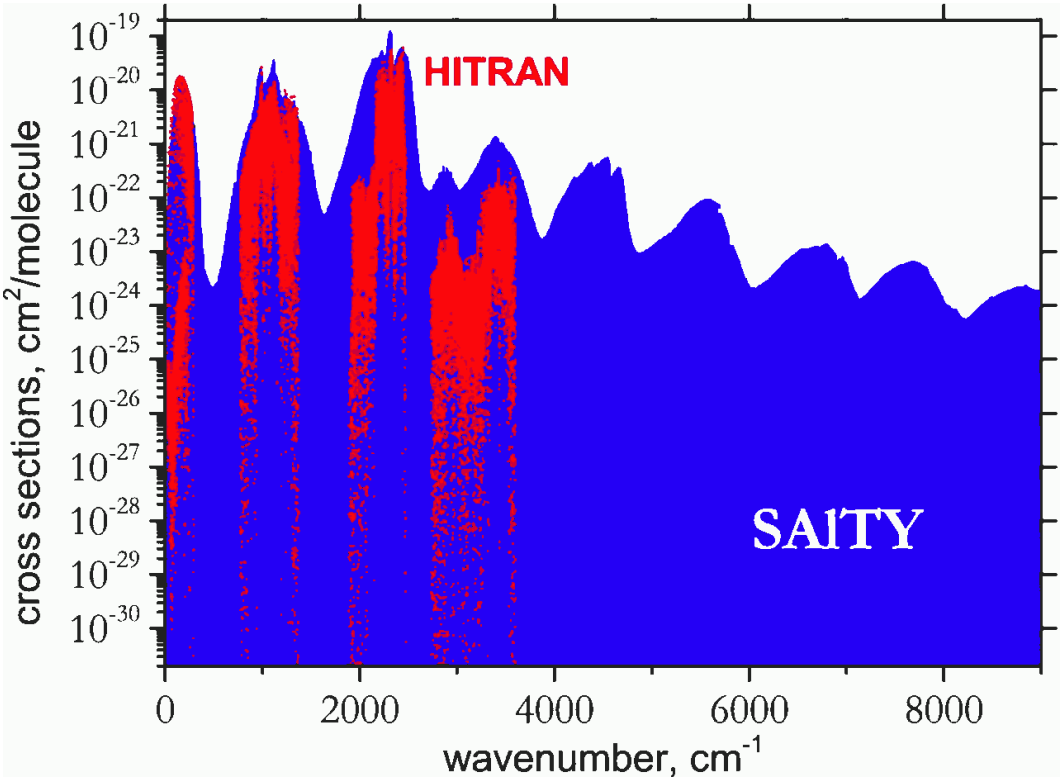

Figure 3.12: The hot PH_3 line list dubbed SAlTY[9] in its entirety compared to currently available experimental data at 1500 K. Red represents HITRAN transitions and the blue represents the SAlTY transitions. Note that HITRAN is not designed for usage at this temperature.

Chapter 4

Formaldehyde

4.1 Introduction

Formaldehyde, H_2CO, is a poisonous molecule in the aldehyde group. On Earth it plays a part in troposphere chemistry dynamics as the main source of OH via photo-dissociation and is formed from photo-oxidation in the atmosphere or through the incomplete burning of biomass [109]. Traces of formaldehyde have tentatively been detected in the Martian atmosphere [110] where it is believed to be derived from the oxidation of methane (CH_4) [111].

Formaldehyde was the first polyatomic molecule to be detected in the interstellar medium (ISM) [112] and is extremely abundant [113]. This has made it useful in investigating the isotope composition of carbon in the Galaxy [114]. The proposed mechanism of production is via the successive hydrogenation of CO [115] on icy grain mantles:

$$\begin{aligned} H + CO &\rightarrow HCO \\ H + HCO &\rightarrow H_2CO. \end{aligned} \tag{4.1}$$

Further hydrogenation produces methanol through an intermediate methyl radical $H+H_2CO \rightarrow CH_3O \rightarrow H+ CH_3O \rightarrow CH_3OH$. Common reactions include that with ammonia which produces amines [116] and polymerisation with other H_2CO molecules. As a result, formaldehyde is believed to be the major precursor for the formation of complex organic molecules in the ISM that include interstellar glyco-

laldehyde [117] and amino acids [116].

Formaldehyde's astrophysical relevance does not end in the ISM. Recently, it has been detected in comets [118], such as 103P/Hartley 2 [119], C/2007 N3[120] and Hale-Bopp [121], where it is thought to originate from the degredation of polyoxymethylene [122]. It is also present in protoplanetary discs around low mass young stars (Taurus-Auruga Class I/II) [123, 124, 125] as circumstellar ice with an abundance ratio of $\approx 2\%$ compared to the more ubiquitous water-ice.

Because of H_2CO's role as a precursor to complex organic molecules, it is considered a possible biomarker. The RNA world hypothesis suggests an early Earth with a CO_2, H_2O and N_2 rich atmosphere [126]. Illuminating this mix with ultraviolet (UV) radiation should lead to a large amount of formaldehyde being fixed in the atmosphere before being deposited into the prebiotic oceans [126]. Alternatively, the source of prebiotic chemical compounds may be derived without need of illuminating UV radiation via glancing icy body impacts [127]. Such impacts would produce shock-compression conditions that lead to formation of HCN molecules. These HCN molecules can be hydrolyzed to form formaldehyde and from there produce amino acids. Thus a planet rich in formaldehyde may indicate one undergoing the stages of pre-life.

Finally, formaldehyde masers [128, 129] are a reliable and proven tracer for high-density environments such as star-forming regions in galaxies due to its ubiquity and large number of long wavelength transitions [130]. Currently, there are 19 extragalactic sources [130] of these masers including IRAS 18566 + 0408, which is notable for detection of the first H_2CO maser flare [131]. Formaldehyde masers (and maser flares) have mostly been observed via the $1_{10} \rightarrow 1_{11}$ and $2_{11} - 2_{12}$ K-doublet transitions at 6.1 cm^{-1} and 2.2 cm^{-1}, respectively.

The wide-range of interactions in atmospheric, terrestrial, astrophysical and astrobiological phenomena makes formaldehyde a relevant molecule in the chemistry of exoplanets and their atmospheres. Therefore a complete, high-resolution, line list for H_2CO should provide an important aid for characterisation and modelling of formaldehyde. These considerations led formaldehyde to be included as

part of the ExoMol project [8].

4.1.1 Available data

High-resolution, room-temperature formaldehyde spectra have been well-studied in the laboratory [132, 133, 11, 134, 135, 13, 136, 137]; the early work was reviewed by Clouthier and Ramsay [138]. Currently, the major source of publicly available spectroscopic data on H_2CO is the HITRAN database [6] which has recently been updated to include long-wavelength data from the CDMS database [43]. The spectral regions covered in the database are 0 – 100 cm^{-1}, 1600 – 1800 cm^{-1} [139] and the 2500 – 3100 cm^{-1} [139] for lines with transition intensities greater than 10^{-29} cm/molecule for T=296 K. However, this compilation accounts for only 40 000 transitions extending up to $J = 64$ and covers only four of the six fundamental vibrational bands as well as the ground state rotational spectrum. This deficiency arises from an apparent lack of absolute intensities in the 100 – 1600 cm^{-1} range. Additional observed transitions are available [139] and include line positions [140, 141, 14], and intensities [140, 142, 15] of some of the fundamental bands and hot bands [143, 144, 145]. The incompleteness and low rotational excitations available in HITRAN limits the applicability of this data for temperatures above 300 K. The theoretical spectra presented in this thesis aims to provide a more complete and accurate picture of the spectrum of formaldehyde up to 10 000 cm^{-1} and for temperatures up to 1500 K. The line list should therefore be useful for modelling higher temperature environments as well as studies on non-LTE transitions such as those observed in masers.

Theoretically, electric dipole transition intensities of H_2CO were studied by Luckhaus et al. [146] and Carter et al. [147]; see also the review by Yurchenko [148]. Luckhaus et al. [146] used an *ab initio* MP2/6-311G** DMS to simulate the photoacoustic spectrum of high C-H stretching overtones of H_2CO. Carter et al. [147] generated an *ab initio* coupled-cluster CCSD(T)/aug-cc-pVTZ dipole moment surface (DMS) for H_2CO; they used an effective charges representation to compute (relative) rovibrational line intensities for H_2CO reproducing the HITRAN data [42] with reasonable agreement. Poulin et al. [136] computed an *ab initio* DMS using

Table 4.1: Vibrational modes and observed band centres in cm $^{-1}$ by Carter et al. [22].

Mode	Band Centres	Symmetry	Description
ν_1	2782.46	A_1	symmetric C-H stretching
ν_2	1746.01	A_1	C-O stretching
ν_3	1500.18	A_1	symmetric O-C-H bending
ν_4	1167.26	B_1	out-of-plane bending
ν_5	2843.33	B_2	asymmetric C-H stretching
ν_6	1249.10	B_2	asymmetric O-C-H bending

the QCISD/6-31111G(d,p) level of theory and presented it as an expansion.

Despite these works there is no comprehensive line list for formaldehyde available in the literature. The goal of this work is to bridge this gap. An initial potential energy surface (PES) obtained 'spectroscopically' by Yachmenev et al. [97] and a new *ab initio* dipole moment surface (DMS) for formaldehyde and generate an extensive line list for $H_2{}^{12}C^{16}O$ applicable for the temperatures up to $T = 1500$ K. In the following, H_2CO and formaldehyde will refer to the main isotopologue $H_2{}^{12}C^{16}O$.

H_2CO is a near-prolate asymmetric top molecule that belongs to the C_{2v} molecular symmetry group [20]. The group has four irreducible representations A_1, A_2, B_1 and B_2. Once the H atom nuclear spin is taken into account the 'para' A representations are singly degenerate $g_{ns} = 1$ and the 'ortho' B representations are triply degenerate $g_{ns} = 3$. As H_2CO has four atoms, it has six vibrational modes; Table 4.1 shows the vibrational modes and their corresponding symmetries, band centers and descriptions. Coriolis interactions occur strongly between the ν_4 and ν_6 modes, and weakly between the ν_3 and ν_4 modes [149] which couples their energy levels and wavefunctions. This manifests itself in the ν_3, ν_4 and ν_6 mode interaction as overlapping bands which make these three bands difficult to distinguish from each other.

4.2 Potential Energy Surface

The potential energy surface was produced by Yachmenev et al. [97], so only a brief summary is presented here. An *ab initio* six-dimensional surface was computed with 30,840 geometries with energies up to 44,000 cm^{-1} using MOLPRO [67].

The CCSD(T) level of theory was utilized in the quantum-chemical calculations with the frozen-core approximation and augmented correlation-consistent aug-cc-pVQZ basis. The potential is represented analytically as a Taylor expansion around the out-of-plane motion minimum energy path (MEP):

$$V(r_1,r_2,r_3,\theta_1,\theta_2,\tau) = \sum_{ijklmn} f_{ijklmn}(1-e^{-\Delta r_1})^i(1-e^{-\Delta r_2})^j(1-e^{-\Delta r_3})^k(\Delta\theta_1)^l(\Delta\theta_2)^m(1+\cos\tau)^n \tag{4.2}$$

with minimum expansion order of $i+j+k+l+m+n=6$. The three stretching co-ordinates are represented as:

$$\begin{aligned} \Delta r_1 &= r_{CO} - r_{CO}^{ref}(\tau) \\ \Delta r_2 &= r_{CH_1} - r_{CH_1}^{ref}(\tau) \\ \Delta r_3 &= r_{CH_2} - r_{CH_2}^{ref}(\tau) \end{aligned} \tag{4.3}$$

and the two bending co-ordinates are represented as:

$$\begin{aligned} \Delta\theta_1 &= \theta_{OCH_1} - \theta_{OCH_1}^{ref}(\tau) \\ \Delta\theta_2 &= \theta_{OCH_2} - \theta_{OCH_2}^{ref}(\tau) \end{aligned} \tag{4.4}$$

where r_{CO}, r_{CH_1} and r_{CH_2} are bond lengths, θ_{OCH_1} and θ_{OCH_2} are bond angles and τ is the out-of-plane bending angle. Figure 4.1 visually describes these co-oridnates. Here the MEP reference geometries were determined *ab initio* using the CCSD(T)/aug-cc-pVQZ level of theory and then expanded around $1+\cos\tau$:

$$\begin{aligned} r_{CO}^{ref}(\tau) &= \sum_{n=0}^{4} a_n^{CO}(1+\cos\tau)^n \\ r_{CH}^{ref}(\tau) &= \sum_{n=0}^{4} a_n^{CH}(1+\cos\tau)^n \\ \theta_{OCH}^{ref}(\tau) &= \sum_{n=0}^{4} a_n^{OCH}(1+\cos\tau)^n \end{aligned} \tag{4.5}$$

Where a_n are reference geometry expansion parameters. With the MEP defined, the potential expansion parameters (f_{ijklmn}) are computed using a least squares

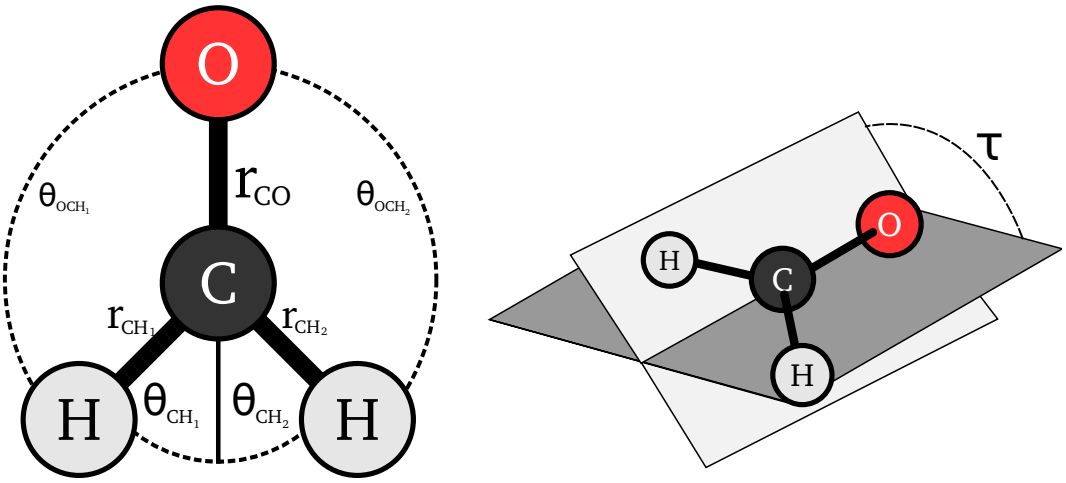

Figure 4.1: The internal co-ordinates of H_2CO.

fit to computed *ab initio* energies with weighting as suggested by Partridge and Schwenke [150]:

$$w_i = \frac{\tanh[-0.0006(V_i - 16000)] + a}{b \times V_i^{(w)}} \tag{4.6}$$

where $a = 1.002002002$, $b = 2.002002002$ and $V_i^{(w)} = max(16000, V_i)$ where V_i is the *ab initio* energy at the i-th geometry.

This *ab initio* PES has an rms error for all term values below 7200 cm^{-1} of 5.1 cm^{-1} . A refinement process was carried out via the methodology outlined by Yurchenko et al. [98] and Chapter 2.3.5. The input data for the refinement contains 319 experimental energies for $J = 0, 1, 2$ and $J = 5$ with fitting weights of $w_i = 100$, $w_i = 1.0$ and $w_i = 0.1$ assigned to purely rotational states, excited vibrational states and data from Bouwens et al. [151], respectively. The rms error against experimental energy levels with $J \leq 5$ of this semi-empirical PES, called H2CO-2011, is 0.04 cm^{-1}.

4.3 Variational computation

The calculations outlined here were performed with TROVE[30] (See Chapter 2.3) For H_2CO, a kinetic expansion order of 6 and a potential energy expansion order of 8 was chosen. The internal co-ordinate system used in the kinetic energy expansion

Table 4.2: Linearized internal co-ordinates and equilibrium geometry, bond lengths are in angstroms and bond angles are in degrees

ξ_i^ℓ	ξ_i	Equilibrium (ξ_i^e)
r_1^ℓ	r_{CO}	1.20367
r_2^ℓ	r_{CH_1}	1.10290
r_3^ℓ	r_{CH_2}	1.10290
θ_1^ℓ	θ_{OCH_1}	121.7810
θ_2^ℓ	θ_{OCH_2}	121.7810

Table 4.3: Primitive basis functions, co-ordinate borders, number of grid points for Numerov-Cooley integration and number of solutions

Basis func.	Borders	No. grid points	No. soln
$\phi_{n_1}(r_1^\ell)$	-0.35–1.80	1000	16
$\phi_{n_2}(r_2^\ell)$	-0.50–1.00	1000	8
$\phi_{n_3}(r_3^\ell)$	-0.50–1.00	1000	8
$\phi_{n_4}(\theta_1^\ell)$	-1.20–1.20	1000	14
$\phi_{n_4}(\theta_1^\ell)$	-1.20–1.20	1000	14
$\phi_{n_6}(\tau)$	-120.0°–120.0°	2000	16

for H_2CO are linearized co-ordinates [30, 20] of the form:

$$\xi_i^\ell = \xi_i - \xi^e \tag{4.7}$$

where the ξ^e is the equilibrum geometry for a co-ordinate ξ_i defined in Table 4.2. The out of plane bending co-ordinate is simply τ. The primitive basis-set consists of the six functions: $\phi_{n_1}(r_1^\ell)$, $\phi_{n_2}(r_2^\ell)$, $\phi_{n_3}(r_3^\ell)$, $\phi_{n_4}(\theta_1^\ell)$, $\phi_{n_5}(\theta_2^\ell)$, and $\phi_{n_6}(\tau)$ which are obtained by solving the one-dimensional Schrodinger equation in Eq. (2.49). The paramaters used in computing these functions are given in Table 4.3.

The direct product of the 1D basis functions is contracted using the polyad condition:

$$P = 2(n_2+n_3)+n_1+n_4+n_5+n_6 \leq P_{\max}, \tag{4.8}$$

where $P_{\max} = 16$. The relative simplicity of the molecule means that this gives well-converged results. In terms of the normal mode quantum numbers v_i reads

$$P = 2(v_1+v_5)+v_2+v_3+v_4+v_6 \leq P_{\max}. \tag{4.9}$$

This polyad rule is based on the approximate relationship between the H_2CO fundamental frequencies (see Table 4.1):

$$\nu_1 \approx \nu_5 \approx 2\nu_2 \approx 2\nu_3 \approx 2\nu_4 \approx 2\nu_6. \tag{4.10}$$

The vibrational basis set is further optimized by solving Eq. (2.52) for $\{r_1\}$, $\{r_2, r_3\}$, $\{\theta_1, \theta_2\}$, and $\{\tau\}$ to produce four sets of wavefunctions $\Phi_{n_1}^{(1)}(r_1)$, $\Phi_{n_2,n_3}^{(2,3)}(r_2, r_3)$, $\Phi_{n_4,n_5}^{(4,5)}(\theta_1, \theta_2)$, and $\Phi_{n_6}^{(6)}(\tau)$, respectively. Eq. (2.54) is then solved and the solutions are symmetrized according to the C_{2v}(M) molecular symmetry group. The resulting eigenfunctions $\Psi_i^{J=0,\Gamma}$ obtained for each C_{2v}(M) symmetry $\Gamma = A_1, A_2, B_1$ and B_2 and form the wavefunction representation given in Eq. (2.55). The latter is defined according with the Eckart conditions [152] and is oriented approximately along the CO bond. In C_{2v}(M) symmetry, K and τ_{rot} correlate with the customary K_a and K_c rotational quantum numbers as

$$K = K_a, \quad \tau_{\text{rot}} = \text{mod}(|K_a - K_c|, 2). \tag{4.11}$$

The $J = 0$ basis set is truncated using the energy threshold of $E_{\text{max}}^{J=0} = hc\,15\,400$ cm^{-1} and thus consists of 868, 570, 628, and 791 functions for the A_1, A_2, B_1 and B_2 symmetries, respectively. The largest J computed was $J = 70$ which required the diagonalisation of matrices in the order of $\approx 122\,000$ for eigenvalues and eigenvectors. The diagonalization was performed using LAPACK and SCALAPACK.

4.4 Dipole moment surface and intensities

Intensity calculations require a high-quality electric DMS. An *ab initio* DMS was computed at the CCSD(T)/aug-cc-pVQZ level of theory in the frozen-core approximation using CFOUR [68]. Three symmetry-adapted projections of the dipole moment Cartesian components, μ_{A_1}, μ_{B_1}, and μ_{B_2}, are given in the analytical representations with each component expanded in Taylor series (185 parameters in total) in terms of internal coordinates around the equilibrium configuration using the form developed by Yachmenev et al. [153] to represent the dipole moment of

H_2CS. These parameters reproduce the *ab initio* dipole moment values of the μ_{A_1}, μ_{B_1}, and μ_{B_2} components with rms errors of 0.0002 Debye for each component. The equilibrium value of the dipole moment is 2.3778 D (at r^e_{CO} = 1.2033742 Å, r^e_{CH} = 1.10377 Å, θ^e_{OCH} = 121.844°), which can be compared to the experimental value of the ground vibrational state dipole moment of μ=2.3321(5) D measured by Fabricant et al. [154].

The eigenvectors, obtained by diagonalization, are used in conjunction with the DMS to compute the required linestrengths (and from that the Einstein-*A* coefficients and absolute intensities) via the methodology described in Chapter 3. The transitions must satisfy Eq. (3.5). The symmetry of the dipole moment operator $\Gamma(\mu_A)$ is A_2. Therefore Eq. (3.4) is non-zero for the symmetries:

$$A_1 \leftrightarrow A_2, \;\; B_1 \leftrightarrow B_2. \tag{4.12}$$

The transitions were computed using the energy limits *hc* 6 000 and *hc* 15 600 cm^{-1} for the lower and upper states, respectively with a maximum transition wavenumber of 9 600 cm^{-1}.

4.5 AYTY-0

AYTY-0 was the name retrospectively given to the first line-list produced using the H2CO-2011 PES. This line list was developed over the course of a year and contained 5 million energy states and around 3 billion transitions. Comparing band intensities of the fundamentals highlighted in Table 4.4 demonstrated good agreement with the ν_1, ν_2, ν_3, ν_5 and ν_6 bands. The weak ν_4 band however was an order of magnitude too strong. Due to the lack of experimental intensities in this region, it was difficult to determine whether it was due to a problem with the linelist or of some experimental difficulty with this particular region. Cross-sections of this region are available [10, 11]. Figures 4.2(a) and 4.2(b) show simulated cross-sections of the AYTY-0 linelist with HWHM at 0.112 cm^{-1} and 1.185 cm^{-1} respectively compared against the available experimental data. It is worth noting that the data from Figure 4.2(b) was extracted from Fig. 3 of [11] and scaled by a reasonable factor to

Table 4.4: Band Intensities in 10^{-17} cm $^{-1}$ /(molecule cm $^{-2}$)

Band	Obs[a]	AYTY-0	(O-T)/O (%)
ν_1	1.008	0.993	1.5
ν_2	1.219	1.527	-25.2
ν_3[b]	0.184	0.058	68.5
ν_4[b]	0.069	0.846	1108.9
ν_5	1.120	1.243	11
ν_6	0.173	0.202	16.7

[a] Perrin et al. [139] [b] Perrin et al. [140]

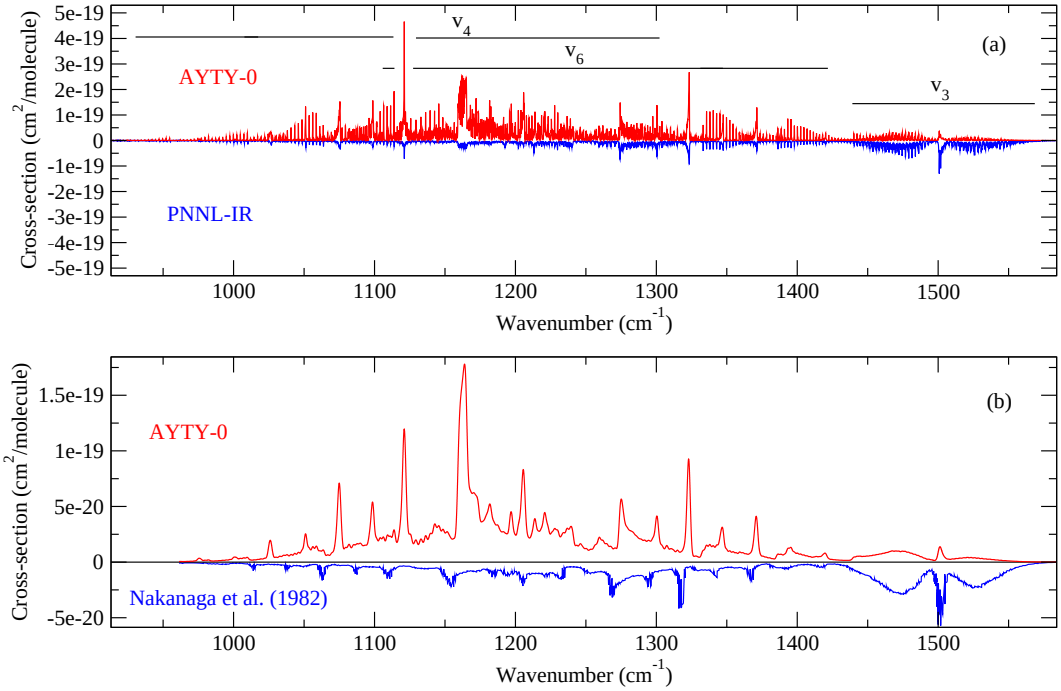

Figure 4.2: Cross-sections of the ν_3, ν_4 and ν_6 region for H_2CO against experimental data. (a) is from PNNL-IR [10] (b) is from [11]. Indicated in the plot is the location and extent (indicated by the line under the band label) of each band. Here the overlap between the ν_4 and ν_6 band and their unusually strong intensities compared to the ν_3 band can be seen.

highlight structure and relative intensities.

From both comparisons, we see that the general features of the band are reproduced well and overall structure and position are in good agreement. The problem arises in both the absolute and relative intensities of the ν_3 and ν_6 bands compared to the ν_4 band. Both experimental data sets suggest that the ν_3 band should be stronger than either the ν_4 and ν_6 band and the combined ν_4 and ν_6 integrated cross-sections

around 72% stronger than the ν_3 which is in agreement with Table 4.7 at 68%. The AYTY-0 line list suggests 1806% and in general seems to be an order of magnitude off. This effect was not observed with any other fundamental band.

Initially the blame fell upon the DMS. It was then discovered that calculations based on the *ab initio* PES did not display this effect. The aggressive refinement process PES used in H2CO-2011 changed the transition moments drastically for ν_4. The reason for this is unknown but may be due to the potential surface becoming distorted unphysically.

As it is not known how many transitions were affected by this, it was decided that the PES would be re-refined with careful observation of the transition moments. A new PES dubbed H2CO-2014 was produced with comparable accuracy and was used in the final hot-linelist AYTY.

4.6 AYTY

The AYTY-0 line-list was produced over the course of a year, during that time two major developments occured. First, the development of an external MPI diagonalizer to diagonalize matrices produced by TROVE using SCALAPACK. Second, was the development of the GAIN code described in Chapter 3. The former meant that the extremely large matrices with sizes in the order ≈100,000 could be diagonalized efficiently and were no longer constrained to the available memory on a single computer. With this the basis-set truncation $E_{\max}^{J=0}$ was raised to 18,000 cm^{-1} making the average size of basis-set 1,920 and maximum matrix size at $J = 70$ ≈300,000. The latter development afforded us the ability to raise the energy limits to hc 8 000 and hc 18 000 cm^{-1} for the lower and upper states, respectively increasing the maximum transition wavenumber to 10 000 cm^{-1}.

After completing the new PES, a new line list was produced in only 4 months. This new line-list labelled AYTY, now offers over 10 million states and around 14 billion transitions.

The transitions are sorted in increasing transition wavenumber and then converted into the ExoMol format [155]. Spectra at arbitrary temperatures can be com-

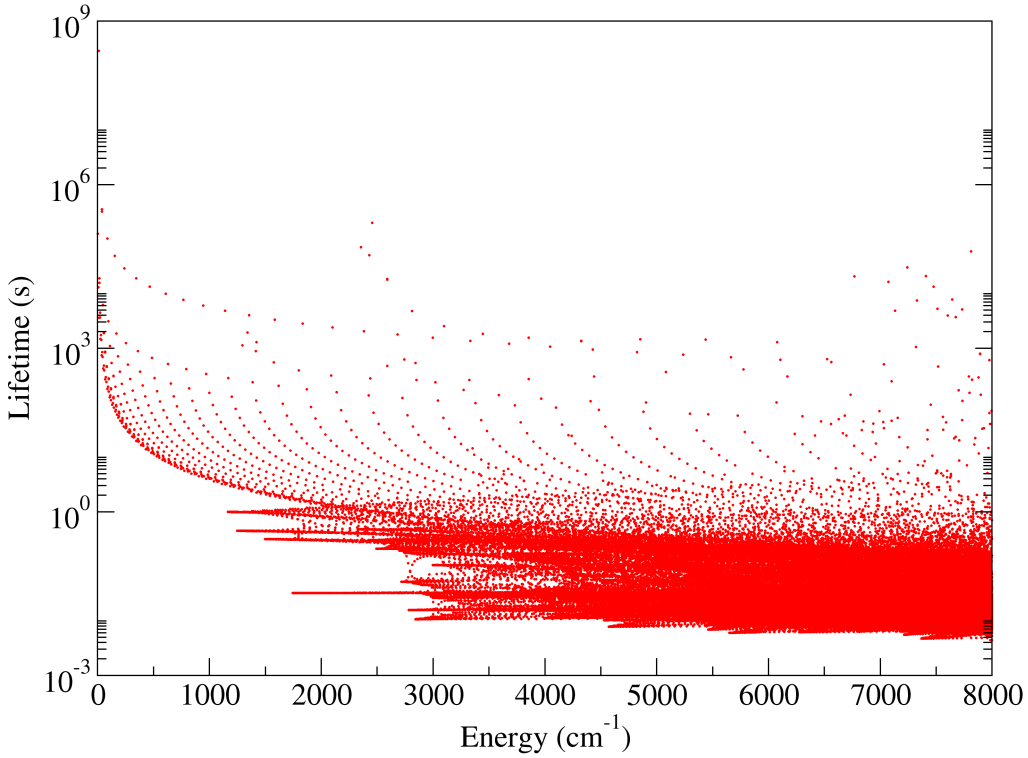

Figure 4.3: Radiative lifetimes computed for H_2CO using the method described by Tennyson et al. [12]

puted using the Einstein-A coefficients from the transition files. The theoretical error is estimated by the fitting rms deviation as 0.18 cm^{-1}. This means transition frequencies and energy levels should be reliable to about 0.2 cm^{-1} for low-lying levels and levels for vibrational states for which there is no available laboratory data. The pure rotational transitions are much more accurate than this. Radiative lifetimes were also computed with the methodology described by Tennyson et al. [12] and are presented in Figure 4.3.

The completeness of the line list as a function of temperature can be determined by checking the convergence of the temperature-dependent partition function Q given in Eq. (3.9), which is computed via explicit summation [46] of the 10.3 million energy levels available. As T increases, a greater proportion of these states are required as their contribution towards Q becomes more important. Figure 4.4 shows computed partition functions as a function of the maximum J value

($J_{\max}$) used in the calculation. As $J_{\max}$ increases, each J contributes progressively less until convergence is reached. The partition function at $T = 296$ K converges to better than 1% at $J \approx 34$ with the limit of $Q = 2\,844.621$ at $J = 58$. For $= 1\,500$ K, it converges to about 0.005% at $J = 70$ with a Q value of 130 190.25. These partition functions can be used to evaluate the effect of lower energy state threshold of 8000 cm^{-1} on the completeness of the line list by comparing $Q_{\rm limit}$, which sums energies up to this threshold, with the full partition sum. Figure 4.5 shows that the two partition functions are essentially the same up to 800 K and that $Q_{\rm limit}$ is 92.3 % of Q at $T = 1500$ K. Therefore T=1500 K is a 'soft' limit to the applicability of the line list. Use of the line list at higher temperatures will lead to the progressive loss of opacity although the ratio $Q_{\rm limit}/Q$ can be used to estimate the proportion of this missing contribution [52].

Table 4.6 compares AYTY partition functions with those from CDMS [43] and those used in HITRAN [23]. At temperatures $T \leq 300$ K we agree to better than 1% with CDMS and HITRAN. At 500 K the difference with CDMS is much higher at 8.9%, due to the explicit sum running over a much larger number levels, but agreement with HITRAN remains good. There are bigger differences at higher temperatures: at 1500 K the partition function is lower by about 1.2% and at 3000 K by 9.7%. This may be caused by the lack of the high energy contributions due to the J cut-off of used in the line list, see Sousa-Silva et al. [47] and Neale and Tennyson [156] for a discussion of the importance of contributions from the excited ro-vibrational states up to the dissociation. The full partition function evaluated on a 1 K grid is given on the ExoMol website.

We use the analytical representation suggested by Vidler and Tennyson [46] as given by

$$\log_{10} Q(T) = \sum_{n=0}^{8} a_n [\log_{10} T]^n . \tag{4.13}$$

The expansion parameters given in Table 4.5 reproduce the partition function better than 0.3% for temperatures ranging up to 3000 K.

The dependence of the cross-sections on temperature is illustrated in Figure 4.6, the features in the simulated spectra become smoother as the temperature in-

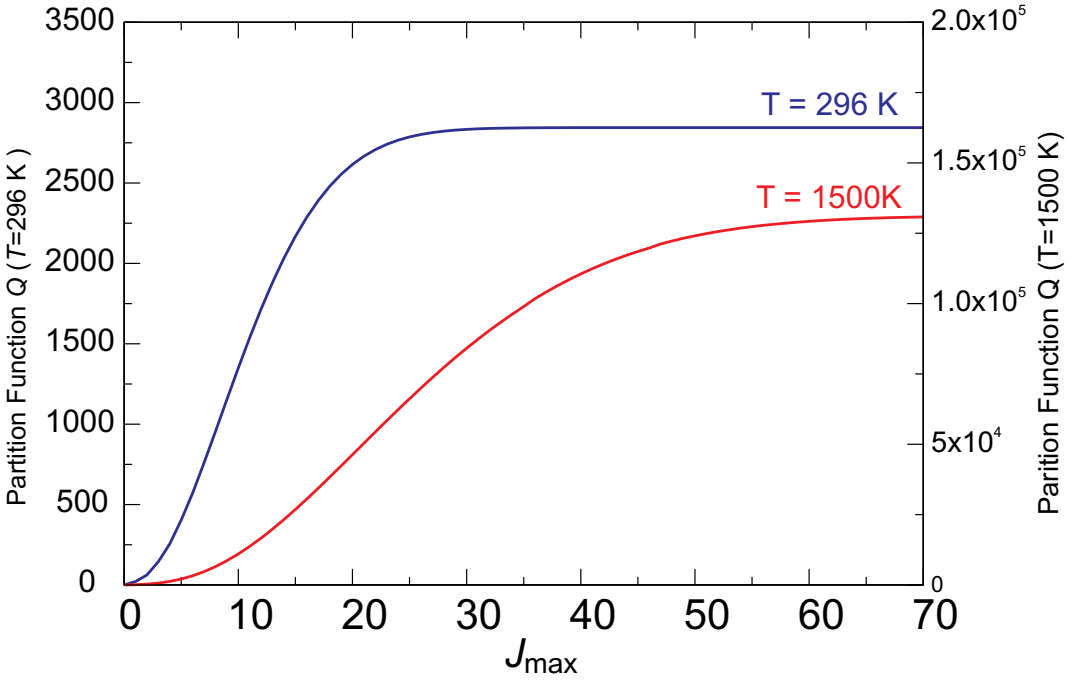

Figure 4.4: Partition functions of H_2CO at two temperatures as a function of inclusion of rotational states: all J up to J_{max} for $T = 296$ K (left hand scale) and $T = 1500$ K (right hand scale).

Table 4.5: Parameters used to represent the H_2CO partition function, see Eq. (4.13), which is valid for temperatures up to 3000 K.

Parameter	Value
a_0	1.12789807683
a_1	−5.35067939866
a_2	10.33684323700
a_3	−4.92187455147
a_4	−2.28234089365
a_5	3.61122821799
a_6	−1.64174365325
a_7	0.33727543206
a_8	−0.02654223136

creases. This is a result of the vibrationally excited states becoming more populated and the increasing width of the rotational envelope with temperature. Figure 4.7 shows a simulated $T = 296$ K spectrum computed from the line list against the available laboratory absorption spectra up to 10 000 cm^{-1}.The logarithmic scale used shows the density of transitions in the AYTY line list and reveals the significant gaps and limitations in the HITRAN 2012 database. Comparing specific

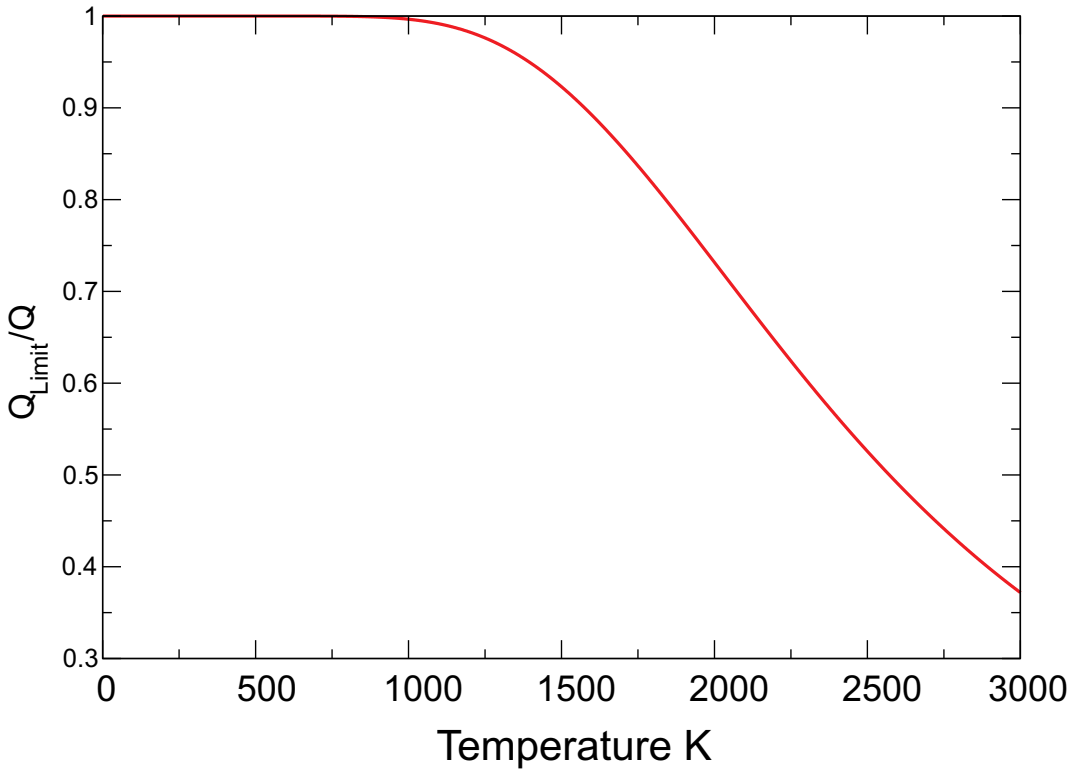

Figure 4.5: Plot of Q_{limit}/Q against temperature where Q_{limit} is the partition function computed using only energy levels below the lower state threshold of 8000 cm^{-1}.

Table 4.6: Comparisons of H_2CO partition functions as a function of temperature for this work, CDMS [43] and those used in HITRAN [23].

T / K	AYTY	CDMS	HITRAN
2.725	2.0165	2.0166	
5.000	4.4833	4.4832	
9.375	13.801	13.8008	
18.75	44.6835	44.6812	
37.5	128.6581	128.6492	
75	361.7053	361.7195	362.07
150	1019.9549	1019.9706	1020.47
225	1874.4679	1872.6221	1875.67
300	2904.1778	2883.0163	2906.32
500	6760.2315	6208.3442	6760.99
1500	128635.40		130190.25
3000	2741283.3		3038800.0

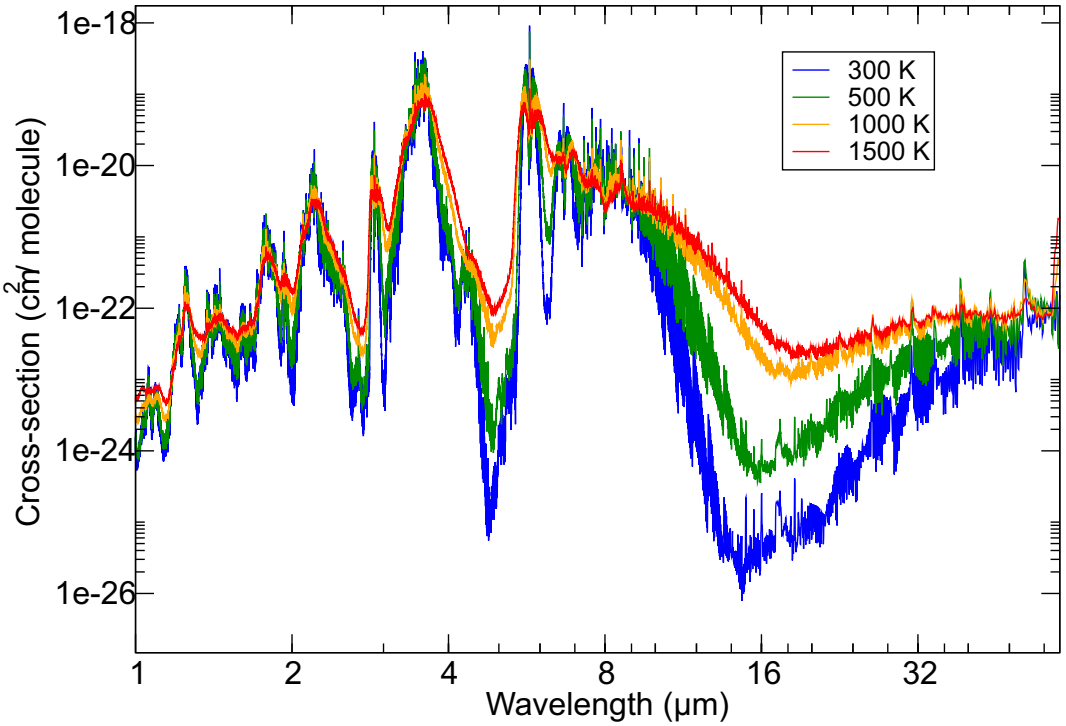

Figure 4.6: Cross-sections of the entire AYTY line list as a function of temperature: The curves in the 16 μm region increase in opacity with increasing temperature.

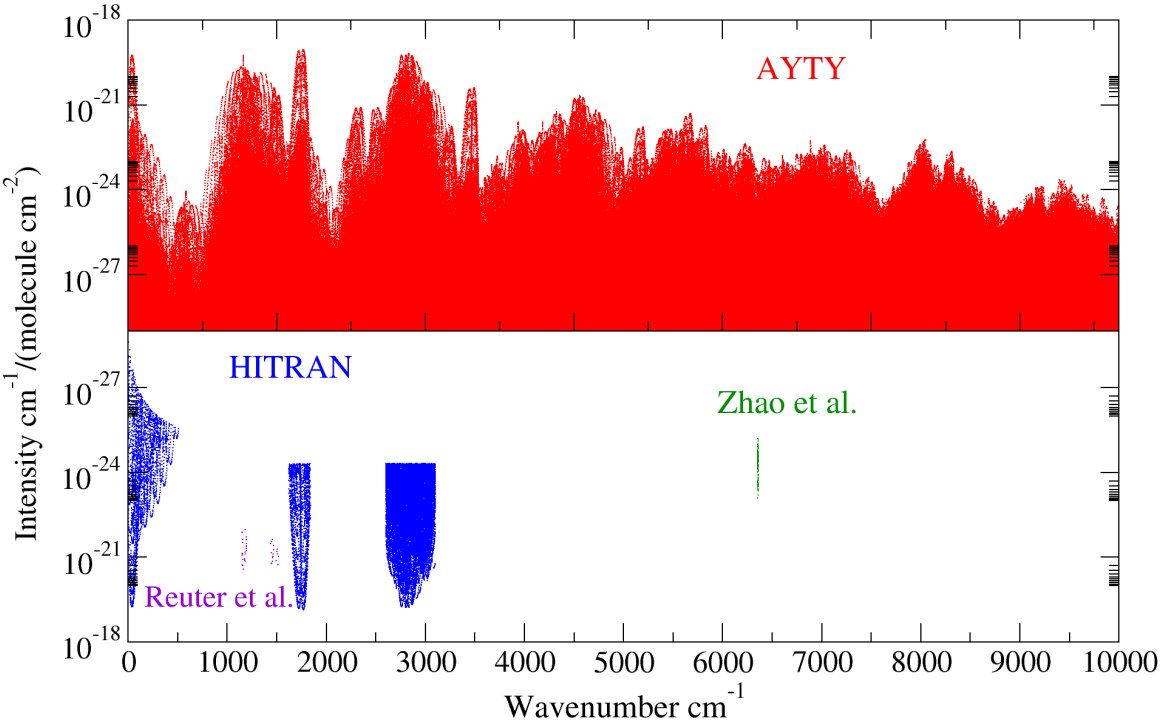

Figure 4.7: Overview of the AYTY synthetic spectrum at T =296 K against HITRAN [6], Reuter et al. [13] and Zhao et al. [14].

regions, our line list accurately replicates both the line positions and intensities of the three available bands, as illustrated in detail in Fig. 4.8. Additional lines are present as the computed spectra contains all possible transitions within the region including hot bands. Fig. 4.8d and Table 4.8 show agreement with the line positions and absolute intensities from Reuter et al. [13] with an rms deviation of 0.099 cm^{-1} for the line positions. There are some limitations with our line list. Higher J transitions at around the $J > 50$ range begin to show a slight drift of ≈ 0.3 cm^{-1} in predicted line position; this does not occur for the rotational band. In practice, errors in the ro-vibrational energy levels grow with K (as opposed to J); the discrepancies in transition frequencies become more pronounced in $|K' - K''| = 1$ transitions than those that involve the same K ($K' = K''$). This can be seen in the lack of drift in the pure rotational band as it is mostly comprised of $K' = K''$ transitions due to both ground and excited states being of A_1 symmetry. B_1 and B_2 vibrational bands however are mostly comprised of $|K' - K''| = 1$ transitions which makes their errors more sensitive to the quality of the model.

Computing band intensities requires simulating spectra at a chosen temperature and accumulating all transitions that correspond to the specific band. Table 4.7 highlights the AYTY band intensities against those available in the literature. Each band intensity required spectra simulated to the parameters used by each reference. In general, AYTY agrees well with all band intensities which confirms that the H2CO-2011 PES was responsible for the initial issues with the ν_4 band intensities as the DMS is unchanged from AYTY-0. Overall all bands are more intense in AYTY compared to the experimental. This may be due to the fact that AYTY sums over orders of magnitude more lines in a given band compared to synthetic spectra from effective Hamiltonians. Table 4.7 also shows the total band intensity for the 3.5 μm region compared to that by [133, 11] and HITRAN. The value is 13 % stronger than HITRAN, (matches the discrepancy for the ν_1 and ν_5 bands in Table 4.7), 18 % stronger than Nakanaga et al. [11] and 40 % stronger than that by Brown et al. [133]. Absolute intensities and bands not available in the HITRAN database or literature can be evaluated against cross-sections. For the ν_3, ν_4 and ν_6 bands,

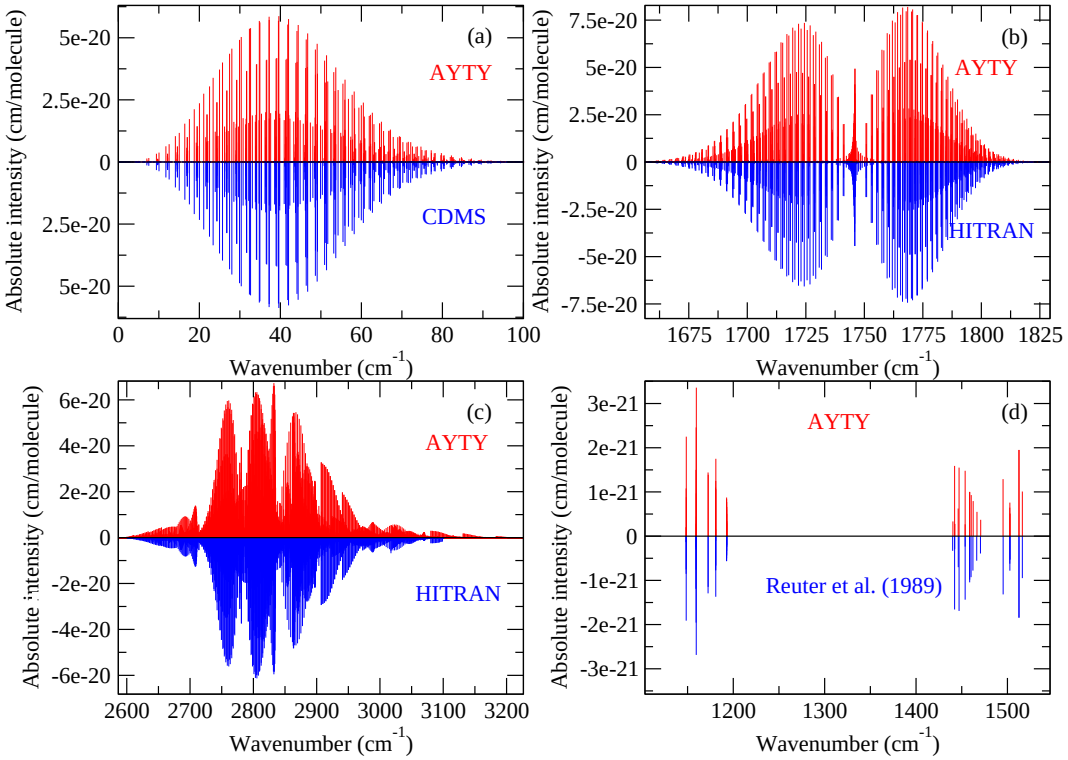

Figure 4.8: The fundamental bands compared to currently available experimental intensities [6, 13, 43] at $T = 296$ K. (a) Rotational Band (b) ν_2 (c) ν_1 and ν_5 (d) ν_3, ν_4 and ν_6 .

further evaluation of these bands can be made against cross-sections available from the PNNL-IR database [10] and Nakanaga et al. [11] using a Gaussian profile with a HWHM (half-width-half-maximum) of 1.1849 cm^{-1} and 0.1120 cm^{-1}, determined from their respective experimental profiles. Figure 4.9a compares the AYTY line list with a spectrum extracted from Fig. 3 of Nakanaga et al. [11] and scaled to match the AYTY line list. Good agreement is seen in both structure and position in the band with a slight drift occurring as an artifact from the extraction process. Figure 4.9b shows an even better agreement with the spectral structure as well as the cross-section intensity.

The total cross-section over the region 6.2 – 10.5 μm for AYTY and PNNL is 8.02×10^{-17} cm^2/molecule and 8.20×10^{-17} cm^2/molecule respectively, making PNNL overall around 8% stronger. PNNL covers regions beyond those currently available in HITRAN. Figure 4.10(a) depicts the $2\nu_2$ band at 2.88 μm. Good agreement is seen in structure, position and cross-sections with the integrated cross-

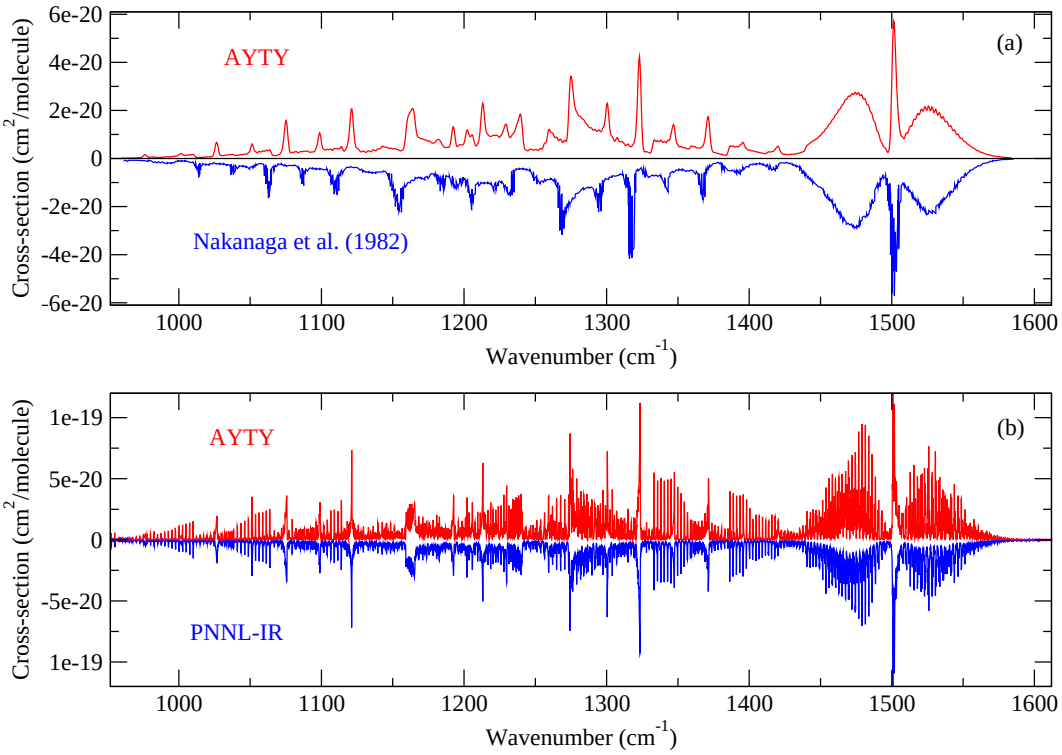

Figure 4.9: Cross-section comparison of AYTY against experimental data for the ν_3, ν_4 and ν_6 band regions: (a) Nakanaga et al. [11] at 296 K with HWHM = 1.1849 cm^{-1}; Extracted from image and scaled to match the AYTY cross-section; (b) PNNL-IR data at 323.15 K [10] with HWHM = 0.1120 cm^{-1}.

sections differing by only 10%.

The regions below 2.8 μm in PNNL become increasingly polluted with noise but band features are still visible as seen in Figures 4.10(b),(c) and (d). In particular, Figure 4.10(b), the AYTY cross-section reproduces peaks in features present in the PNNL-IR data. This region was also studied by Flaud et al. [15]. Their absorbance spectrum produces certain transitions with double the intensity compared to AYTY. These are due to splitting caused by two transitions with the same quanta but with swapped Γ_f and Γ_i giving the two lines very similar transition frequencies and absolute intensity which make them difficult to resolve experimentally.

Further bands include the integrated cross-section for the $2\nu_5$ $^{Q}\mathrm{R}_1(10)$ line at 5676.21 cm^{-1} for AYTY and Barry et al. [157] at 6.4×10^{-22} cm/molecule and 5.6×10^{-22} cm/molecule respectively making AYTY 11% stronger. In Table 4.7 we also compare theoretical (AYTY) overtone band intensities obtained by the di-

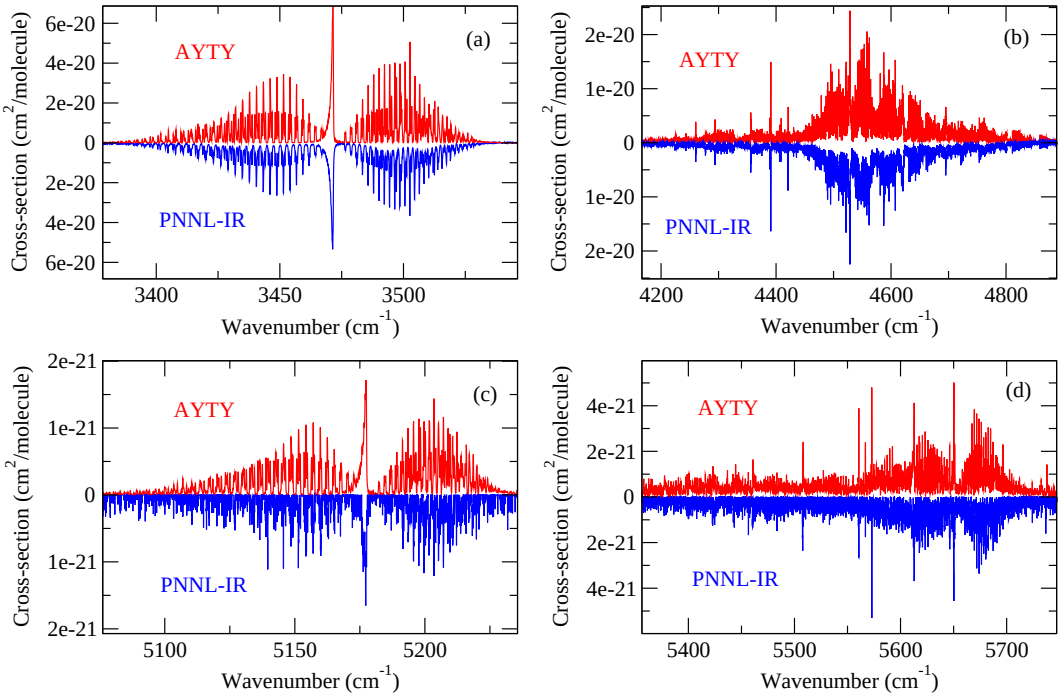

Figure 4.10: Additional bands in PNNL at wavelengths below 3.2 μm with HWHM at 0.1120 cm^{-1}. (a) $2\nu_2$ band; (b) Bands covered by Flaud et al. [15]; (c) $3\nu_2$; (d) Various bands including $\nu_1 + \nu_5$. Note: (c) and (d) Negative PNNL values have been removed.

rect summation with the corresponding experimentally derived values from Perrin et al. [142], Flaud et al. [15]. The agreement with the data obtained by Flaud et al. [15] is very good. Those from Perrin et al. [142] are in fact a compilation of different sources [158, 133, 11, 159, 135], some of which were obtained at low resolution, which could explain the slightly worse agreement with AYTY. Comparing the total integrated band intensity for the band at 1.5 μm we obtain 3.11×10^{-17} cm/-molecule against 2.19, 2.62, and 2.73 $\times 10^{-17}$ cm/ molecule by Perrin et al. [142], Nakanaga et al. [11] and from HITRAN, respectively.

Finally, Ito et al. [143] presented the relative band intensities as the ratio of the vibrational transition moments between $2\nu_4$ and $2\nu_6$ of 0.755(48), which can be compared to AYTY's absolute value of 0.6264.

The hot line list and relevant supplementary material has been published [105] and are also included in Appendix A. Currently it is being used in radiative transfer calculations by Gray et al. [160] in order to investigate formaldehyde maser action.

Table 4.7: Band intensities, in 10^{-17} cm $^{-1}$ /molecule cm $^{-2}$.

Band	Ref.	Obs	AYTY	(O-A)/O (%)
ν_1	[a]	1.008	1.057	-4.9
ν_2	[a]	1.219	1.348	-10.6
ν_3	[b]	0.184	0.185	-0.5
ν_4	[b]	0.069	0.089	-27.8
ν_5	[a]	1.120	1.282	-14.6
ν_6	[a]	0.173	0.204	-17.9
$\nu_2+\nu_3$	[c]	0.0025	0.0019	22.7
$\nu_2+\nu_6$	[c]	0.0790	0.1222	-54.6
$2\nu_3$	[c]	0.0260	0.0428	-64.5
$\nu_2+\nu_4$	[c]	0.1100	0.1379	-25.4
$\nu_3+\nu_6$	[c]	0.1940	0.3274	-68.8
$\nu_3+\nu_4$	[c]	0.0290	0.0300	-3.4
$2\nu_6$	[c]	0.0220	0.0214	2.9
$\nu_4+\nu_6$	[c]	0.0062	0.0014	77.7
$2\nu_4$	[c]	0.0060	0.0047	22.4
$\nu_1+\nu_6$	[d]	0.0015	0.0022	-45.0
$\nu_2+\nu_4+\nu_6$	[d]	0.0006	0.0007	-4.6
$\nu_3+\nu_5$	[d]	0.0097	0.0098	-1.2
$2\nu_3+\nu_6$	[d]	0.0036	0.0027	24.4
$\nu_2+\nu_5$	[d]	0.0377	0.0446	-18.2
$2\nu_2+\nu_6$	[d]	0.0108	0.0123	-14.0
$\nu_1+\nu_2$	[d]	0.0243	0.0275	-13.2
$3\nu_2$	[d]	0.0022	0.0026	-21.4

[a] Perrin et al. [139] [b] Perrin et al. [140]
[c] Perrin et al. [142] [d] Flaud et al. [15]

Table 4.8: Residuals, in cm^{-1}, for line positions for the ν_3, ν_4 and ν_6 bands. Observed data from Reuter et al. [13].

Band	J'	J''	AYTY	Obs.	Obs.-Calc.
6	17	18	1148.4322	1148.3346	-0.0976
6	17	18	1148.4578	1148.3600	-0.0978
4	11	10	1148.4115	1148.3453	-0.0662
4	3	4	1148.5548	1148.4702	-0.0846
4	16	16	1148.6150	1148.5082	-0.1068
4	1	1	1159.2222	1159.1356	-0.0866
4	2	2	1159.3587	1159.2716	-0.0871
4	28	28	1159.3222	1159.3070	-0.0152
4	15	14	1159.4760	1159.3917	-0.0843
6	6	7	1159.5539	1159.4132	-0.1407
4	9	8	1159.5115	1159.4396	-0.0719
4	3	3	1159.5594	1159.4715	-0.0879
4	18	18	1172.4242	1172.3864	-0.0378
4	6	6	1172.6002	1172.5256	-0.0746
6	12	13	1180.6607	1180.6446	-0.0161
6	4	5	1180.8080	1180.7328	-0.0752
4	24	23	1180.8209	1180.8082	-0.0127
4	11	11	1180.8777	1180.8324	-0.0453
6	13	14	1180.9109	1180.8834	-0.0275
6	10	10	1192.6923	1192.6086	-0.0837
6	3	4	1192.6678	1192.6267	-0.0411
6	9	9	1192.7477	1192.6657	-0.0820
6	8	8	1192.7985	1192.7181	-0.0804
6	18	19	1192.7781	1192.7369	-0.0412
6	7	7	1192.8441	1192.7651	-0.0790

Chapter 5

Hydrogen Peroxide

5.1 Introduction

Terrestrial hydrogen peroxide exists as a trace molecule in the Earth's atmosphere and contributes to the atmosphere's oxidising budget as well as ozone production and water chemistry [161, 162, 163, 164] and its concentration is now being routinely observed [163]. Astrophysically there have been multiple detections of H_2O_2 in the atmosphere of Mars [165, 166, 167, 168] with seasonal variation, possibly formed by triboelectricity in dust devils and dust storms [167] and may well act as an agent in the oxidization of the Martian surface. Hydrogen peroxide has also been detected in the atmosphere of Europa [169] in the 3.5 μm region. The first detection of interstellar H_2O_2 was made by Bergman et al. [170] and is believed to play an important role in astrophysical water chemistry similar to that on Earth. Du et al. [171] suggest that H_2O_2 is produced on dust-grains via the hydrogenation of grain HO_2 and released into the gas-phase through surface reactions. On the dust-grain, H_2O_2 acts as an intermediate in the formation of water and aids in the production of other species such as H_2CO, CH_3OH, and O_2.

Hydrogen peroxide belongs to the peroxide group of molecules with an HO-OH bond dissociation enthalpy of 17050 cm^{-1}[172] at 0 K. H_2O_2 is an asymmetric prolate rotor molecule and is the simplest molecule that exhibits internal rotation. This torsional motion gives rise to a double minimum potential curve with respect to its internal rotation co-ordinates as well as two alignments of the O-H bonds: *cis* and

trans. The consequence of this motion means that there are four sub-levels for each torsional excitation which are characterized by their symmetry. This necessitates the use of an additional quantum number, τ, to unambiguously describe its motion. The molecular states can be classified using the C_{2h}^{+}(M) symmetry group which best describes the torsional splitting caused by the *cis* and *trans* tunneling [173]. H_2O_2 has six vibrational modes: ν_1 and ν_5 represent the symmetric and asymmetric O-H stretching respectively, ν_3 and ν_6 represent the O-H bending modes, ν_2 represents the O-O stretch and the ν_4 mode represents the torsional excitation with the more common notation of n.

Experimental studies of ro-vibrational H_2O_2 spectra have mostly probed the torsional motion in the ground [27], the ν_3 [26] and ν_6 [28, 174] vibrational modes. Conversely, the higher-lying O-H stretching modes, ν_1 and ν_5, are poorly studied using high resolution techniques. The difference between the two stretching bands is about 8 – 10 cm^{-1} and torsional splitting from the double minimum of the potential gives rise to doubling [175] in the form of 'quasi'-degenerate states [176] that are difficult to resolve with a degree of accuracy. Olson et al. [27] give an estimate of 3610 - 3618 cm^{-1} for ν_5 and 3601 - 3617 cm^{-1} for ν_1 whilst a Raman study gives a lower value of 3607 cm^{-1} for the ν_1 band-centre [175] but determining the accuracy to better than 0.1 cm^{-1} is difficult.

H_2O_2 has been a benchmark system for developing methods aiming to treat large amplitude motion [177, 178, 179, 180]. Recent calculations on the ro-vibrational states for H_2O_2 include the *ab initio* computation using CCSD(T)-F12 electronic structure calculations of band frequencies accurate to about 4.0 cm^{-1} by Rauhut et al. [176], models of the peroxide stretches by Bacelo and Binning [181], a discrete variable representation (DVR) calculation for levels up to 6000 cm^{-1} by Chen et al. [182], Lin and Guo [183] and finally, potential energy surface (PES) calculations by Koput et al. [184] and Kuhn et al. [185]. Calculations which also consider transition intensities are rather rarer but a recent example is provided by Carter et al. [186]. The peroxide system was used to benchmark the large amplitude calculations of MULTIMODE [187] up to $J = 20$ and showed good agreement

against HITRAN line intensities but the PES used had an rms of ≈ 20 cm^{-1} against experimental band centres. However, this PES has been superseded by the higher accuracy *ab initio* potential energy surface (PES) of Małyszek and Koput [188] which was further modified by Polyansky et al. [84].

Experimental transition frequencies and intensities for H_2O_2 are available in the HITRAN 2012 database [6] but only for room temperature modelling up to 1800 cm^{-1}. This region covers the torsional, O-H bending modes and O-O stretch but misses the O-H stretches in the 3750 cm^{-1} region. Only a few studies deal with absolute intensities of H_2O_2 in the far-infrared [189, 24, 174] with only PNNL-IR [10] data providing integrated intensities in the mid-infrared region [190]. The thermal decomposition of hydrogen peroxide at 423 K makes it difficult and dangerous to study at higher temperatures.

Theoretical line lists can be used to fill in gaps in the experimental data both in terms of wavelength and temperature coverage. Two line-lists were computed for hydrogen peroxide, a room temperature and hot line-list.

5.2 *Ab-Initio* Potential Energy Surface

The *ab initio* PES is based of the high-accuracy *ab initio* calculations of Małyszek and Koput [188]. The Born-Oppenheimer PES was computed using the CCSD(T)-F12 method with aug-cc-pV5Z basis-sets at 1762 points near equilibrium geometry. The PES was then corrected for core-electron correlation, scalar relativistic and higher order valence-electron correlation effects. Additionally adiabatic effects were accounted for by computing the Born-Oppenheimer diagonal correction terms using Hartree-Fock and CCSD methods on all *ab initio* points. The points are then fit to the functional form:

$$V(q_1,q_2,q_3,q_4,q_5,\tau) = \sum_{ijklmn} f_{ijklmn} q_1^i q_2^j q_3^k q_4^l q_5^m \cos nq_6 \tag{5.1}$$

where $q_i (i = 1,2,3,4,5)$ are expansions around the equilbrium geometries of the O-O bond length (R_e), O-H bond length (r_e) and O-H bond angle (θ_e). The definitions

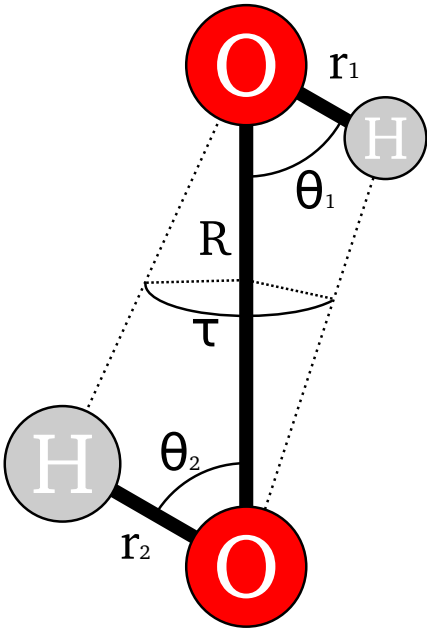

Figure 5.1: The internal co-ordinates of HOOH.

for q_i are

$$\begin{aligned} q_1 &= \frac{r_1 - r_e}{r_e} \\ q_2 &= \frac{r_2 - r_e}{r_e} \\ q_3 &= \frac{R - R_e}{R_e} \end{aligned} \tag{5.2}$$

which are the Simons-Parr-Finlan stretching co-ordinates in terms of the bond lengths: O-O (R), O-H_1 (r_1) and O-H_2 (r_2) and

$$\begin{aligned} q_4 &= \theta_1 - \theta_e \\ q_5 &= \theta_2 - \theta_e \end{aligned} \tag{5.3}$$

are the bending co-ordinates in terms of the bond angles for O-H_1 (θ_1) and O-H_2 (θ_2). $q_6 = \tau$ is the torsional angle. The co-ordinates are described in Figure 5.1. This PES has an rms of $\approx$10 cm^{-1} for vibrational band origins. Polyansky et al. [84] further improved the PES by utilizing the larger aug-cc-pV7Z basis-set for certain parts of the PES as well as a small adjustment of the *ab initio* equilibrium geometry and height of the torsional barrier. The PES reproduces observed energies with an

Table 5.1: Primitive basis functions, co-ordinate borders, number of grid points for Numerov-Cooley integration and number of solutions

Basis func.	Borders	No. grid points	No. soln
$\phi_{n_1}(r_1)$	-0.50–1.00	1000	16
$\phi_{n_2}(r_2)$	-0.40–1.20	1000	8
$\phi_{n_3}(r_3)$	-0.40–1.20	1000	8
$\phi_{n_4}(\theta_1)$	-1.40–1.40	1000	14
$\phi_{n_4}(\theta_1)$	-1.40–1.40	1000	14
$\phi_{n_6}(\tau)$	0.0°–720.0°	10000	42

rms of 1 cm^{-1} for rotational levels up to $J = 35$ within low-lying vibrational states. By utilizing empirical band centre shifting as given in Eq. (2.58), this is reduced to 0.02 cm^{-1} making it a good starting point for computing variational energies.

5.3 Variational computation

The kinetic energy is expanded around the reference geometry in terms of five linearized co-ordinates of the form:

$$\zeta_i = x_i^l - x_i^e \tag{5.4}$$

where x_i^l and x_i^e represent linearized version and equilibrium geometry of the bond lengths and angles respectively. Here, $i = 1$, $i = 2$, $i = 3$, $i = 4$ and $i = 5$ represent R, r_1, r_2, θ_1 and θ_2 respectively and $i = 6$ is the sixth co-ordinate, $\zeta_6 = \tau$. Similarly the potential is expanded but the stretches are represented in terms of Morse-type functions for ζ_1,ζ_2 and ζ_3 and bending functions for ζ_4 and ζ_5. For both line-lists the kinetic energy expansion order is 6 and the potential expansion order is 8.

Like formaldehyde, the basis-set was constructed from the methods outlined in Chapter 2.3.3.2 with parameters outlined in Table 5.1 that is truncated via polyad number $P_{\max} = 42$ [84]. The allowed modes follow the polyad rule:

$$P = 4v_1 + 8(v_2 + v_3 + v_4 + v_5) + v_6 \leq P_{\max}. \tag{5.5}$$

The six dimensional co-ordinate space is then divided into four reduced subspaces: (ζ_1), (ζ_2, ζ_3), (ζ_4, ζ_5) and (ζ_6) based on symmetry and solved for Eq. (2.52) to

obtain the contracted vibrational basis-functions $\Phi_1(\zeta_1)$, $\Phi_2(\zeta_2,\zeta_3)$, $\Phi_3(\zeta_4,\zeta_5)$ and $\Phi_4(\zeta_6)$. These basis-functions are then symmetrised according to $\mathscr{D}_{2h}$ symmetry and the final vibrational basis-set is formed from the product of the four contracted basis-functions which are truncated via Eq. (5.5) and symmetrized again. Finally the contracted basis-sets are used to solve Eq. (2.54) for the $J = 0$ wavefunctions. Using these symmetrized wavefunctions also has the benefit of the Hamiltonian matrix being factorized into independent blocks according to $\mathscr{D}_{2h}$ symmetry. The $\mathscr{D}_{2h}$ is isomorphic to the C_{2h}^{+}(M) symmetry group which best describes the torsional splitting caused by the *cis* and *trans* tunneling [173]. The irreducible representations of this group are A_g, A_u, B_{1g}, B_{1u}, B_{2g}, B_{2u}, B_{3g} and B_{3u}. However, the states corresponding to B_{2g}, B_{2u}, B_{3g} and B_{3u} have zero statistical weight and therefore their matrix blocks are not constructed and diagonalized for $J > 0$. It is usual to describe the H_2O_2 torsional modes using the notation (n,τ), where n describes the excitation of the torsional mode. The excitations of the torsional (v_4/n) mode are representated by A_g, A_u, B_{2g} or B_{2u} symmetry which correspond to the quanta $\tau = 1$,$\tau = 4$,$\tau = 2$ and $\tau = 3$ respectively.

The τ quantum number can be preserved in the quantum number assignment in TROVE by utilizing the following form:

$$v_4 = 4n + i, \tag{5.6}$$

where n is the excitation and i is the symmetry where $i = 0,1,2,3$ is A_g, B_{2g}, B_{2u} and A_u respectively. To retrieve n and τ is simply:

$$\tau = (v_4 \bmod 4) + 1, \quad n = \frac{v_4}{4} \tag{5.7}$$

5.4 Dipole Moment Surface

An *ab initio* DMS computed at the CCSD(T)-f12b/aug-cc-pV(T+d)Z [81] level of theory in the frozen-core approximation using CCSD(T) [191] on a grid of 50 000 geometries in conjunction with the finite electric field method and field of 0.005

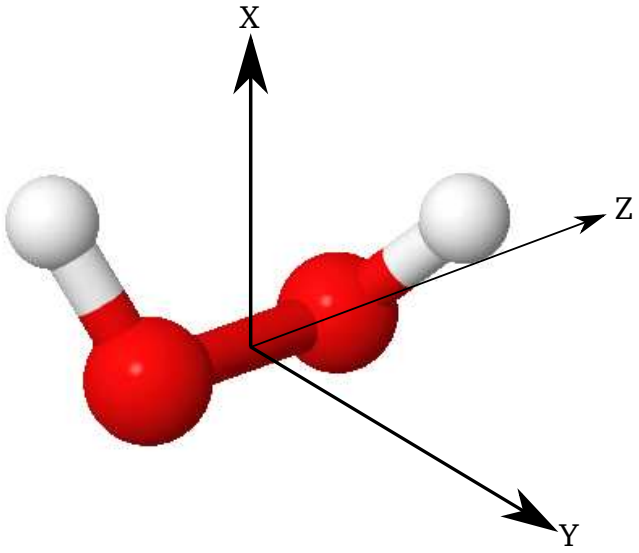

Figure 5.2: The principal axes for HOOH used in the DMS expansion.

a.u. The $\mathscr{D}_{2h}$ symmetry-adapted projections of the dipole moment Cartesian components μ_x, μ_y, and μ_z are given in the analytical representations with each component expanded in Taylor series (312 parameters in total) in terms of internal coordinates around the equilibrium configuration using a molecule-fixed axis system as follows. The z axis is aligned along the O-O bond, and the x axis lies in the plane bisecting the two O-O-H planes (i.e. planes containing the O-O and O-H bonds). The y axis is oriented such that the xyz axis system is right-handed. These xyz axes are not exact but are close to the principal axis system shown in Figure 5.2. With the chosen axes, the x, y, and z components of the dipole moment span the B_{1u}, B_{3u}, and B_{2u} representations, respectively. The three electronically averaged dipole components are represented by the following analytical functions:

$$\bar{\mu}_x = \cos(\tau/2) \sum_{i_1,i_2,\ldots,i_6} \mu^{(x)}_{i_1,i_2,\ldots,i_6} \xi_1^{i_1} \xi_2^{i_2} \xi_3^{i_3} \xi_4^{i_4} \xi_5^{i_5} \xi_6^{i_6}, \tag{5.8}$$

$$\bar{\mu}_y = \sin(\tau/2) \sum_{i_1,i_2,\ldots,i_6} \mu^{(y)}_{i_1,i_2,\ldots,i_6} \xi_1^{i_1} \xi_2^{i_2} \xi_3^{i_3} \xi_4^{i_4} \xi_5^{i_5} \xi_6^{i_6}, \tag{5.9}$$

$$\bar{\mu}_z = \sum_{i_1,i_2,\ldots,i_6} \mu^{(z)}_{i_1,i_2,\ldots,i_6} \xi_1^{i_1} \xi_2^{i_2} \xi_3^{i_3} \xi_4^{i_4} \xi_5^{i_5} \xi_6^{i_6}, \tag{5.10}$$

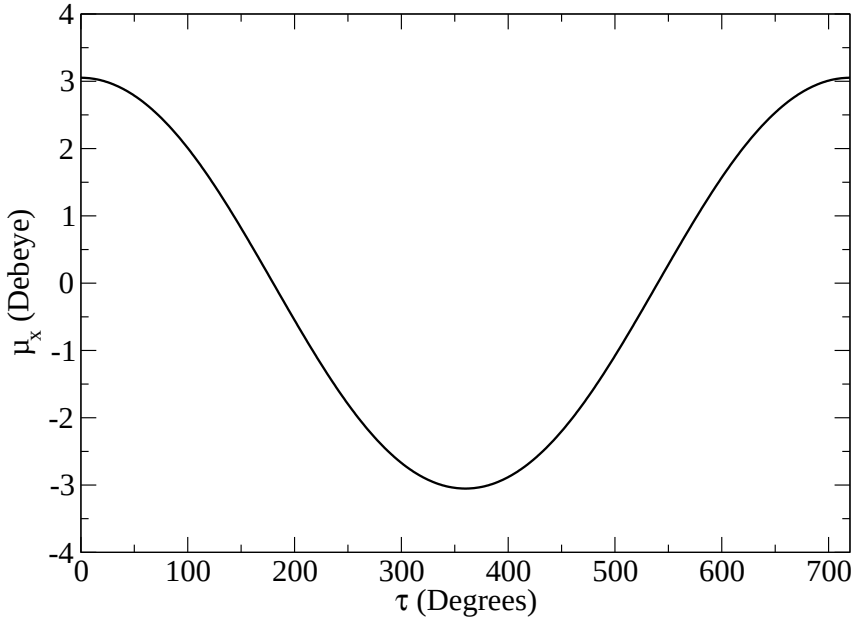

Figure 5.3: The μ_x dipole moment component for H_2O_2, computed at the torsional geometries shown

where

$$\xi_1 = \Delta R e^{-(\Delta R)^2}, \tag{5.11}$$

$$\xi_2 = \Delta r_1 e^{-(\Delta r)^2}, \tag{5.12}$$

$$\xi_3 = \Delta r_2 e^{-(\Delta r)^2}, \tag{5.13}$$

$$\xi_4 = \Delta\theta_1, \tag{5.14}$$

$$\xi_5 = \Delta\theta_2, \tag{5.15}$$

$$\xi_6 = \cos\tau. \tag{5.16}$$

The expansion parameters of the y and z components of the dipole obey the following permutation rule:

$$\mu^{(\alpha)}_{i_1,i_2,i_3,i_4,i_5,i_6} = -\mu^{(\alpha)}_{i_1,i_3,i_2,i_5,i_4,i_6} \tag{5.17}$$

corresponding to the permutation of the two hydrogen atoms and therefore $\mu^{(\alpha)}_{0,0,0,0,0,i_6} = 0$ ($\alpha = y,z$) for any i_6. The dependence of μ_x component against the torsional angle is shown in Figure 5.3.

The 130, 90, and 92 symmetrically independent expansion parameters

$\mu^{(\alpha)}_{i_1,i_2,i_3,i_4,i_5,i_6}$ ($\alpha = x,y,z$) were obtained in a least-squares fit to the $3 \times 20\ 842$ *ab initio* dipole moment values (corresponding to energies below $hc12\ 000$ cm^{-1}) with the rms error of 0.0013, 0.0002, and 0.0010 D, respectively.

The vibrational transition moment for the ground vibrational state is 1.5683 D and compares well with measured value of 1.5728 D [192]. This is very different from the equilibrium value of the *ab initio* dipole moment $\bar{\mu}^{\rm e}_x$ = 1.738 D (at $R =$ 1.4554 Å, $r_1 = r_2 = 0.96257$ Å, $\theta_1 = \theta_2 = 101.083^\circ$), showing strong non-rigid character of the H_2O_2 dipole moment.

Information on transition moments for H_2O_2 is limited; Table 5.2 compares available experimentally derived values at different torsional excitations from experiment [24]. Our calculated values reproduce the experimental with a maximum deviation of 2.4%. Some papers report effective transition dipole moments as a torsional expansion in terms of τ, e.g. $\cos\tau\phi_z$ [24] which are difficult to compare to fully averaged transition dipoles.

Table 5.2: Absolute values of experimental [24] and *ab initio* transition moments, in debye, for $(n',\tau = 1) \leftrightarrow (n'',\tau = 3)$.

n'	n''	Calc	Obs	(Obs-Calc)/Obs (%)
0	0	1.5683	1.5723	0.25
0	1	0.3332	0.3413	-2.40
1	0	0.6031	0.6136	-1.72
1	1	1.1664	1.1751	0.74
2	1	1.1664	1.1628	0.31
2	2	1.2638	1.2825	1.46
3	2	1.3276	1.3535	-1.91

5.5 Room temperature line-list

The basis-sets used is described in Section. 5.3. Here the $J = 0$ wavefunctions with eigenvalues up to 8000 cm^{-1} are utilized reducing the size of the Hamiltonian. The original primitive basis-set was of size 2 789 400, this was reduced to 23 078 in the contracted form and finally to 2875 using the $J = 0$ representaion. The room temperature rovibrational energies utilise the empirical band-centre shifting outlined in Eq. (2.58). Table 5.3 lists all the band-centers that were utilized in the empirical

shifts together with the *ab initio* values before the shift.

Table 5.3: Experimental [25, 26, 27, 28] band centres used in the empirical shift.

v_1	v_2	v_3	n	v_5	v_6	τ	Symmetry	Ab-Initio (cm^{-1})	Shifted/Obs (cm^{-1})
0	0	0	0	0	0	4	A_u	11.312	11.437
0	0	0	0	0	0	3	B_{2u}	11.312	11.437
0	0	0	1	0	0	1	A_g	255.529	254.55
0	0	0	1	0	0	2	B_{2g}	255.532	254.55
0	0	0	1	0	0	4	A_u	371.589	370.893
0	0	0	1	0	0	3	B_{2u}	371.590	370.893
0	0	0	2	0	0	1	A_g	570.809	569.743
0	0	0	2	0	0	2	B_{2g}	570.818	569.743
0	0	0	2	0	0	4	A_u	777.432	776.1221
0	0	0	2	0	0	3	B_{2u}	777.458	776.1148
0	0	1	0	0	0	1	A_g	865.539	865.939
0	0	1	0	0	0	2	B_{2g}	865.539	865.939
0	0	1	0	0	0	4	A_u	877.470	877.934
0	0	1	0	0	0	3	B_{2u}	877.470	877.934
0	0	0	3	0	0	1	A_g	1002.666	1000.882
0	0	0	3	0	0	2	B_{2g}	1002.869	1000.93
0	0	0	0	0	1	1	B_{1u}	1265.003	1264.583
0	0	0	0	0	1	3	B_{1g}	1285.879	1285.121
0	0	0	1	0	1	1	B_{1u}	1506.164	1504.872
0	0	0	1	0	0	3	B_{1g}	1649.977	1648.367
0	0	0	2	0	1	1	B_{1u}	1855.823	1853.634
0	0	0	2	0	1	3	B_{1g}	2075.366	2072.404

Hamiltonian matrices are constructed up to the limit of $J = 40$ and diagonalized using the $J = 0$ contracted basis set for all eigenvalues and eigenvectors but only eigenvectors up to 8000 cm^{-1} are stored and used in producing the transitions. The required linestrengths were produced from the DMS outlined in Section 5.4 via the methodology described in Chapter 3. Here the symmetry selection rules are:

$$A_g \leftrightarrow A_u \ , B_{1g} \leftrightarrow B_{1u} \tag{5.18}$$

with nuclear statistical weights $g_{\rm ns} = 1$ for A_g and A_u, $g_{\rm ns} = 3$ for B_{1g} and B_{1u} and $g_{\rm ns} = 0$ for the B_{2g}, B_{2u}, B_{3g} and B_{3u} symmetries. The transitions were computed using the energy limits hc 4 000 and hc 8 000 cm^{-1} for the lower and upper states, respectively to acheive a target $\tilde{\nu}_{if}$ limit of 8 000 cm^{-1}. The TROVE-GAIN code

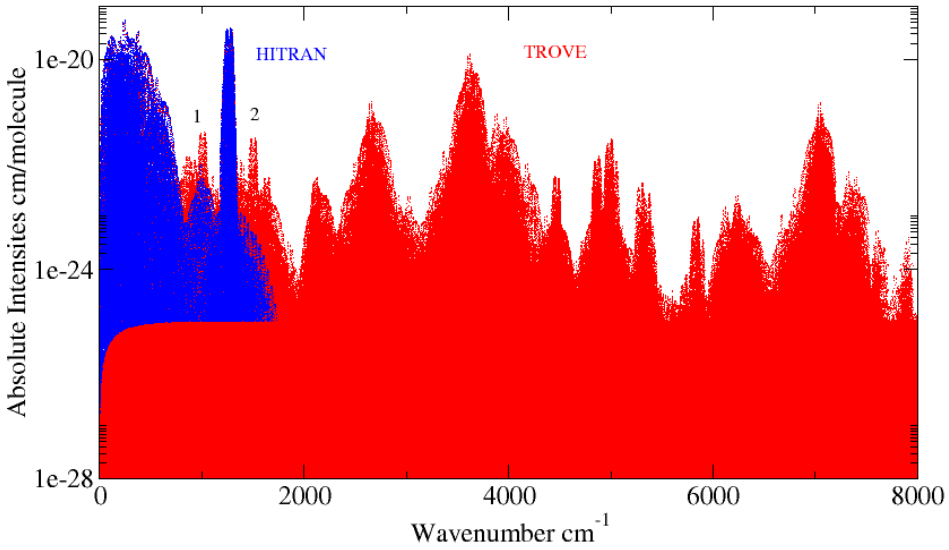

Figure 5.4: Overview of synthetic spectrum at T =296 K against HITRAN [6]

was utilized to compute 1 487 073 009 transitions within 6 hours.

5.5.1 Results

The line-list obtains a partition function of 9840.91 at $T = 296$ K which compares well to the HITRAN value 9819.80 [23].

Figure 5.4 is the synthetic spectrum at 296 K for all 1.4 billion lines and comparing against the 126,983 lines from HITRAN highlights the significant degree of completeness the line-list provides. However two regions (1) and (2) show disagreement in line intensity, which can be attributed to the lack of the experimental data used for producing the HITRAN intensities by Perrin et. al [24, 174].

The PNNL-IR [10] database provides additional cross-sections above 1800 cm^{-1}. Figure 5.5 compares the line list's and HITRAN's simulated cross-sections to PNNL using a Gaussian convolution with HWHM at 0.312 cm^{-1} at T=323.15 K and demonstrates that the line list agrees much better in intensity and structure indicating problems with HITRAN intensities for both regions.

Figure 5.6 highlights a band in the 3.5 μm region which is a combination of the $(0,\tau=1) \rightarrow (\nu_3+6\nu_4,\tau=1)$, $(0,\tau=1) \rightarrow (2\nu_3+4\nu_4,\tau=1)$, $(0,\tau=2) \rightarrow$

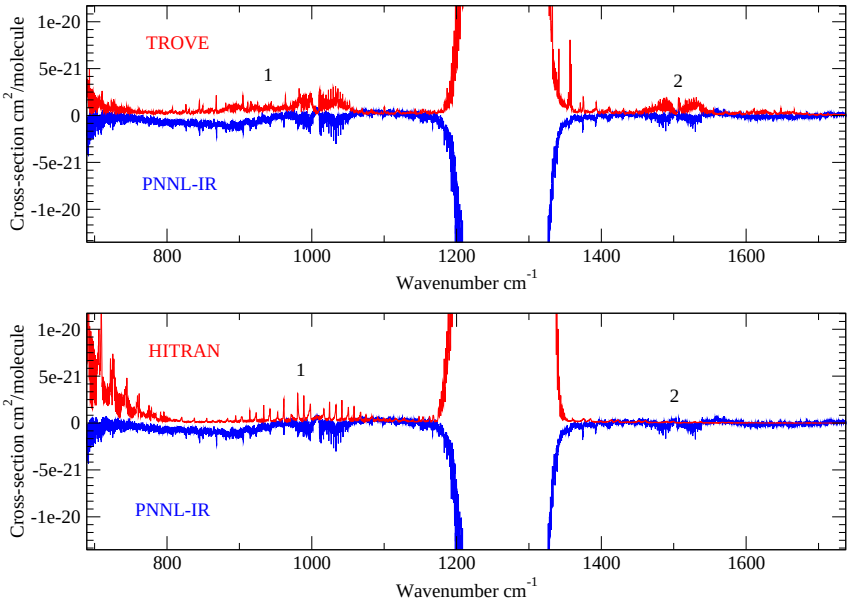

Figure 5.5: Cross-section comparison of the room temperature line-list against experimental PNNL-IR [10] data at the ν_1 and ν_5 band at 323.15 K with HWHM = 0.3120 cm^{-1}.

$(3\nu_3, \tau = 1)$, $(0, \tau = 2) \rightarrow (\nu_3 + \nu_4 + \nu_5, \tau = 2)$, $(0, \tau = 3) \rightarrow (\nu_3 + 4\nu_4, \tau = 4)$ and other weaker hot bands. Good agreement is seen in both structure and overall intensity but the estimated line profile utilized in the convolution may not be adequate enough to properly replicate the PNNL-IR cross-section leading to some minor differences in the overall cross-section.

Figure 5.7 further states the quality of both the line-positions and absolute intensities by comparing the ν_1 and ν_5 (ours vs PNNL-IR's) bands in the 2.7 μm region. As this is the region of most contention in the literature, it is hopeful that this line-list may provide a means with which to identify the confusing spectra in this region.

Finally, the importance of the band shift previously discussed is illustrated in Figure 5.8. Here the TROVE-I spectra is purely using the *ab initio* band centers while TROVE-II utilizes the experimental band centers from Table 5.3. The *ab initio* deviation of 1.12 cm^{-1} reduces significantly to 0.005 cm^{-1} using this empirical shifting method.

The room temperature line-list has been published[16] and is also available in Appendix A.

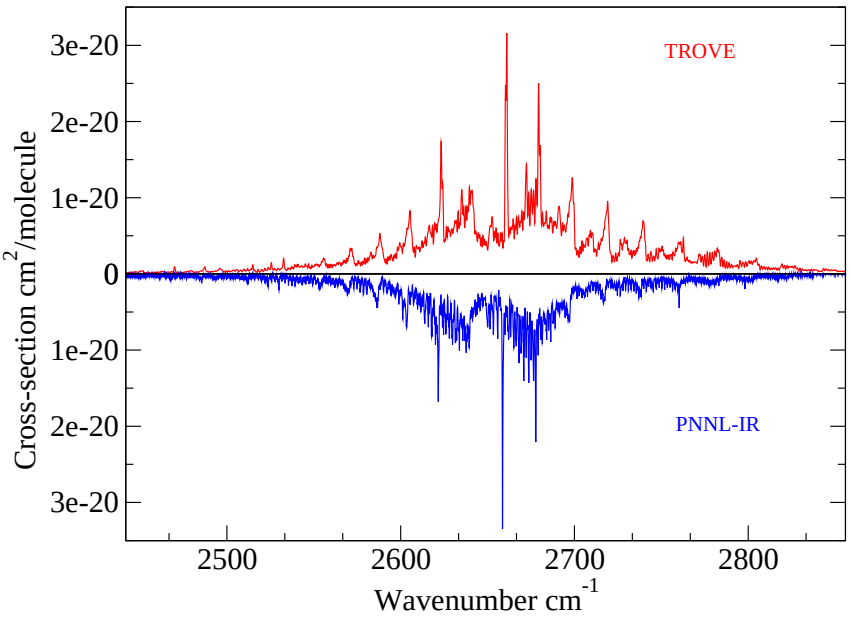

Figure 5.6: Cross-section comparison of the room temperature line-list against experimental PNNL-IR [10] data at 323.15 K with HWHM = 0.3120 cm^{-1}.

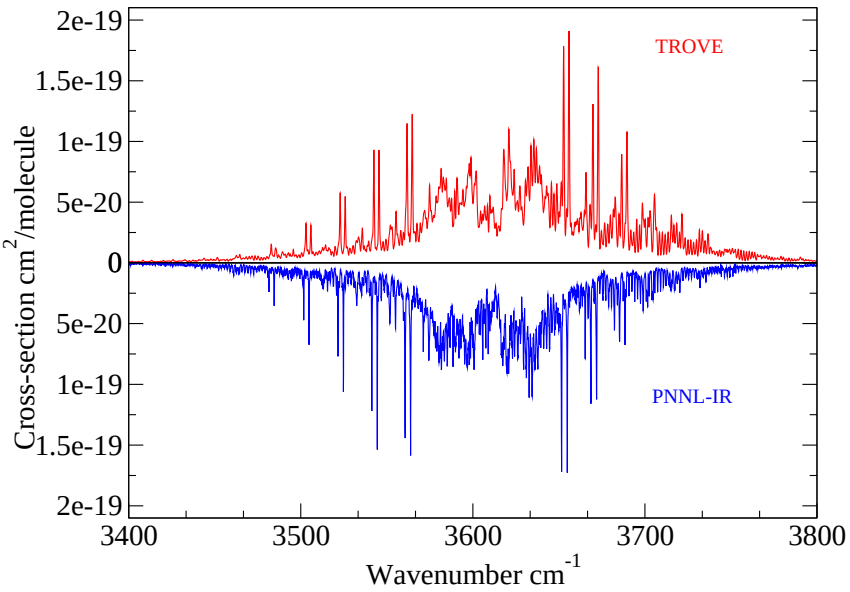

Figure 5.7: Cross-section comparison of the room temperature line-list against experimental PNNL-IR [10] data at the ν_1 and ν_5 band at 323.15 K with HWHM = 0.3120 cm^{-1}.

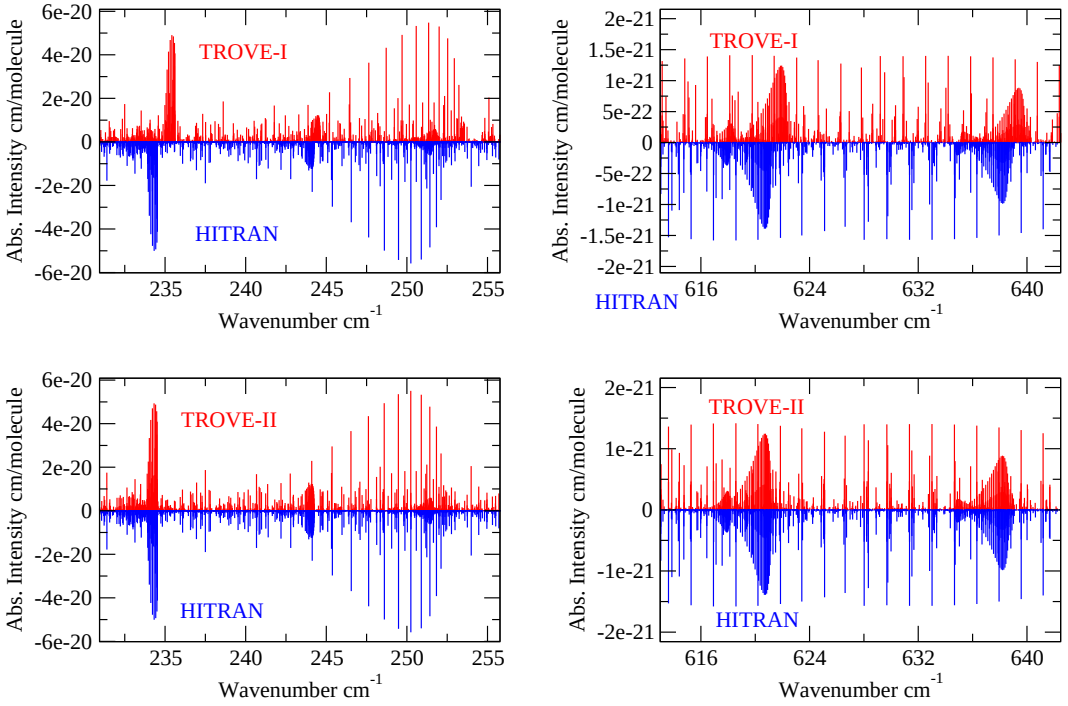

Figure 5.8: Comparing two versions of the synthetic spectrum against HITRAN at T=295 K. TROVE-I is the *ab initio*, TROVE-II is using the empirical band-center shifts.

5.6 Hot line-list

The room temperature line-list provides about 1 billion transitions at up to 8,000 cm^{-1}. However it is limited as the rotational excitation of $J = 40$ makes it inadequate for high temperature modelling and the lower energy cut-off means that coverage above 4,000 cm^{-1} rapidly becomes incomplete. This new line-list aims to build upon this line-list by refining the PES towards spectroscopic accuracy and extending the temperature and frequency range applicability.

5.6.1 Refined Potential Energy Surface

The previous room-temperature H_2O_2 line list [16] was computed using the PES stated in Section 5.2. The *ab initio* PES combined with the empirical shifting reproduces the known empirical energy levels with a root mean square (rms) of about 0.001 – 0.1 cm^{-1}. As previously stated in Chapter 2.3.5, this can be considered an ad-hoc PES and will therefore label it as the 'H2O2-2015' PES. However, its predictive ability for bands whose centres have not been experimentally defined was

limited. In particular, the disagreements from the ν_1 and ν_5 band-centers. Essentially, their band energies have an accuracy that is closer to *ab initio* than experimental. In order to improve them towards experimental accuracy, refinement was performed as outlined in Chapter 2.3.5. Here the original 282 expansion coefficients of the *ab initio* PES are reduced to 163 by removing the symmetry-related O-H stretching (ij) and bending (kl) terms from the input PES and simply linking them in the computation of potential energy terms in the Hamiltonian. This ensures that the symmetry of these terms is preserved during the fitting process. The quality of the fit is determined by the quality and vibrational diversity of the input dataset. Two sources of experimental data come from line-positions provided in the literature and transitions from HITRAN. The HITRAN dataset sources come from observations by Perrin et al. [174], Perrin et al. [24], Perrin et al. [28] and Klee et al. [193] with literature line-positions from Flaud et al. [25], Olson et al. [27] Giguere [194], Zumwalt and Giguere [189] and Camy-Peyret et al. [26]. This empirical dataset provides the ν_4, $\nu_3+\nu_4$, $\nu_4+\nu_6$ and ν_2 vibrational terms. Unfortunately there is little data on the ν_1 and ν_5 energy levels which hampers the vibrational diversity that would construction of an extensive fitted PES, and the reported band-centre values vary significantly in literature making them unsuitable for inclusion. However, these terms can be indirectly improved by the high J values from other vibrational states included in the fit.

Our input dataset includes all energies for $J \leq 4$ up to 4000 cm^{-1}. The weights w_i used have an arbitrary range of values that are normalized in the fit. The energies given in literature are the simplest to include in the refinement process and are given the highest weighting. Here the pure torsional band at $J > 0$ from Camy-Peyret et al. [26] and Olson et al. [27] are given the highest weighting of $w_i = 100$. The ν_2, ν_3, ν_6, $\nu_3+\nu_4$, $\nu_4+\nu_6$ energies and H_2O_2 band centers (except for ν_1 and ν_5) from Camy-Peyret et al. [26], Giguere [194], Perrin et al. [28], Flaud et al. [25] and Zumwalt and Giguere [189] are given weights $10 \leq w_i \leq 20$.

Transitions from HITRAN require additional work. In order to determine the upper state of a transition requires the assignment of lower state energy. Fortunately

Table 5.4: Comparison of N weighted experimental data-points in the fit and non-weighted root mean squared deviation of both H2O2-2016 (this work) and H2O2-2015 [16] for each dataset.

Weight	N	H2O2-2016 rms (cm^{-1})	H202-2015 rms (cm^{-1})	Comment
100	43	0.001	0.000	$J > 0$ pure torsional states
10-20	144	0.004	0.007	Band centers and $\nu_3 + \nu_4$, $\nu_6 + \nu_4$ states
1-9	186	0.539	1.369	Upper state extracted from HITRAN with corroborated lower states

HITRAN provides the lower state energy for all transitions in the database. However, lower state energies require corroboration from literature data and/or the *ab initio* energies for the upper state energies to be included in the fit with $1 \leq w_i \leq 9$ based on confidence of the datum. Each input datum must be correlated with a theoretically computed energy level which, in this present work, was straightforward due to the good agreement given by the initial *ab initio* PES.

Special measures must be taken in order to ensure that the refinement process does not lead to unphysical shapes for the new PES due to a limited sampling of the experimental data not covering all the complexity of the potential energy surface of HOOH. For example the high stretching or bending overtones are poorly represented in the experimental data and therefore it is important to retain the *ab initio* quality of the original PES by Polyansky et al. [84]. To this end the PES is constrained around the *ab initio* energies at each geometry [100, 98, 97, 9]

The new potential energy surface is called H2O2-2016. Table 5.4 describes the rms for states of a particular weight. The energies for H2O2-2016 are computed without any empirical band shifts and shows that this new semi-empirical PES performs better overall than the *ab initio* band-shifted PES especially for the lower weighted states. $w_i \geq 10$ relate to vibrational states that were involved in the band-shifting which gives H2O2-2015 its low rms values. Comparing weights lower than 10 suggests that the predictive ability of the H2O2-2016 PES is greatly enhanced.

The overall comparison as a function of rotational quantum number J with a weighted rms is given in Table 5.5. Overall H2O2-2016 improves the rms deviations of H2O2-2015 by more than a factor of 2. Table 5.6 highlights residuals for $J \geq 30$

Table 5.5: Comparision of N experimental data-points in the fit and weighted root mean squared deviation of both H2O2-2016 (this work) and H2O2-2015 [16].

J	N	H2O2-2016 wrms (cm^{-1})	H202-2015 wrms (cm^{-1})
0	34	0.238	0.254
1	47	0.079	0.320
2	81	0.096	0.345
3	116	0.183	0.404
4	132	0.154	0.287
Total		0.150	0.321

for the ν_3 and ν_6 line positions from Camy-Peyret et al. [26] and Perrin et al. [28] and shows excellent agreement with an overall rms of 0.0642. The rms deviation for all 2734 states in HITRAN up to $J = 49$ and energy up to 3461.02 cm^{-1} is 0.834 cm^{-1}. Vibrational terms that correspond to the highest weighted states have an rms of 0.192 cm^{-1}. Around 12 states related to higher excited torsional modes $n > 3$ have an rms of 5.2 cm^{-1} and may well be due to misassignments. This PES is therefore of improved accuracy and is the one used below.

5.6.2 APTY

The variational calculation uses the same co-ordinates, basis-sets and DMS as the previous room temperature line-list. The difference is that the $J = 0$ wavefunctions are retained at up to to 12,000 cm^{-1} and Hamiltonian matrices are constructed up to the limit of $J = 85$.

The transitions were computed using the energy limits hc 6 000 and hc 12 000 cm^{-1} for the lower and upper states, respectively giving complete coverage of the region 0 cm^{-1}– 6 000 cm^{-1}.

The intensities were computed using an enhanced version of TROVE that utilizes the nVidia graphics processing units (GPU) allowing for the computation of 5,000–30,000 transitions per second on a single GPU. The GPUs utilized were the nVidia M2090, K20 and the K40 models.

The final hot line list named APTY contains 7 560 352 states and almost 20 billion transitions that completely covers the 0 – 6 000 cm^{-1} region. An extended line list is provided which contains an additional 8 billion transitions in the 6 000–

Table 5.6: Residuals in cm^{-1} for energies computed from the H2O2-2016 PES. Observed data is from Camy-Peyret et al. [26] and Perrin et al. [28]. The overall rms is 0.0642 cm^{-1}

J	K	ν_3	n	τ	ν_6	Obs	Calc	O-C
30	0	0	0	1	0	789.58	789.64	-0.06
30	1	0	0	2	0	793.05	793.12	-0.07
30	2	0	0	1	0	829.29	829.34	-0.05
30	2	0	0	4	0	841.34	841.32	0.02
30	3	0	0	2	0	876.03	876.07	-0.04
30	3	0	0	2	0	876.03	876.07	-0.04
30	4	0	0	1	0	940.03	940.07	-0.04
30	6	0	0	4	0	1134.61	1134.58	0.03
30	2	0	1	4	0	1198.50	1198.58	-0.08
30	7	0	0	2	0	1241.30	1241.33	-0.03
30	7	0	0	2	0	1241.30	1241.33	-0.03
30	7	0	0	3	0	1253.29	1253.25	0.04
30	4	0	1	4	0	1308.34	1308.43	-0.09
30	11	0	0	2	0	1898.17	1898.11	0.06
30	11	0	0	3	0	1910.47	1910.41	0.06
30	6	1	0	2	0	1978.18	1978.19	0.00
30	5	0	1	2	1	2525.44	2525.50	-0.06
30	0	0	2	1	1	2638.21	2638.21	0.00
30	2	0	2	1	1	2678.98	2678.96	0.02
30	6	0	1	4	1	2768.27	2768.35	-0.08
30	9	0	0	3	1	2825.75	2825.74	0.01
30	9	0	1	3	1	3180.32	3180.35	-0.03
30	11	0	0	3	1	3196.72	3196.74	-0.02
30	10	0	1	4	1	3355.50	3355.61	-0.11
31	3	0	0	2	0	928.92	928.97	-0.04
31	4	0	0	1	0	992.69	992.73	-0.04
31	6	0	0	4	0	1187.24	1187.21	0.03
31	1	0	1	3	0	1230.77	1230.86	-0.09
31	2	0	1	4	0	1253.76	1253.86	-0.10
31	7	0	0	2	0	1293.78	1293.81	-0.03
31	7	0	0	2	0	1293.78	1293.81	-0.03
31	0	0	2	1	0	1410.64	1410.73	-0.09
31	0	0	2	1	0	1410.64	1410.73	-0.09
31	8	0	0	4	0	1442.61	1442.56	0.05
31	9	0	0	3	0	1597.29	1597.23	0.06
31	1	1	0	2	0	1717.38	1717.60	-0.23
31	2	1	0	1	0	1736.75	1736.79	-0.05
31	10	0	0	4	0	1769.53	1769.46	0.07
31	3	1	0	2	0	1783.76	1783.79	-0.03
31	4	1	0	1	0	1847.46	1847.49	-0.03

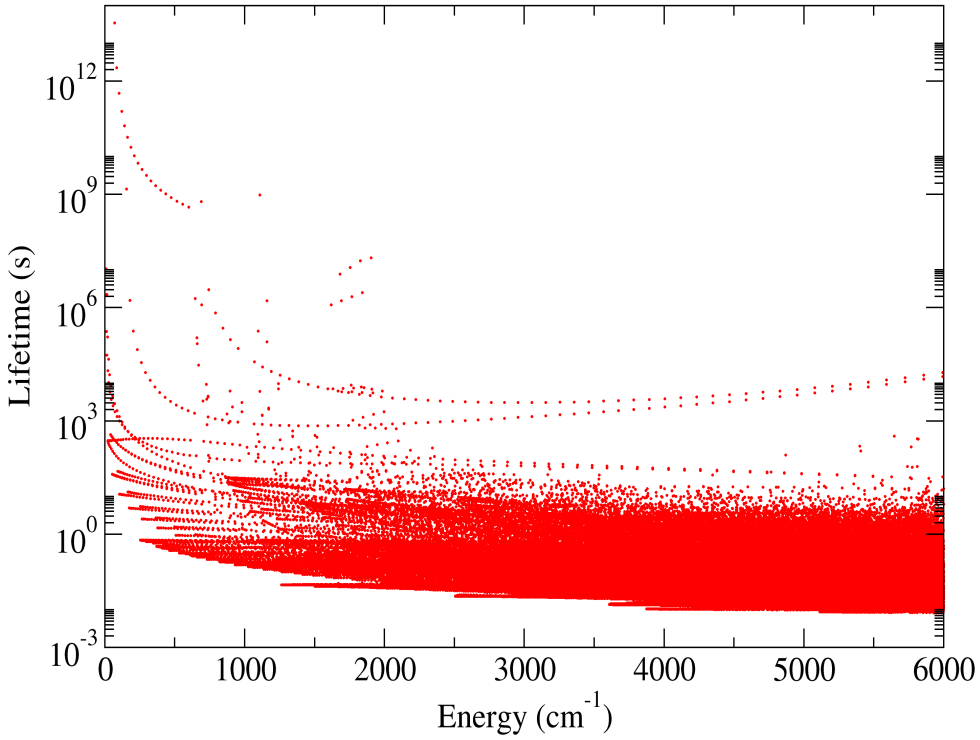

Figure 5.9: H_2O_2 lifetimes computed [12] for states up to 6,000 cm^{-1}

8 000 cm^{-1} region with reduced completeness for higher temperature. Figure 5.9 presents the lifetimes computed[12] for states up to 6,000 cm^{-1}.

An estimate of the temperature applicability of the line list can be performed by checking the partition function convergence of Eq. (3.9), which is computed via explicit summation [46]. The convergence can be measured by computing $(Q_J - Q_{J-1})/Q_J$ where Q_J is the partition function for all energy levels up to rotational excitation J. For 296 K, the partition function converges to 0.001% at $J = 37$ which matches the room temperature line lists J limit of $J = 40$. At higher temperatures, it is well-converged up to at least 1500 K where the estimated error is only 0.2 % at $J = 85$. This can be attributed to the good coverage of J states computed that contribute to the overall population. A second partition Q_{lim} can be evaluated by only including states that fall below the hc 6 000 cm^{-1} lower state energy limit of the line list and compared against Q by computing the ratio Q_{lim}/Q to assess the completeness of the full line list. Figure 5.10 shows that up to 800 K, the partition functions are essentially the same. At 1250 K about 90% of the population of states

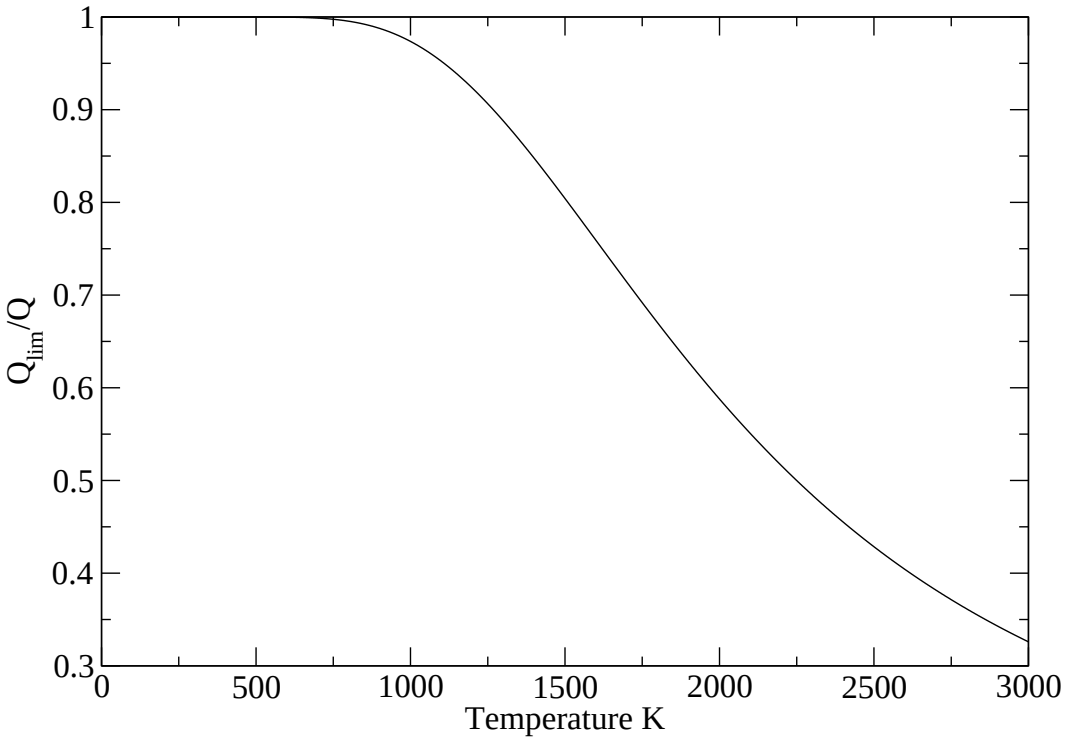

Figure 5.10: $Q_{\rm lim}/\ Q$ against temperature where $Q_{\rm limit}$ is the partition function computed using only energy levels below the lower state threshold of 6000 cm $^{-1}$ and Q is the estimate of the full partition function.

is represented by Q_{lim} but this falls to ≈80% at 1500 K giving the upper temperature for which APTY is reasonably complete as 1250 K. Usage of the line list at higher temperatures runs the risk of losing opacity due to missing contributions. The ratio Q_{lim}/Q can be used to estimate this. However, the decomposition of H_2O_2 means that it is unlikely to be an important species above 1000 K.

Table 5.7 compares the APTY partition function against HITRAN; we see that at temperatures less than 1000 K we agree better than 1%. For temperatures 1000 – 1500 K, the APTY partition function is greater by 2–4% suggesting that the explicit summation method gives higher, and probably better values, than the more approximate method used by HITRAN [23]. However, at 3000 K APTY's Q is 30% lower than the HITRAN value which can be attributed to the eigenvalue cutoff of 12 000 cm^{-1} and $J = 85$. Studies on ammonia and phosphine have shown that considerably extended lists of energy levels are required to get converged partition sums at these elevated temperatures [47].

Table 5.7: Comparisons of H_2O_2 partition functions as function of temperature for this work those used in HITRAN [23].

T / K	APTY	HITRAN
75	895.506	894.866
150	2 815.866	2 811.187
255	7 360.598	7 336.856
300	10 126.961	10 087.090
500	31 246.17	30 990.11
1000	232 439.8	226 152.5
1500	1 031 673.6	993 983.8
3000	21 847 680	15 151 254

Figure 5.11 is a simulated spectrum of the APTY line list computed at $T =$ 296 K. This highlights the coverage and sheer number and density of transitions available compared to the current edition of the HITRAN database [6]. Figure 5.12 compares our results with specific regions in the HITRAN database, the torsional and ν_6 bands. Comparisons of the two show excellent agreement in replicating both line position and intensities.

Our line list in the ν_1 and ν_5 band regions can be validated by simulating absorption cross sections in the the 2.7 μm region and comparing against PNNL-IR data [10]. Figure 5.13 illustrates how the structure and positions are in good agreement with the overall integrated intensity for APTY in this region being 3% stronger than PNNL. The improvement given by the H2O2-2016 PES can be demonstrated by comparing with the previous room temperature line list and with the PNNL-IR data; see Figure 5.14. The wavelengths of the largest two peaks in this band at 2.738 μm and 2.736 μm are correctly reproduced by APTY but are shifted by about 0.001 μm for H2O2-2015, showing the improvement in this band due to use of the refined PES. Overall the integrated cross-sections for this band differs only by 3% from PNNL; indeed the entire spectrum up to 6,000 cm^{-1} only differs by 3%.

Band intensities can be computed by explicit summation of all transitions within a band and compared against available data. Table 4.7 shows that for the limit available empirical band intensities we agree with most regions to $\leq$7.3%

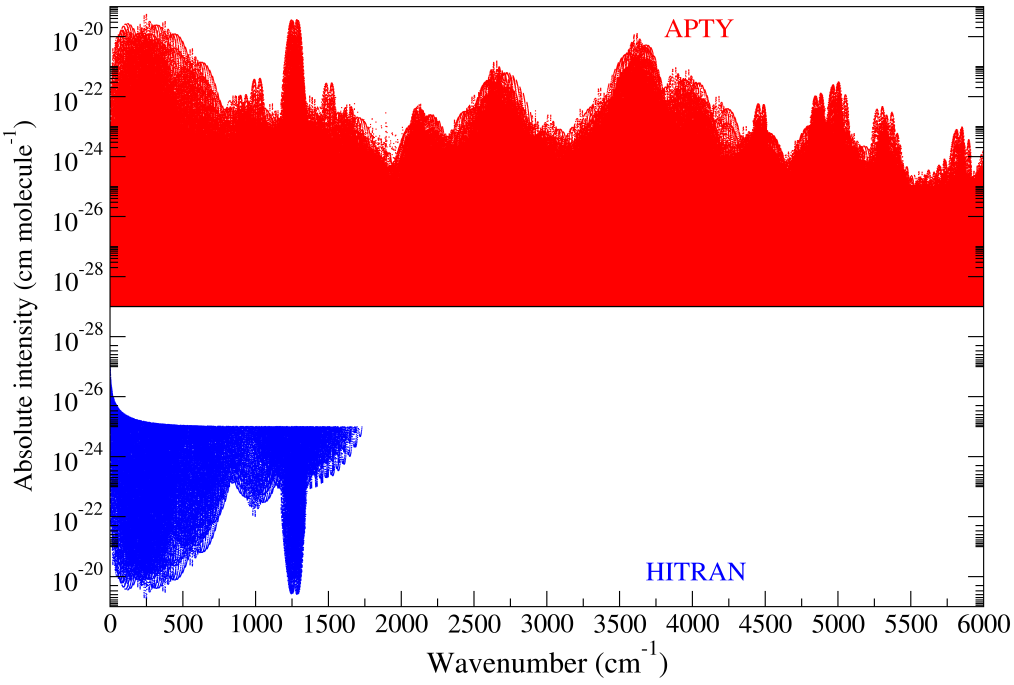

Figure 5.11: Overview of the synthetic spectrum at $T = 296$ K against HITRAN data [6].

which is below the estimated experimental uncertainty of $\approx \pm 10\%$. Two discrepancies are with the ν_6 band from Perrin et al. [174] and the $\nu_2 + \nu_6$ band from Johnson et al. [190]. The former conflicts with other measurements due to Johnson et al. [190] and Klee et al. [193] where integrated absorption intensities were measured directly, whilst the band intensities of Perrin et al. [174] were obtained by summing a synthetic spectrum of only 27 276 transitions. The $\nu_2 + \nu_6$ band intensity is 33.78% weaker than the experimentally derived value. Johnson et al. [190] suggests that the assignment of this band is $\nu_2 + \nu_6$ compared to the $\nu_2 + \nu_3 + \nu_4$ assignment by Giguere [194]. This is based on a Q-branch peak observed at 2,658.62 cm^{-1}. The assignments from APTY suggest that the peak observed is actually a convolution of Q-branches of the $(0,4) \rightarrow \nu_2 + (0,4)$, $(0,4) \rightarrow \nu_2 + (0,2) + \nu_6$, $(0,4) \rightarrow \nu_3 + (4,4)$, $(0,4) \rightarrow \nu_3 + (2,2)$, $(0,4) \rightarrow 2\nu_3 + (2,3)$ and $(0,4) \rightarrow \nu_3 + (2,2) + \nu_6$ transitions with an average separation between them at $\approx$0.02 cm^{-1}. Computing the band intensities of all of these bands in this region give an answer that agrees with value given by Johnson et al. [190] to 3.97%. The Kitt Peak Archive provides FTIR spec-

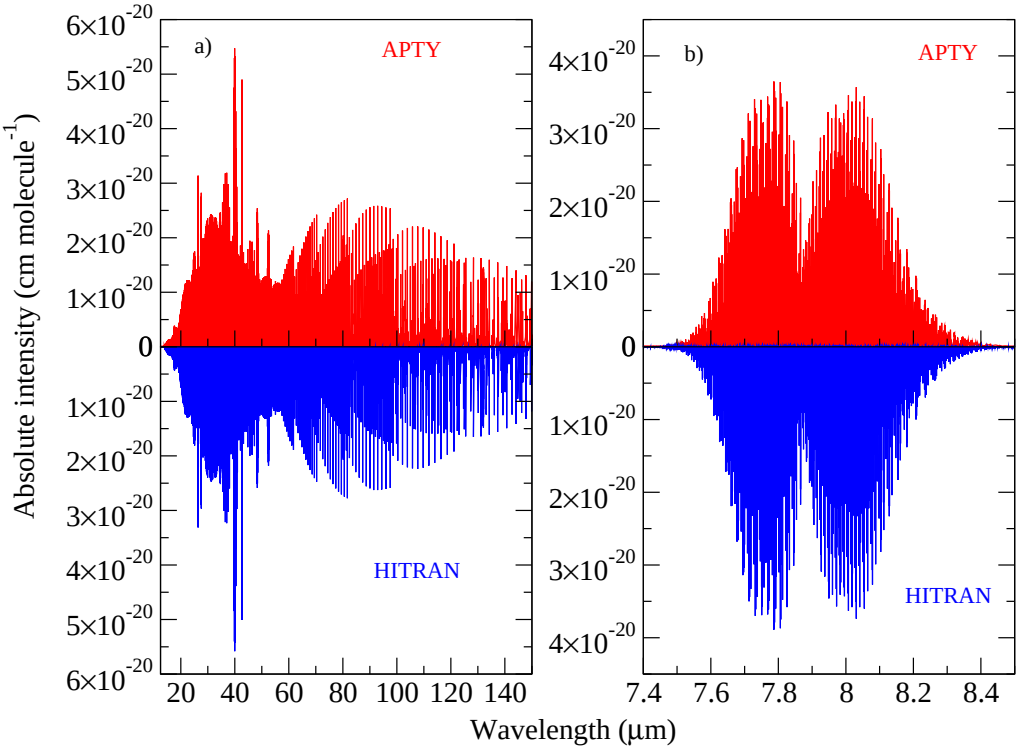

Figure 5.12: The fundamental bands compared to the HITRAN database [6] at $T = 296$ K. (a) Torsional band (b) ν_6

tra of H_2O_2 covering the wavenumber region up to 6422 cm^{-1}, which is only partly assigned. Figure 5.15 (lower display) shows an uncalibrated spectrum of H_2O_2 in the 1.78 μm region (800628R0.002) recorded by R. H. Hunt in 1980, which covers the ν_1 and ν_5 fundamental bands of the hydrogen peroxide. These two bands have not been spectroscopically analysed. The upper display of this figure presents the APTY absorption spectrum at 296 K simulated using the Doppler line profile. The APTY synthetic spectrum resembles all the main features of the experimental data. It is hoped that this will encourage a spectroscopic analysis of the Kitt Peak H_2O_2 spectra in the IR and near-IR regions currently not present in HITRAN. APTY's capability of providing absolute intensities and quantum numbers can assist in the assignment of these spectra.

Figure 5.16 presents integrated absorption cross-sections computed using a Doppler profile [195] for a range of temperatures. The figure shows how the opac-

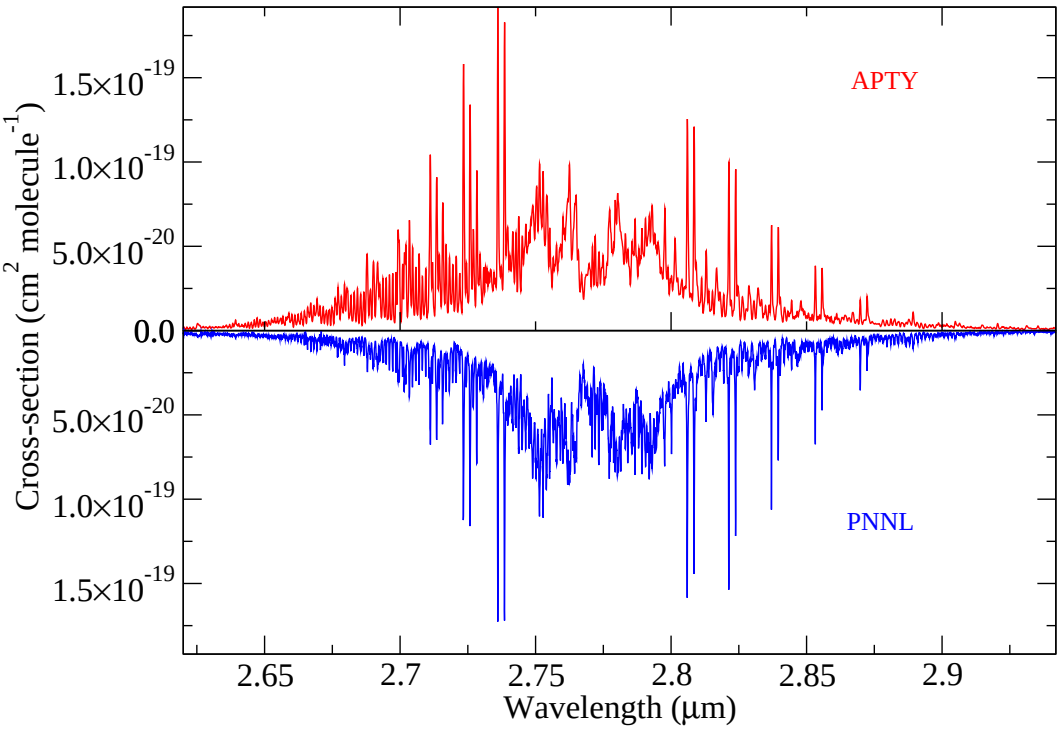

Figure 5.13: The ν_1 and ν_5 band region with APTY against PNNL-IR data at 323.15 K [10] with HWHM = 0.300 cm^{-1}

Table 5.8: Band intensities, in 10 $^{-17}$ cm $^{-1}$/(molecule cm $^{-2}$).

Band	Freq range (cm^{-1})	Ref.	Obs	Calc	(O-C)/O (%)
Torsional	0–1,427	[24]	4.0400	3.7450	7.3
ν_3	750–1,100	[190]	0.0157	0.0165	-5.43
ν_6	1,135–1,393	[190]	1.7458	1.7651	-1.10
ν_6	1,170–1,380	[193]	1.8500	1.7633	4.68
ν_6	1,170–1,380	[174]	1.0030	1.7633	-75.80
$\nu_2 + \nu_6$	2,300–2,900	[190]	0.0830	0.055	33.78
Multiple bands	2,300–2,900	[190]	0.0830	0.0797	3.97
ν_1,ν_5 region	3,300–3,800	[190]	0.8356	0.8724	-4.40

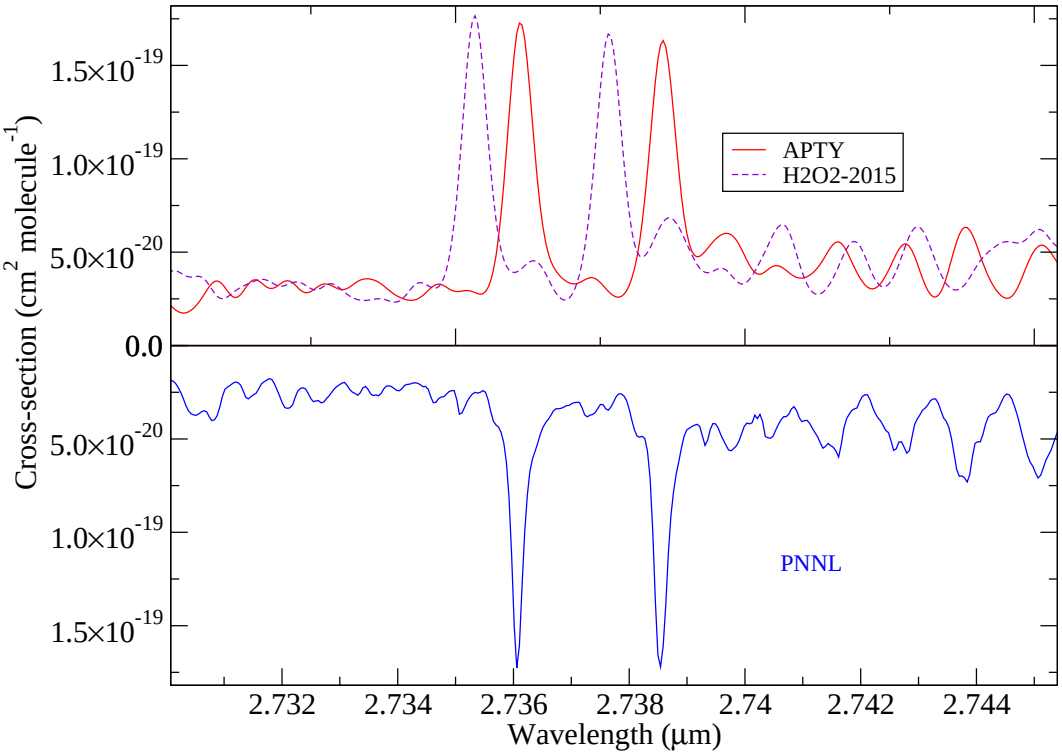

Figure 5.14: Cross-section comparison of peaks in the ν_1 and ν_5 band region with APTY (this work) and H2O2-2015 [16] against PNNL-IR data at 323.15 K [10] with HWHM = 0.300 cm^{-1}

ity changes with increasing temperature. We note the particularly dramatic effect raising the temperature has on the absorption by H_2O_2 in the 13.7 μm region. This smoothing in the overall spectra can only be modelled if there is adequate coverage and population of rotationally and vibrationally excited states. We also note the strength of the OH stretch feature at about 2.75 μm; these features are absent from line databases such as HITRAN because of the absence of assigned spectra in this region. Hopefully APTY can be used to help analyse spectra in this region, as the BYTe NH_3 line list is being used to analyse ammonia spectra [197, 198].

The APTY line list has been published [196] and can also be found in the location listed in Appendix A.

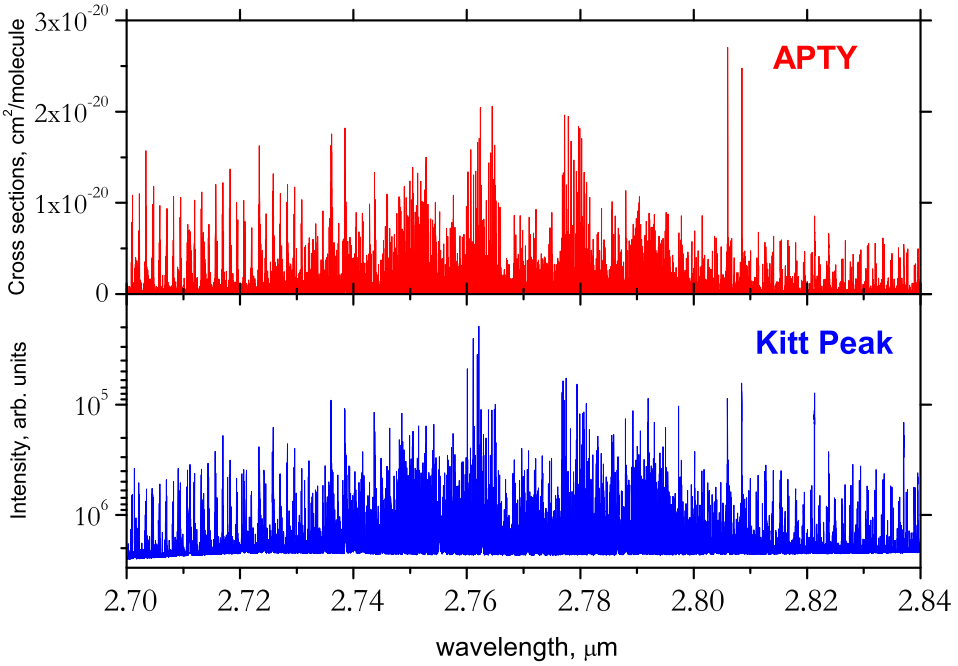

Figure 5.15: The 2.77 μm band of H_2O_2 at room temperature. Upper display: APTY cross-sections (296 K) generated using a Doppler profile; Lower display: an uncalibrated Kitt Peak spectrum of H_2O_2 (Archive Name is 800628R0.002, Date is 28/06/1980; Range is 1599.010271–6421.422198 cm^{-1}, Observer is Hunt; 1.77 m cell).

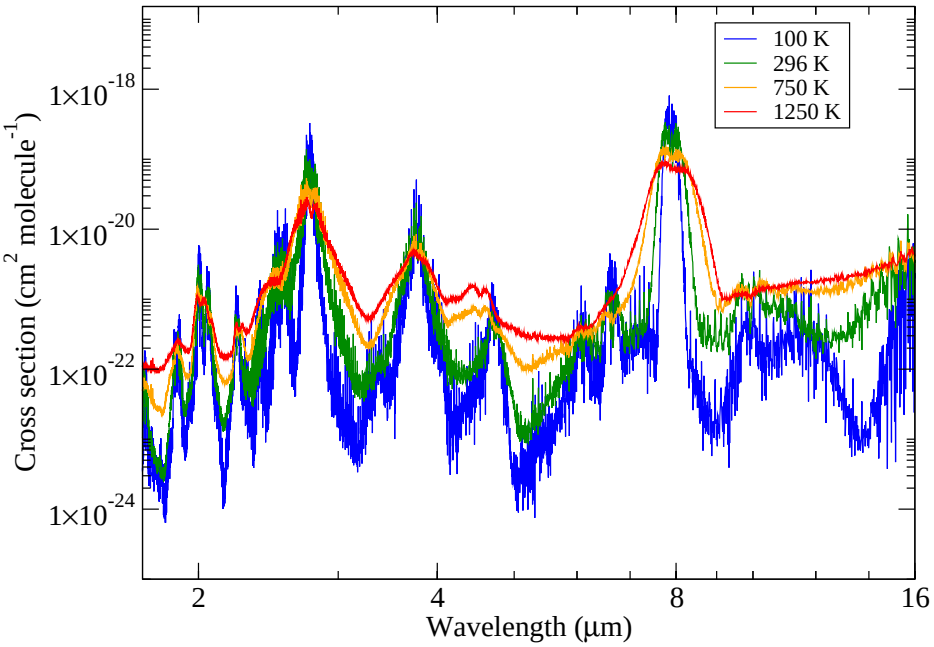

Figure 5.16: Cross-sections using a Doppler profile for the APTY line list as a function of temperature.

Chapter 6

Modelling Cross Sections and Opacities

The production of three molecular line lists and the computation of transition intensities for phosphine were completed within the allotted three years. During the spare time that was afforded, other interesting opportunities in improving computational efficiency were explored within the ExoMol Project. In particular, the impressive results presented by GAIN has inspired an endeavour in applying the same approach into other areas of spectroscopic modelling. As previously mentioned in Chapter 1, computing the opacities of a molecule requires integrating the contribution of transition intensities of a molecule for a specific frequency. Here, the frequencies are binned on an equally spaced grid for the range that we wish cover. If a grid of N points is constructed for a frequency range from $\tilde{\nu}_{\rm min}$ to $\tilde{\nu}_{\rm max}$, then the frequency at grid-point i is:

$$\tilde{\nu}_i = i\Delta\tilde{\nu} \tag{6.1}$$

where $\Delta\tilde{\nu}$ is defined as:

$$\Delta\tilde{\nu} = \frac{\tilde{\nu}_{\rm max} - \tilde{\nu}_{\rm min}}{N} \tag{6.2}$$

and represents the 'resolution' of the generated spectra. Generally this should match the experimental resolution that one wishes to model. If we desire a resolution of 0.001 cm^{-1} for the spectral range of the BT2 [29] H_2O line list (i.e 0.0 – 30,000 cm^{-1}) then $N = 3,000,000$. The process of computing the opacities requires evalu-

ating for each transition j at each grid point i:

$$\sigma_{ij} = I_j F(\tilde{\nu}_j - \tilde{\nu}_i) \tag{6.3}$$

where I_j corresponds to the absorption intensity of a transition with frequency $\tilde{\nu}_j$ computed from Eq. (3.7). The function F is the line profile, this was briefly mentioned in a qualitative manner in the introduction chapter. For brevity we will only discuss the Voigt profile as it is the most commonly used, this has the form:

$$F(x,y) = \frac{y}{\pi} \int_{-\infty}^{\infty} \frac{e^{-t^2}}{y^2 + (x-t)^2} dt \tag{6.4}$$

where we have defined $x = \sqrt{\ln 2} \frac{\tilde{\nu}_j - \tilde{\nu}_i}{\gamma_D}$ with Doppler width γ_D and $y = \sqrt{\ln 2} \frac{\gamma_L}{\gamma_D}$ with Lorentzian width γ_L. Here the exponential part comes from the Doppler contribution and the fractional comes from the Lorentzian. The convolution of the two requires integrating their combination. The solution to this integral is the real part of the complex error function often referred to as the Faddeeva function:

$$\begin{aligned} z &= x + iy \\ w(z) &= \mathrm{erf}(z) \\ F(x,y) &= \Re[w(z)] \end{aligned} \tag{6.5}$$

Both complex and real solutions have been the subject of various articles to numerically approximate [199, 17, 200] quickly.

The total opacity σ_i at a grid point i is the sum contribution of all transitions:

$$\sigma_i = \sum_j \sigma_{ij} \tag{6.6}$$

considering there are around half a billion transitions in BT2 and over tens of billions in other ExoMol line lists, this calculation poses a staggering computational cost. In order to address this problem, we will take step through a typical calculation and note the time taken for each transition:

- Read transition information t_r
- Call Voigt function t_f
- Compute Voigt on all bins t_v

the time taken for N_t transitions is therefore:

$$T = (t_r + t_f + t_v)N_t \tag{6.7}$$

of course the contribution of t_r is related to I/O and there is little that can be improved as each transition is used once, therefore, assuming that the actual calculations take no time, the minimum time to complete this is:

$$T_{\min} = N_t t_r \tag{6.8}$$

however the Voigt profile evaluation does contribute to the time, even if $t_v < t_r$. If we assume that $t_v = \frac{t_r}{2}$ (i.e the calculation takes half as long as the reading), then $T = 1.5T_{\min}$, if reading takes 4 hours, then the total time will be 6 hours, 8 becomes 12. The question is: what can be done to remedy this? Whilst it may be obvious to simply improve the calculation of the Voigt function, this is an extremely grand endeavour as it will require a radically new method of evaluation and extremely efficient implementation to approach the minimum time. Another commonly used method is to reduce the number of transitions, and this can be done by setting an intensity cut-off. However, such a cut-off is arbitrarily decided and may risk opacity loss from weak but numerous transitions. Additionally, this doesn't reduce the IO time as one still needs to read the transition first. Instead it may be more interesting to ask: What can be done to hide this?

Let us assume that we implement a system in such a way that we can read the next transition whilst simultaneously computing the Voigt profile for the previous, the time function becomes:

$$T = [t_r + \max(\{t_f + t_v\} - t_r, 0)]N_t \tag{6.9}$$

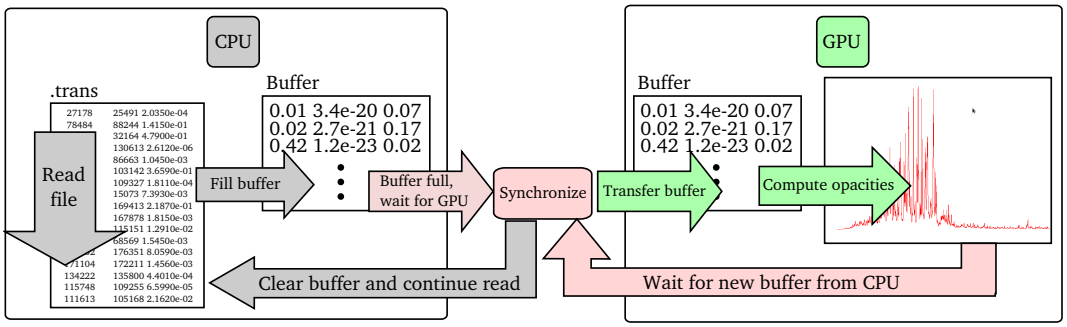

Figure 6.1: The double buffering system used by GEXS. Here the CPU and GPU are working simultaneously but must synchronize when the CPU buffer is full and when the GPU has completed its computation.

where the max function will select the higher of the two values (in the case where the expression is negetive, it will be 0). Here, if $t_v < t_r$ (assuming t_f is zero), then the time taken is the minimum. In a sense, the reading masks as much of the Voigt calculation as possible and can reduce the 'apparent' calculation time. Therefore, the requirement for minimum runtime is that the Voigt profile calculation takes as long as 'read', which is a possiblity with GPUs.

6.1 GEXS

GEXS (**G**PU **E**xoMol **X**ross **S**ections) is a hybrid CPU+GPU code that utilizes this aspect to compute opactities for extremely large line-lists without an intensity cut-off. The key to the code relies on a 'double buffering' system not dissimilar to the technique used in generating computer graphics. The system works by having two equally sized buffers, one in the CPU and one in the GPU. The CPU will continously read until it fills its own buffer until it exhausts the space at which point it hands it to the GPU buffer. The GPU performs work on its own buffer while the CPU simultaneously refills its own. A synchronization step must occur, the CPU after filling its buffer must wait for the GPU to finish its calculation and the GPU, after finishing work, must wait for the CPU to supply it a new buffer. Figure 6.1 shows a graphical description of this method. Technically it's possible to pass transitions immediately to the GPU as its read. The problem comes from t_f, the time taken to call a function. Each call to the GPU incurs a slight overhead, and this

overhead can contribute signficantly to the run time and even dominate the Voigt profile calculation itself if numerous. Therefore buffering reduces this by batching the transitions into a single call. The actual Voigt implementation uses the Humlicek method to compute, it is fairly fast and its results can be compared to currently available codes in the ExoMol project which aids in debugging.

GEXS can also perform multiple temperature and pressure combinations at the same time through usage of streams. Generally, there is good performance with around two to three simultaneous temperature-pressure calculations whilst still achieving ≈40% of the minimum runtime. This is offset by the fact that we are achieving 2-3 times throughput in calculation. Table 6.1 shows five different runs in calculating the cross sections for the half a billion transiions of the BT2 [29] line list with 3,000,000 grid-points in the 0.0 – 30,000 cm^{-1} spectral range. The system used was a 20 core Xeon(R) CPU E5-2687W running at 3 GHz with 256 GB memory and a single nVidia K40 GPU. The read-only time is considered T_{min} of the calculation and refers to reading the transitions without calculation. The OpenMP serial run describes the typical run as described previously with a 10^{-30} cm molecule^{-}1 absolute intensity cutoff and where the actual Voigt uses OpenMP. Finally the GEXS runs uses no cutoff and describes the number of temperature-pressure (T-P) combinations being done simultaneously. Here we see that the GEXS performs better than the CPU serial read version in all cases and the single (T-P) is extremely close to minimum time. Whilst there is a reduction in performance with multiple T-P combinations, the time for a single T-P actually reduces. The 2 and 3 T-P computation effectively does 1 T-P in 731.83 and 680.58 seconds respectively, which is less than the minimum time. Essentially we get more value performing multiple T-P combinations for each read. For the simpler profiles such as Doppler, up to 10 different temperatures can be computed with almost no compromise in time taken. Figure 5.16 was computed with GEXS for the entire APTY line list using 6,000 grid points in the 0.0–6,000 cm^{-1} spectral range. Here, 10 different temperatures were calculated at the same time on the same run with a run-time ≈99% of the minimum. Four of the 10 cross sections were selected for the plot.

Table 6.1: Table showing various times to calculate 3,000,000 grid points for the full BT2 [29] line list. 'Read only' refers to purely IO time. OpenMP serial is an implementation to calculate the Voigt profile using OpenMP but in a serial reading fashion. Here a 10^{-30} cm molecule^{-1} intensity cutoff was used. The GEXS times uses no cutoff and each describe the number of temperature-pressure combinations being calculated. The minimum ratio describes the ratio between the minimum completion time (Read only) and the run time of the calculation. The speed up is in relation to the OpenMP version.

Calculation	Time (s)	Minimum ratio	Speed Up
Read only ($T_{\min}$)	786.56	1.00	14.29
OpenMP Serial read (10^{-30} cutoff)	11242.80	0.07	1.00
GEXS 1 T-P (No cutoff)	818.97	0.96	13.73
GEXS 2 T-P (No cutoff)	1463.66	0.54	7.68
GEXS 3 T-P (No cutoff)	2041.76	0.39	5.51

6.2 The current state of GEXS

As of the time of writing, the GEXS code is in the prototype stage of development and can be viewed in the location given in Appendix A. Currently the code suffers a severe performance drop with extremely large numbers of grid points (≈30,000,000) causing the CPU to stall frequently waiting for the GPU to finish. Multiple GPUs do alleviate this issue as it gives each GPU more time to complete their calculations before a new buffer is filled. The problem is due to the fact that each grid point requires a large range of transitions to be considered and therefore increases the memory reads and Voigt evaluations. Steps are being taken to introduce higher degree of caching to reduce the memory reads but so far this has paradoxically increased the run-time with no obvious explanation. A more thorough debugging and profiling process should reveal the issue but in theory this caching methodology should give better scaling. Additionally an intensity cut-off could be implemented to reduce the number of transitions for each grid point but this goes against the philosophy of the code.

In terms of evaluating the Voigt profile, the Humlicek algorithm [199] was only used to ensure that the buffering system worked by comparing with other ExoMol codes. However it is fairly inefficient in a GPU setting and unnecessarily calculates the imaginary parts. Two more promising methods are currently being implemented and tested. The first is a 10–40 point Gauss-Legendre Quadrature integration; this

is inspired by the ExoCross [201] code and boasts low numerical error in all regions and seems well suited to GPU use. The second method utilizes a combination of Algorithm 916 by Zaghloul and Ali [17] and a 3 point Gauss-Hermite quadrature similar to the methodology used by Grimm and Heng [202]. Its efficient implementation and tuneable errors has had promising results and improvements to run time but is suffering from opacity loss due to a possible error in the code (see Figure 6.2). Additionally, GEXS utilizes the GPU exclusively which is wasteful as there

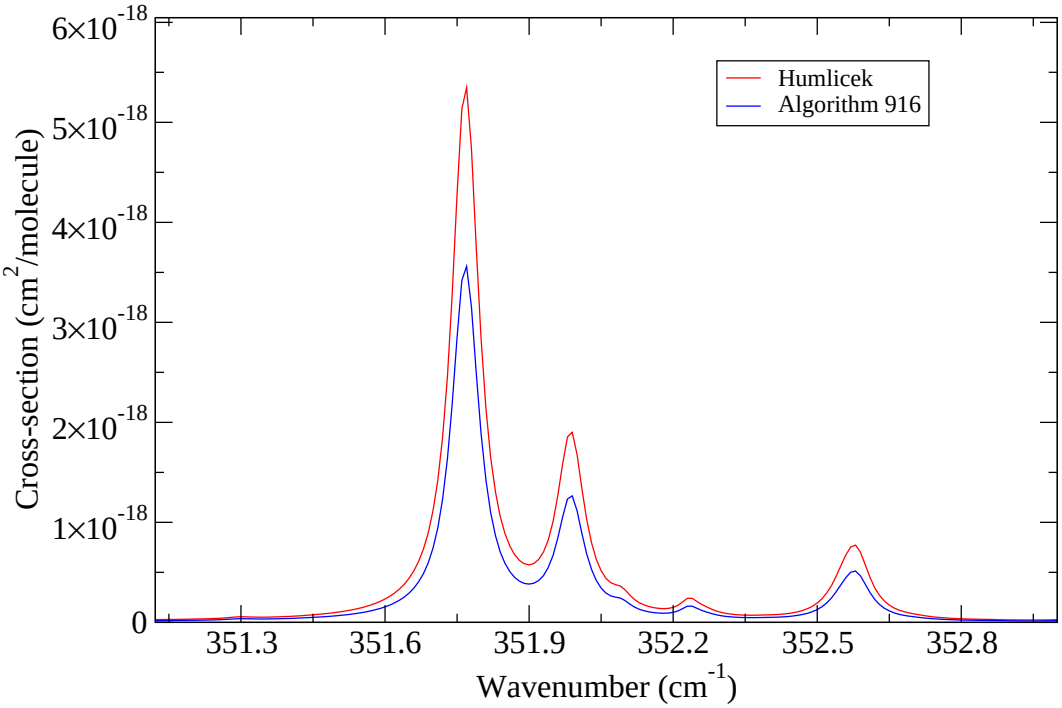

Figure 6.2: A comparison between the Humlicek method and Algorithm 916 [17], whilst improving run-time, there is a coding issue which is giving reduced opacity.

are idle CPU cores which may be commandeered to perform work. However, using OpenMP directives to split the reading and computing cores apart is difficult without affecting code readability and organization. Therefore a prototype was developed using the C++ 2011 standard `std::thread` class. This has been promising to use with a single temperature pressure combo as it also gives the GPU ample time to conduct its calculations but the current implementation causes far too many threads to be spawned reducing the throughput of calculations. An implementation using a job queue is being considered to alleviate this problem.

Lastly the current version of GEXS does not comply with the new ExoMol format [203] standard so further work must be done in order to support future line lists.

Chapter 7

Discussion and Conclusion

Simulating atmospheres of hot exoplanets requires tremendous amounts of spectroscopic data. The ExoMol project attempts to solve this problem by providing a database of these much needed molecular opacities constructed through theoretical means. In particular, the currently available data for formaldehyde and hydrogen peroxide lacked both spectral and temperature coverage, hindering their usage. Both of these molecules are prime candidates for production of their hot line lists. However, this requires the computation of millions of states and tens of billions of transitions. Such a task is not insurmountable but requires not only the application of well established methodologies, but also the development of new ones. The work in this thesis not only contributes high-accuracy line lists for these molecules, but also a new state-of-the-art code for computing transition linestrengths.

GAIN started off as small code to speed up the computation of H_2CO transitions by experimenting with GPUs. It has escalated into becoming a complementary code to TROVE. The task of porting the FORTRAN code into CUDA C was non-trivial and required intimate knowledge of the hardware in order to perform properly. However, by exploiting their huge number of cores, their smaller but faster cache memory and TROVEs basis set arrangement, a substantial increase in performance was achieved with over $1000\times$ speed-up in the preprocessing step of the calculation. Improvements in computational efficiency was achieved by using multiple GPUs to perform the linestrength calculations and MPI to cache all of the large eigenvectors while providing an even greater number of GPUs to perform

work. This new code dubbed GAIN-MPI was utilized extensively in completing the SAlTY hot PH_3 line list and, for example, turned the $J = 21 \leftrightarrow 22$ transitions from a 1.5 month calculation into less than a few hours. GAIN-MPI has since been integrated into the ExoMol project ecosystem and has been used to compute over 70 billion transitions for a range of molecules. The code will see further usage in future molecular line lists including C_2H_2, C_2H_4, NH_3 and CH_4. The task of using accelerators seems impregnable, but newer programming standards such as OpenACC (**Open Acc**elerators) aim to provide the power of accelerators with the simplicity of OpenMP. The maturity of these standards will see an exponential rise in their usage in a range of scientific fields.

The first line list produced was for hot formaldehyde. A previously calculated semi-empirical PES and new *ab initio* DMS was used in conjunction with TROVE to variationally compute high-accuracy line positions and intensities. H_2CO differs from the usual ExoMol system of producing polyatomic line lists as no room temperature version using an *ab initio* PES was produced first. This procedure differs from previous line lists produced with TROVE such as PH_3 [108] and SO_3 [204]. The first line list that was computed, AYTY-0, contained 5 million energy states and over 3 billion transitions up to 9,600 cm^{-1} and applicable to 1000 K. However, AYTY-0 suffered from a major flaw in its line strengths that completely destroyed its viability for spectroscopic usage. It is interesting to imagine that if we had followed the usual production schedule, it may have allowed us to fix the issues due to refinement that was encountered with AYTY-0 early on. However, this in the end has actually benefited formaldehyde and in turn, future polyatomic line lists to be produced. Common knowledge dictates that the PES is responsible for the line positions and the DMS is responsible for the transition intensity. However, what occurred in AYTY-0 demonstrates that the procedure of refining the PES also affects the line strengths as well, even when all the energies seem physically appropriate. Knowledge of this means that future line lists, especially molecules with little experimental study, can avoid this problem altogether during refinement of the PES. Additionally, the time lost in its production allowed for new MPI di-

agonalisers and GAIN to be developed. This meant that significantly more states and transitions could be computed producing a 'hotter' hot line list. The *ab initio* PES was re-refined to produce a new semi-empirical PES that gave an rms range of 0.001 cm^{-1} to 0.2 cm^{-1}. The resulting calculations with TROVE and GAIN produced over 10 million states and around 14 billion transitions. This is a significant increase compared to the original and as a result, pushed the maximum frequency from 9,600 cm^{-1} to 10,000 cm^{-1} and the temperature applicability from 1,000 K to 1,500 K.

Fortunately, H_2O_2 did follow the usual production pattern and an *ab initio* PES was used to produce a room temperature line list. Whilst no real problems were encountered in the actual production of the line list, the hydrogen peroxide molecule is a far more complicated system than formaldehyde. The internal torsional motion and tunnelling between configurations results in extremely small splittings. These splittings have caused a number of issues in interpreting its spectrum, in particular those involving the ν_1 and ν_5 bands. Therefore the treatment of the torsional motion in the variational calculation required a significantly larger basis set in order to accurately resolve the splitting. Retaining high accuracy with an *ab initio* PES required using the empirical band shift method in order to bring the calculated band centres in line with the experimental. An rms deviation of 0.001 – 0.1 cm^{-1} was easily achieved with no modification to the potential. The calculation of up to $J = 40$ states and 1.4 billion transitions resulted in a temperature applicability of 296 K and a spectral range of up 8,000 cm^{-1}. The spectral completeness only reaches up to 4,000 cm^{-1}, however, and any bands past this point are incomplete in structure. The benefit of such an extended range is as a guide to unobserved spectral regions of interest and to aid in interpretation in experimental studies.

Producing a high accuracy hot line list for H_2O_2 required far more than performing a band shift. The problem is that for the bands where a definite observation is unavailable, they gain no benefit in their line positions through this procedure. This means that their energies are essentially at *ab initio* levels of accuracy. One of the uses of a H_2O_2 line list would be to aid in definitively assigning bands such as ν_1

and ν_5. In order to improve these bands, a refinement procedure was performed on the PES with 410 experimentally observed energies. This new semi-empirical PES 'H2O2-2016' boasts an rms close to 0.001 cm^{-1} without shifting on band centres and a significantly improved rms of 0.064 cm^{-1} at higher rotational excitations as well. The predictive ability has improved significantly and compares extremely well to experimentally observed opacities compared to the shifted *ab initio* PES. Whilst the majority of the variational calculation did not change, the maximum J and energy were increased to 85 and 12,000 cm^{-1} respectively. This resulted in the APTY line list with over 7 million states and over 20 billion transitions. The spectral and temperature completeness reaches up to 6,000 cm^{-1} and 1250 K respectively. As of time of writing, the publication of this line list is currently in press. It will be exciting to see the future usage of APTY in the scientific community.

7.1 Further work

Whilst much has been accomplished concerning the computation of theoretical line lists. There are still plenty of opportunities for future projects and work to explore the utilization of both AYTY and APTY line lists as well as tackling the challenge of producing line lists for larger molecules.

7.1.1 Experimental Assignments

As previously discussed throughout this thesis, there is a limited amount of experimental data for both H_2CO and H_2O_2. Both the AYTY and APTY line list can aid in this aspect by suggesting regions of interest for experimental observation as well as providing a robust means of assigning complicated spectra and line intensity. For formaldehyde, one of the biggest deficiencies is the lack ν_3, ν_4 and ν_6 bands intensities. However, Reuter et al. [13] provides an extensive set of line positions for the ν_3,ν_4 and ν_6 bands. As the experimental PNNL opacities agree to about 10% with AYTY (and by Eq. (6.6) the absolute intensities in general), the experimental line positions can be matched with the AYTY transition intensities using quantum number assignments. This combination would complete the set of fundamental bands for H_2CO and allow the experimental data to become integrated into the HITRAN

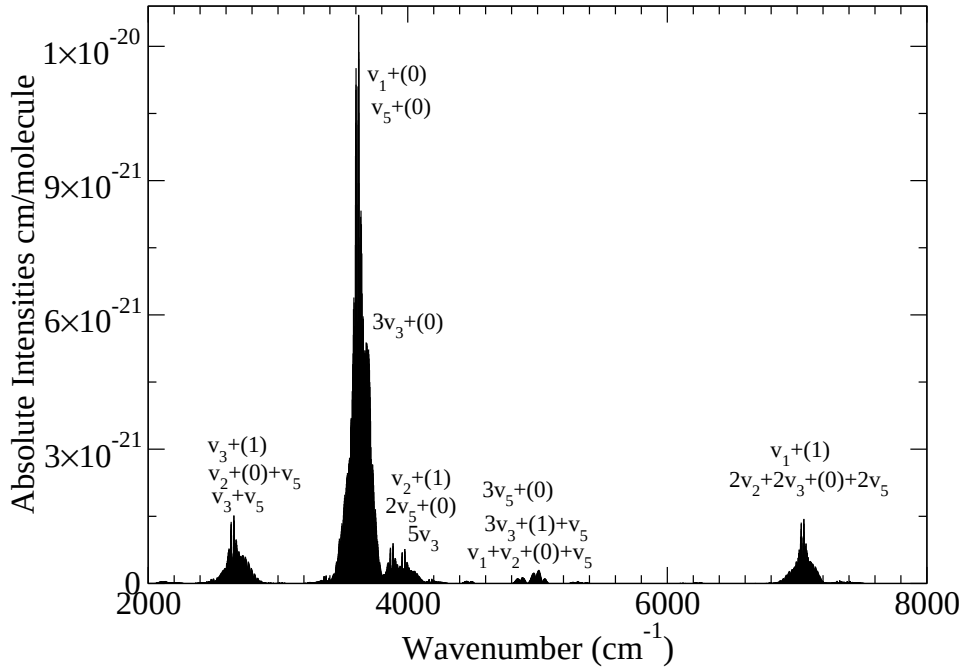

Figure 7.1: A simulated 296 K spectrum up to 8000 cm^{-1} for H_2O_2 using the room temperature line list. Preliminary assignments are shown with brackets indicating torsional excitation *n*.

database. Additionally, there are further overtone and combination bands worth investigating that have been hinted at by both AYTY and the PNNL[10] database that include the $2\nu_3$ and $3\nu_4$ at 3472.22 cm^{-1}, $\nu_1+\nu_2$ and $\nu_2+\nu_5$ at 4545.12 cm^{-1}, $4\nu_2$ at 5181.1 cm^{-1} and $2\nu_5$ at 5649.0 cm^{-1} as well as the $\nu_5+3\nu_6$ bands at 6361.3 cm^{-1} by Zhao et al. [14]. Hydrogen peroxide past 2000 cm^{-1} contains many strong bands that require further experimental observation. Figure 7.1 shows a simulated spectrum using the room temperature line list, this was chosen over the hot line list due to the larger spectral range and easier computation. Highlighted are the major bands with preliminary assignments. Included in this range are the ν_1 and ν_5 bands fundamental bands as well as a range of overtones and combination bands. The most interesting are the ν_3 and ν_1 torsionally excited bands at 2,700 cm^{-1} and 7,060 cm^{-1} which display similar band structures. An experimental study in combination with the APTY line list and the extended line list should provide a more robust analysis, assignment and absolute intensities to observed line positions in this region for integration into molecular databases.

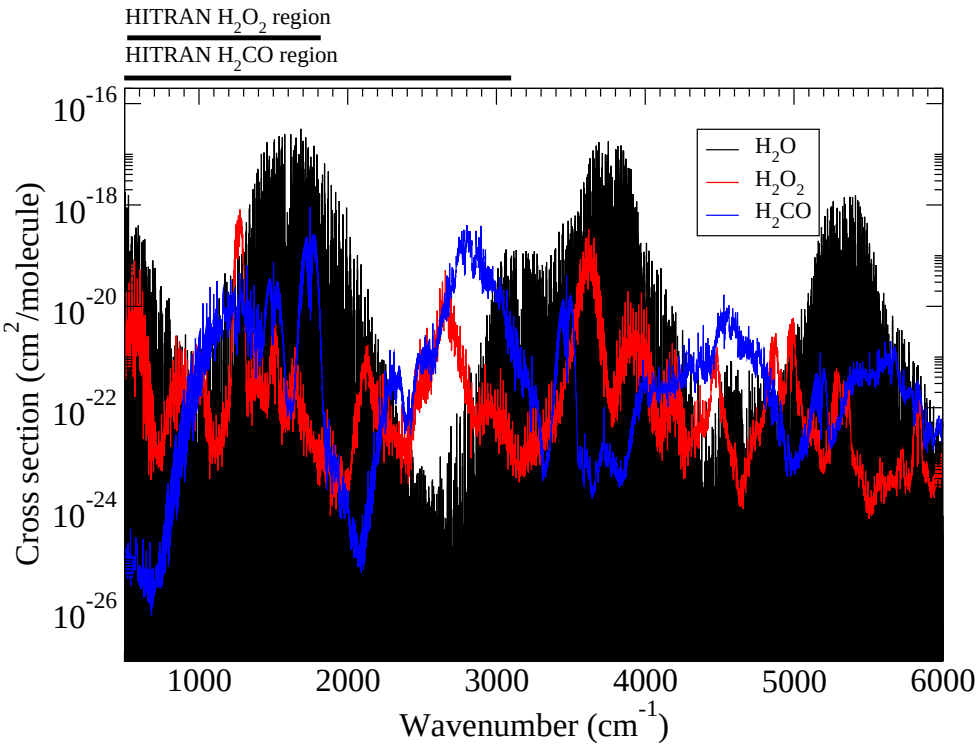

Figure 7.2: Cross sections of water (black) formaldehyde (blue) and hydrogen peroxide (red) using the BT2, AYTY and APTY line lists with a Doppler profile at 296 K. Highlighted above are the spectral regions that HITRAN covers for H_2O_2 and H_2CO.

7.1.2 Atmospheric and Industrial Applications

As previously discussed in the introductions of Chapters 4 and 5, both hydrogen peroxide and formaldehyde molecules are trace species in the Earths atmosphere. IR studies of H_2O_2 and H_2CO in the terrestrial atmosphere can benefit from the produced molecular line lists. In particular, the identification of features that are not masked by water absorption bands. Figure 7.2 are plots of cross sections for water, formaldehyde and hydrogen peroxide using a Doppler profile at 296 K. The masking effect of water is evident, in particular, almost the entire HITRAN spectrum of H_2O_2 is hidden. However, both molecules have strong absorption features that reside in the 'gaps' of waters IR spectra in the 3,000 cm^{-1} and 4,500 cm^{-1} regions. These regions, in particular the latter, should be investigated as a means of conducting IR studies of both of these molecules in the Earth's atmosphere as well as ground based

IR observations of astrophysical objects.

In combustion studies, the assessment of new biofuels [205] as an alternative to hydrocarbons requires monitoring their byproducts. Formaldeyhde is a toxic byproduct of combustion and its concentration in burning is usually monitored by its UV bands[206, 207] using the highly sensitive cavity ring-down spectroscopy (CRDS) technique. However, broadband absorption in the UV can complicate exact observation of features[207]. Recently an alternative to CDRS are Quantum Cascade Lasers (QCL) observations in the mid-infrared region. A comparison of UV-CDRS and MIR-QCL by Nau Patrick and Andreas [208] on a 1100 K flame demonstrated that MIR-QCL provides similar sensitivity in measuring formaldehyde concentration with a simpler equipment setup and easier to identify absorption features. The comparison was done using a simulated spectrum from the HITRAN database in the 1694 cm^{-1} region in combination with the BT2 water line list. This region was chosen in particular due to the simpler ν_2 spectral features at 1100 K. Further studies using the AYTY line list may enhance the sensitivity of MIR-QCL by measuring and identifying the more complex bands at the 3000 cm^{-1} and 4500 cm^{-1} that have almost no interference from water absorption.

In medical fields, vaporised H_2O_2 is effective at sterilizing equipment[209] making its trace-gas-sensing important in minimize exposure risk. Additionally, hydrogen peroxide is a byproduct of tissue damage and inflammation[210] and as such can be used as a possible biomarker in breath condensate [211] to non-intrusively diagnose lung and throat damage. As such, there is a push towards developing new MIR lasers[212] specifically for sensing hydrogen peroxide for both decontamination and medical settings. Future work with APTY may involve collaborations to assist in developing and benchmarking more sensitive lasers. To quote Nancy P. Sanchez and Tittel [212]: “For future development of mid-IR laser-based sensor systems, it is therefore necessary to explore alternative strong, interference-free H_2O_2 absorption lines”.

7.1.3 Large Molecules, The Curse of Dimensionality and ANGMOL

The number of vibrational basis functions N_b required to converge a vibrational overtone can be estimated with a simple expression given by Pavlyuchko et al. [213]:

$$N_b \approx \prod_{i}^{N_f} \frac{(P_t + N_f + i)}{i} \tag{7.1}$$

Where P_t is a target polyad given by Eq. (2.50) and N_f is the number of vibrational degrees of freedom. Four atom polyatomic molecules have 6 degrees of freedom, therefore to achieve high accuracy for the fifth overtones ($P_t = 5$) requires $\approx$ 8,000 basis functions for the $J = 0$ solution. This is similar to what is achieved with the AYTY line list. To reach full convergence for all polyads ($P_t = 16$) would require half a million basis functions, the act of determining the number of basis functions is an act of balancing the accuracy with computational cost. This expression however suffers greatly from the 'curse of dimensionality'. Simply adding an additional atom increases the degrees of freedom to 9 and to achieve accuracy on the fifth vibrational overtone requires $\approx$817,190 basis functions. Barring exceptional cases such as CH_4, full variational treatment of rotating and vibrating molecules with more than four atoms becomes infeasible. This is an issue as there is interest in hot line lists for hydrocarbons such as ethylene (C_2H_4) and benzene (C_6H_6) in both astrophysical and industrial settings. Overcoming this problem requires implementing a new methodology that can significantly reduce the computational cost of the problem whilst maintaining accuracy. A solution to this comes from the ANGMOL code [18], it utilizes a hybrid variational-perturbation method whereby the large vibrational matrix is partitioned into three sub-blocks as seen in Figure 7.3. Region 1 contains the largest contributions towards the energies, the contributions of region 2 are pertubatively added on to this sub-block whilst all contribution from region 3 are discarded. The smaller sub-block is then stored and diagonalized. This method has seen great success in producing a hot line list for nitric acid (HNO_3) [19] on a single workstation that matches well with experimental observations (Figure 7.4)

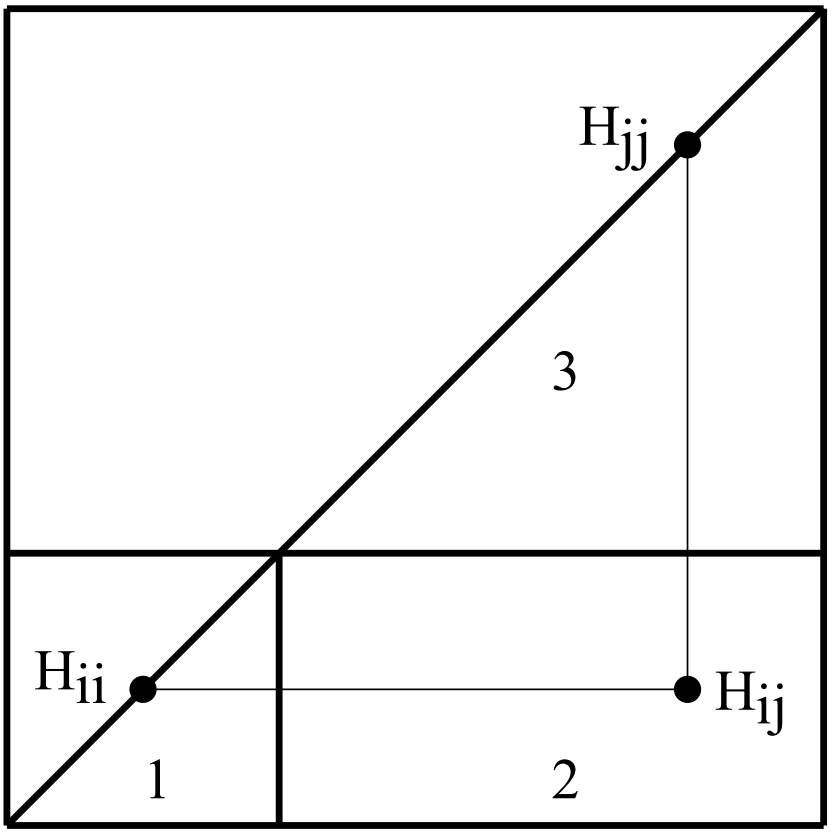

Figure 7.3: Taken from Figure 2 of Pavlyuchko et al. [18]. A vibrational maxtrix split into three sub blocks. Region 1 contains the largest contributions to energy. Region 2 contains small contributions that are added to Region 1 and Region 3 is discarded. The relations between matrix elements are shown.

ANGMOL however is not general like TROVE and exclusively utilizes curvilinear coordinates. Implementing the methods presented by ANGMOL into TROVE will allow for the production of hot line lists for an even greater range of molecules, turning TROVE into an extremely powerful general nuclear motion software suite.

7.2 Conclusion

The greatest limitation with experimental data is that it can only provide what has been observed. Theoretical line lists circumvent this by not only attempting to replicate the experimental data, but also predict it as well. This gives it the advantage of completeness. To experimentally acquire and analyse the tens of billions of

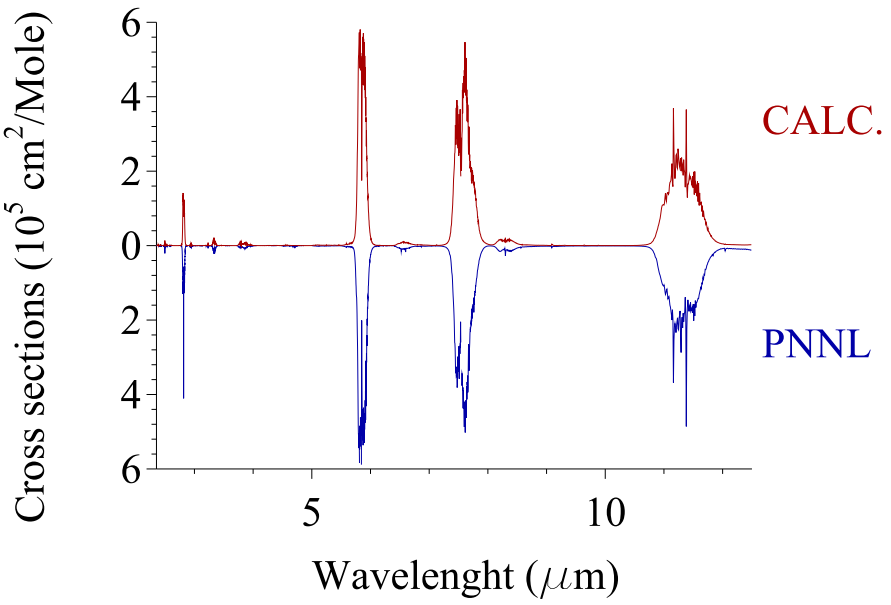

Figure 7.4: Taken from Figure 1 of Pavlyuchko et al. [19]. A comparison between the simulated spectrum from the hot nitric acid line list and PNNL at 296.0 K

transitions needed to compete with theory verges on the impossible. Completeness however comes with a great computational cost. Exploiting to a high degree the hardware currently available is key to pushing theoretical line lists to higher temperatures and greater frequency ranges. The PH_3, H_2CO and H_2O_2 molecular line lists are a testament to what the ExoMol project is capable of achieving in a short timespan.

The works described in this thesis represents three years of effort and required a thorough understanding of quantum mechanics, molecular spectroscopy and atmospheric physics. The usage of high performance computing facilities necessitated an intimate understanding of hardware architecture, programming paradigms, accelerators and parallelization in order implement methodologies efficiently followed by the analysis of terabytes of computed data.

Appendix A

Supplementary Data

Included in this appendix is the location of data produced for each of the listed Chapters.

A.1 Chapter 3

- GAIN-MPI source: `https://github.com/ahmed-f-alrefaie/GAIN-MPI/`

- SAlTY line list: `http://www.exomol.com/data/molecules/PH3/31P-1H3`

A.2 Chapter 4

- AYTY hot line list: `http://www.exomol.com/data/molecules/H2CO/1H2-12C-16O`

A.3 Chapter 5

- 296 K line list: `http://www.exomol.com/data/molecules/H2O2/1H2-16O2`

- APTY hot line list: `http://theoryserv2.phys.ucl.ac.uk/data/molecules/H2O2/1H2-16O2/APTY/`

A.4 Chapter 6

- GEX source: `https://github.com/ahmed-f-alrefaie/gpu_cross`

Bibliography

[1] T. M. Brown. Transmission spectra as diagnostics of extrasolar giant planet atmospheres. *ApJ*, 553(2):1006, 2001.

[2] G. Tinetti, P. Deroo, M. R. Swain, C. A. Griffith, G. Vasisht, L. R. Brown, C. Burke, and P. McCullough. Probing the terminator region atmosphere of the Hot-Jupiter XO-1B with transmission spectroscopy. *ApJL*, 712(2):L139, 2010.

[3] I. P. Waldmann, G. Tinetti, M. Rocchetto, E. J. Barton, S. N. Yurchenko, and J. Tennyson. Tau-Rex i: A next generation retrieval code for exoplanetary atmospheres. *ApJ*, 802(2):107, 2015.

[4] D. Deming, A. Wilkins, P. McCullough, A. Burrows, J. J. Fortney, E. Agol, I. Dobbs-Dixon, N. Madhusudhan, N. Crouzet, J.-M. Desert, R. L. Gilliland, K. Haynes, H. A. Knutson, M. Line, Z. Magic, A. M. Mandell, S. Ranjan, D. Charbonneau, M. Clampin, S. Seager, and A. P. Showman. Infrared transmission spectroscopy of the exoplanets HD209458b and XO-1B using the wide field camera-3 on the hubble space telescope. *ApJ*, 774(2):95, 2013.

[5] D. K. Sing, J. J. Fortney, N. Nikolov, H. R. Wakeford, T. Kataria, T. M. Evans, S. Aigrain, G. E. Ballester, A. S. Burrows, D. Deming, J.-M. Désert, N. P. Gibson, G. W. Henry, C. M. Huitson, H. A. Knutson, A. L. D. Etangs, F. Pont, A. P. Showman, A. Vidal-Madjar, M. H. Williamson, and P. A. Wilson. A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion. *Nature*, 529:59–62, January 2016.

[6] L. S. Rothman, I. E. Gordon, Y. Babikov, A. Barbe, D. Chris Benner, P. F. Bernath, M. Birk, L. Bizzocchi, V. Boudon, L. R. Brown, A. Campargue, K. Chance, E. A. Cohen, L. H. Coudert, V. M. Devi, B. J. Drouin, A. Fayt, J.-M. Flaud, R. R. Gamache, J. J. Harrison, J.-M. Hartmann, C. Hill, J. T. Hodges, D. Jacquemart, A. Jolly, J. Lamouroux, R. J. Le Roy, G. Li, D. A. Long, O. M. Lyulin, C. J. Mackie, S. T. Massie, S. Mikhailenko, H. S. P. Müller, O. V. Naumenko, A. V. Nikitin, J. Orphal, V. Perevalov, A. Perrin, E. R. Polovtseva, C. Richard, M. A. H. Smith, E. Starikova, K. Sung, S. Tashkun, J. Tennyson, G. C. Toon, Vl. G. Tyuterev, and G. Wagner. The *HITRAN* 2012 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transf.*, 130:4 – 50, 2013.

[7] A. A. A. Azzam, S. N. Yurchenko, J. Tennyson, and O. V. Naumenko. ExoMol line lists XVI: A Hot Line List for H_2S. *MNRAS*, page (in press), 2016.

[8] J. Tennyson and S. N. Yurchenko. ExoMol: molecular line lists for exoplanet and other atmospheres. *MNRAS*, 425:21–33, 2012.

[9] C. Sousa-Silva, A. F. Al-Refaie, J. Tennyson, and S. N. Yurchenko. ExoMol line lists VIII: A Hot Line List for Phosphine. *MNRAS*, 446:2337–2347, 2015.

[10] S W Sharpe, T J Johnson, R L Sams, P M Chu, G C Rhoderick, and P A Johnson. Gas-phase databases for quantitative infrared spectroscopy. *Appl. Spectrosc.*, 58:1452–1461, 2004.

[11] T. Nakanaga, S. Kondo, and S. Saeki. Infrared band intensities of formaldehyde and formaldehyded2. *J. Chem. Phys.*, 76:3860–3865, 1982.

[12] J. Tennyson, K. Hulme, O. K. Naim, and S. N. Yurchenko. Radiative lifetimes and cooling functions for astrophysically important molecules. *J. Phys. B: At. Mol. Opt. Phys.*, 49:044002, 2016.

[13] D. C. Reuter, S. Nadler, S. J. Daunt, and J. W. C. Johns. Frequency and in-

tensity analysis of the nu-3-band, nu-4-band and nu-6-band of formaldehyde. *J. Chem. Phys.*, 91:646–654, 1989.

[14] W. Zhao, X. Gao, L. Deng, T. Huang, T. Wu, and W. Zhang. Absorption spectroscopy of formaldehyde at 1.573 μm. *J. Quant. Spectrosc. Radiat. Transf.*, 107:331–339, 2007.

[15] J. M. Flaud, W. J. Lafferty, R. L. Sams, and S. W. Sharpe. High resolution spectroscopy of $H_2^{12}C^{16}O$ in the 1.9 to 2.56 μm spectral range. *Mol. Phys.*, 104:1891–1903, 2006.

[16] A. F. Al-Refaie, R. I. Ovsyannikov, O. L. Polyansky, S. N. Yurchenko, and J. Tennyson. A variationally calculated room temperature line-list for H_2O_2. *J. Mol. Spectrosc.*, 318:84–90, 2015.

[17] M. R. Zaghloul and A. N. Ali. Algorithm 916: Computing the Faddeyeva and Voigt functions. *CoRR*, abs/1106.0151, 2011.

[18] A. I. Pavlyuchko, S. N. Yurchenko, and J. Tennyson. A hybrid variation-perturbation method for calculating rovibrational energy levels of a polyatomic molecules. *Mol. Phys.*, 113:1559–1575, 2015.

[19] A. I. Pavlyuchko, S. N. Yurchenko, and J. Tennyson. ExoMol line lists XI: A Hot Line List for nitric acid. *MNRAS*, 452:1702–1706, 2015.

[20] P. R. Bunker and P. Jensen. *Molecular Symmetry and Spectroscopy*. NRC Research Press, Ottawa, 2 edition, 1998.

[21] D. S. Underwood, J. Tennyson, S. N. Yurchenko, S. Clausen, and A. Fateev. ExoMol line lists XVII: A line list for hot SO_3. *MNRAS*, page (submitted), 2016.

[22] S. Carter, N. Pinnavaia, and N. C. Handy. The vibrations of formaldehyde. *Chem. Phys. Lett.*, 240:400–408, 1995.

[23] J. Fischer, R. R. Gamache, A. Goldman, L. S. Rothman, and A. Perrin. Total internal partition sums for molecular species in the 2000 edition of the HITRAN database. *J. Quant. Spectrosc. Radiat. Transf.*, 82:401–412, 2003.

[24] A. Perrin, J.-M. Flaud, C. CamyPeyret, R. Schermaul, M. Winnewisser, J.-Y. Mandin, V Dana, M Badaoui, and J Koput. Line intensities in the far-infrared spectrum of H_2O_2. *J. Mol. Spectrosc.*, 176:287–296, 1996.

[25] J.-M. Flaud, C. Camy-Peyret, J. W. C. Johns, and B. Carli. The far infrared spectrum of H_2O_2. first observation of the staggering of the levels and determination of the cis barrier. *J. Chem. Phys.*, 91:1504–1510, 1989.

[26] C. Camy-Peyret, J.-M. Flaud, J. W. C. Johns, and M. Noel. Torsion-vibration interaction in H_2O_2: First high-resolution observation of v_3. *J. Mol. Spectrosc.*, 155:84–104, 1992.

[27] W. B. Olson, R. H. Hunt, B. W. Young, A. G. Maki, and J. W. Brault. Rotational constants of the lowest torsional component (0g) of the ground state and lowest torsional component (1g) of the first excited torsional state of hydrogen peroxide. *J. Mol. Spectrosc.*, 127:12–34, 1988.

[28] A. Perrin, J-M. Flaud, C. Camy-Peyret, A. Goldman, F J Murcray, and R D Blatherwick. New analysis of the ν_6 band of H_2O_2 – the $(n, \tau) = (0, 1)$, (1, 1), 2, 1), (0, 3), and (1, 3) torsional subbands. *J. Mol. Spectrosc.*, 142:129–147, 1990.

[29] R. J. Barber, J. Tennyson, G. J. Harris, and R. N. Tolchenov. A high accuracy computed water line list. *MNRAS*, 368:1087–1094, 2006.

[30] S. N. Yurchenko, W. Thiel, and P. Jensen. Theoretical ROVibrational Energies (TROVE): A robust numerical approach to the calculation of rovibrational energies for polyatomic molecules. *J. Mol. Spectrosc.*, 245:126–140, 2007.

[31] S. Seager and D. D. Sasselov. Theoretical transmission spectra during extrasolar giant planet transits. *ApJ*, 537(2):916, 2000.

[32] W. B. Hubbard, J. J. Fortney, J. I. Lunine, A. Burrows, D. Sudarsky, and P. Pinto. Theory of extrasolar giant planet transits. *ApJ*, 560(1):413, 2001.

[33] D. Charbonneau, T. M. Brown, R. W. Noyes, and R. L. Gilliland. Detection of an extrasolar planet atmosphere. *ApJ*, 568(1):377, 2002.

[34] G. Tinetti, A. Vidal-Madjar, M.-C. Liang, J.-P. Beaulieu, Y. Yung, S. Carey, R. J. Barber, J. Tennyson, I. Ribas, N. Allard, G. E. Ballester, D. K. Sing, and F. Selsis. Water vapour in the atmosphere of a transiting extrasolar planet. *Nature*, 448:169–171, July 2007.

[35] M. R. Swain, G. Vasisht, G. Tinetti, J. Bouwman, Pin Chen, Y. Yung, D. Deming, and P. Deroo. Molecular signatures in the near-infrared dayside spectrum of HD 189733b. *ApJ*, 690:L114–L117, 2009.

[36] M. R. Swain, G. Vasisht, and G. Tinetti. The presence of methane in the atmosphere of an extrasolar planet. *Nature*, 452:329–331, 2008.

[37] R. M. Goody and Y. L. Yung. *Atmospheric radiation : theoretical basis*. 1989.

[38] N. Madhusudhan and S. Seager. A temperature and abundance retrieval method for exoplanet atmospheres. *ApJ*, 707(1):24, 2009.

[39] B. Benneke and S. Seager. Atmospheric retrieval for super-earths: Uniquely constraining the atmospheric composition with transmission spectroscopy. *ApJ*, 753(2):100, 2012.

[40] A. Tsiaras, M. Rocchetto, I. P. Waldmann, O. Venot, R. Varley, G. Morello, M. Damiano, G. Tinetti, E. J. Barton, S. N. Yurchenko, and J. Tennyson. Detection of an atmosphere around the super-earth 55 cancri e. *ApJ*, 820(2): 99, 2016.

[41] L S Rothman, D Jacquemart, A Barbe, D C Benner, M Birk, L R Brown, M R Carleer, C Chackerian, K Chance, L H Coudert, V Dana, V M Devi, J-M Flaud, R R Gamache, A Goldman, J-M Hartmann, K W Jucks, A G Maki, J-Y Mandin, S T Massie, J Orphal, A Perrin, C P Rinsland, M A H Smith, J Tennyson, R N Tolchenov, R A Toth, J Vander Auwera, P Varanasi, and G Wagner. The *HITRAN* 2004 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transf.*, 96:139–204, 2005.

[42] L. S. Rothman, I. E. Gordon, A. Barbe, D. Chris Benner, P. F. Bernath, M. Birk, V. Boudon, L. R. Brown, A. Campargue, J. P. Champion, K. Chance, L. H. Coudert, V. Dana, V. M. Devi, S. Fally, J. M. Flaud, R. R. Gamache, A. Goldman, D. Jacquemart, I. Kleiner, N. Lacome, W. J. Lafferty, J. Y. Mandin, S. T. Massie, S. N. Mikhailenko, C. E. Miller, N. Moazzen-Ahmadi, O. V. Naumenko, A. V. Nikitin, J. Orphal, V. I. Perevalov, A. Perrin, A. Predoi-Cross, C. P. Rinsland, M. Rotger, M. Simeckova, M. A. H. Smith, K. Sung, S. A. Tashkun, J. Tennyson, R. A. Toth, A. C. Vandaele, and J. Vander Auwera. The *HITRAN* 2008 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transf.*, 110:533–572, 2009.

[43] H. S. P. Müller, F Schlöder, J Stutzki, and G Winnewisser. The Cologne database for molecular spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. *J. Molec. Struct. (THEOCHEM)*, 742:215–227, 2005.

[44] H. M. Pickett, R. L. Poynter, E. A. Cohen, M. L. Delitsky, J. C. Pearson, and H. S. P. Müller. Submillimeter, millimeter, and microwave spectral line catalog. *J. Quant. Spectrosc. Radiat. Transf.*, 60:883–890, 1998.

[45] L. S. Rothman, I. E. Gordon, R. J. Barber, H. Dothe, R. R. Gamache, A. Goldman, V. I. Perevalov, S. A. Tashkun, and J. Tennyson. HITEMP, the High-Temperature Molecular Spectroscopic Database. *J. Quant. Spectrosc. Radiat. Transf.*, 111:2139–2150, 2010.

[46] M. Vidler and J. Tennyson. Accurate partition function and thermodynamic data for water. *J. Chem. Phys.*, 113:9766–9771, 2000.

[47] C. Sousa-Silva, N. Hesketh, S. N. Yurchenko, C. Hill, and J. Tennyson. High Temperature partition functions and thermodynamic data for ammonia and phosphine. *J. Quant. Spectrosc. Radiat. Transf.*, 142:66–74, 2014.

[48] G D Carney and R N Porter. H_3^+ - ab-initio calculation of vibration spectrum. *J. Chem. Phys.*, 65:3547–3565, 1976.

[49] G D Carney and R N Porter. Abinitio prediction of vibration-spectra of deuterated species of H_3^+. *Chem. Phys. Lett.*, 50:327–329, 1977.

[50] G D Carney and R N Porter. Abinitio prediction of the rotation-vibration spectrum of H_3^+ and D_3^+. *Phys. Rev. Lett.*, 45:537–541, 1980.

[51] S. Miller, J. Tennyson, S. Lepp, and A. Dalgarno. Identification of features due to H_3^+ in the infrared spectrum of supernova 1987a. *Nature*, 355:420–422, 1992.

[52] L. Neale, S. Miller, and J. Tennyson. Spectroscopic properties of the H_3^+ molecule: a new calculated linelist. *ApJ*, 464:516–520, 1996.

[53] P Bergeron, M T Ruiz, and S K Leggett. The chemical evolution of cool white dwarfs and the age of the local galactic disk. *ApJS*, 108:339–387, 1997.

[54] Holger Kreckel, Oldřich Novotný, K. N. Crabtree, H. Buhr, A. Petrignani, B. A. Tom, R. D. Thomas, M. H. Berg, D. Bing, M. Grieser, C. Krantz, M. Lestinsky, M. B. Mendes, C. Nordhorn, R. Repnow, J. Stützel, A. Wolf, and B. J. McCall. Storage ring measurements of the dissociative recombination of H_3^+ : a closer look. *Phys. Rev. A*, 82:042715, 2010.

[55] T. T. Koskinen, A. D. Aylward, C. G. A. Smith, and S. Miller. A Thermospheric Circulation Model for Extrasolar Giant Planets. *ApJ*, 661:515–526, May 2007.

[56] F. Allard, D. Homeier, and B. Freytag. Models of very-low-mass stars, brown dwarfs and exoplanets. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 370(1968): 2765–2777, 2012. ISSN 1364-503X.

[57] S. J. Warren, D. J. Mortlock, S. K. Leggett, D. J. Pinfield, D. Homeier, S. Dye, R. F. Jameson, N. Lodieu, P. W. Lucas, A. J. Adamson, F. Allard, D. Barrado y Navascués, M. Casali, K. Chiu, N. C. Hambly, P. C. Hewett, P. Hirst, M. J. Irwin, A. Lawrence, M. C. Liu, E. L. Martín, R. L. Smart, L. Valdivielso, and B. P. Venemans. A very cool brown dwarf in ukidss dr1. *MNRAS*, 381(4):1400–1412, 2007.

[58] E. F. van Dishoeck, E. Herbst, and D. A. Neufeld. Interstellar water chemistry: From laboratory to observations. *Chemical Reviews*, 113(12):9043–9085, 2013. PMID: 24261880.

[59] J. S. Carr and J. R. Najita. Organic molecules and water in the inner disks of t tauri stars. *ApJ*, 733(2):102, 2011.

[60] R. J. Barber, S. Miller, N. Dello Russo, M. J. Mumma, J. Tennyson, and P. Guio. Water in the near-infrared spectrum of comet 8P/Tuttle. *MNRAS*, 398(3):1593–1600, 2009.

[61] G. L. Villanueva, M. J. Mumma, B. P. Bonev, M. A. DiSanti, E. L. Gibb, H. Böhnhardt, and M. Lippi. A sensitive search for deuterated water in comet 8p/tuttle. *ApJ*, 690(1):L5, 2009.

[62] L.A. Kranendonk, X. An, A.W. Caswell, R.E. Herold, S.T. Sanders, R. Huber, J.G. Fujimoto, Y. Okura, and Y. Urata. High speed engine gas thermometry by fourier-domain mode-locked laser absorption spectroscopy. *Optics Express*, 15(23):15115–15128, 2007. cited By 76.

[63] R. J Mathar. Refractive index of humid air in the infrared: model fits. *Journal of Optics A: Pure and Applied Optics*, 9(5):470, 2007.

[64] S. N. Yurchenko and J. Tennyson. ExoMol line lists IV: The rotation-vibration spectrum of methane up to 1500 K. *MNRAS*, 440:1649–1661, 2014.

[65] J. I. Canty, P. W. Lucas, Sergei N. Yurchenko, Jonathan Tennyson, S. K. Leggett, C. G. Tinney, H. R. A. Jones, Ben Burningham, D. J. Pinfield, and R. L. Smart. Methane and ammonia in the near-infrared spectra of late-t dwarfs. *MNRAS*, 450(1):454–480, 2015.

[66] Bonnefoy, M., Zurlo, A., Baudino, J. L., Lucas, P., Mesa, D., Maire, A.-L., Vigan, A., Galicher, R., Homeier, D., Marocco, F., Gratton, R., Chauvin, G., Allard, F., Desidera, S., Kasper, M., Moutou, C., Lagrange, A.-M., Antichi, J., Baruffolo, A., Baudrand, J., Beuzit, J.-L., Boccaletti, A., Cantalloube, F., Carbillet, M., Charton, J., Claudi, R. U., Costille, A., Dohlen, K., Dominik, C., Fantinel, D., Feautrier, P., Feldt, M., Fusco, T., Gigan, P., Girard, J. H., Gluck, L., Gry, C., Henning, T., Janson, M., Langlois, M., Madec, F., Magnard, Y., Maurel, D., Mawet, D., Meyer, M. R., Milli, J., Moeller-Nilsson, O., Mouillet, D., Pavlov, A., Perret, D., Pujet, P., Quanz, S. P., Rochat, S., Rousset, G., Roux, A., Salasnich, B., Salter, G., Sauvage, J.-F., Schmid, H. M., Sevin, A., Soenke, C., Stadler, E., Turatto, M., Udry, S., Vakili, F., Wahhaj, Z., and Wildi, F. First light of the vlt planet finder sphere. *A&A*, 587: A58, 2016.

[67] H. J. Werner, P. J. Knowles, R. Lindh, F. R. Manby, and M. Schütz. MOLPRO, a package of ab initio programs, 2010. see `http://www.molpro.net/`.

[68] M. E. Harding, T. Metzroth, J. Gauss, and A. A. Auer. Parallel calculation of ccsd and ccsd(t) analytic first and second derivatives. *J Chem. Theory Comput.*, 4:64–74, 2008.

[69] R. J. Le Roy. *LEVEL 8.0 A Computer Program for Solving the Radial Schrödinger Equation for Bound and Quasibound Levels*. University of Waterloo Chemical Physics Research Report CP-663, `http://leroy.uwaterloo.ca/programs/`, 2007.

[70] S. N. Yurchenko, L. Lodi, J. Tennyson, and A. V. Stolyarov. Duo: a general program for calculating spectra of diatomic molecules. *Comput. Phys. Commun.*, 202:262–275, 2016.

[71] J. Tennyson, J. R. Henderson, and N. G. Fulton. DVR3D: programs for fully pointwise calculation of ro-vibrational spectra of triatomic molecules. *Comput. Phys. Commun.*, 86:175–198, 1995.

[72] A. F. Al-Refaie, J. Tennyson, and S. N. Yurchenko. GPU Acelerated INtensities (GAIN): a new method of computing Einstein A coefficients. *Comput. Phys. Commun.*, 2016. (to be submitted).

[73] L. Lodi and J. Tennyson. Theoretical methods for small-molecule rovibrational spectroscopy. *J. Phys. B: At. Mol. Opt. Phys.*, 43:133001, 2010.

[74] D. R. Hartree. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. *Proceedings of the Cambridge Philosophical Society*, 24:111, 1928.

[75] J M Bowman. The self-consistent-field approach to polyatomic vibrations. *Acc. Chem. Res.*, 19:202–208, 1986.

[76] R B Gerber and M A Ratner. Self-consistent-field methods for vibrational excitations in polyatomic systems. *Adv. Chem. Phys.*, 70:97–132, 1988.

[77] P. Jensen and P. R. Bunker. *Computational Molecular Spectroscopy*. Wiley, 2000.

[78] L. B. Harding, S. J. Klippenstein, H. Lischka, and R. Shepard. Comparison of multireference configuration interaction potential energy surfaces for H + $O_2 \rightarrow HO_2$: the effect of internal contraction. In Angela K. Wilson, Kirk A. Peterson, and David E. Woon, editors, *Thom H. Dunning, Jr.*, volume 10 of *Highlights in Theoretical Chemistry*, pages 115–121. Springer Berlin Heidelberg, 2015.

[79] J. M. L. Martin and M. K. Kesharwani. Assessment of CCSD(t)-F12 approximations and basis sets for harmonic vibrational frequencies. *J. Chem. Theor. Comput.*, 10:2085–2090, 2014.

[80] H. Kümmel. Origins of the coupled cluster method. *Theoretica chimica acta*, 80(2-3):81–89, 1991.

[81] R. J. Bartlett and M. Musiał. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.*, 79:291–352, 2007.

[82] G. O. Sørensen. A new approach to the hamiltonian of nonrigid molecules. In M. J. S. Dewar et al., editor, *Large Amplitude Motion in Molecules II*, volume 82 of *Topics in Current Chemistry*, pages 97–175. Springer Berlin Heidelberg, Heidelberg, 1979.

[83] A. Yachmenev and S. N. Yurchenko. Automatic differentiation method for numerical construction of the rotational-vibrational hamiltonian as a power series in the curvilinear internal coordinates using the eckart frame. *J. Chem. Phys.*, 143:014105, 2015.

[84] O. L. Polyansky, I. N. Kozin, P. Maĺyszek, J. Koput, J. Tennyson, and S. N. Yurchenko. Variational calculation of highly excited rovibrational energy levels of H_2O_2. *J. Phys. Chem. A*, 117:7367–7377, 2013.

[85] R. Bartholomae, D. Martin, and B.T. Sutcliffe. A nonempirical calculation of the lowest vibrational band origins for the CH_2^+ radical using a novel vibrational wavefunction. *J. Mol. Spectrosc.*, 87(2):367 – 381, 1981. ISSN 0022-2852.

[86] I. N. Kozin, M. M. Law, J. Tennyson, and J. M. Hutson. New vibration-rotation code for tetraatomic molecules WAVR4. *Comput. Phys. Commun.*, 163:117–131, 2004.

[87] A. Yachmenev K. Chubb and S. Yurchenko. Private communications.

[88] P. R. Bunker and P. Jensen. *Fundamentals of Molecular Symmetry*. IOP Publishing, Bristol, 2004.

[89] S N Yurchenko, M Carvajal, P Jensen, H Lin, J J Zheng, and W Thiel. Rotation-vibration motion of pyramidal XY3 molecules described in the Eckart frame: Theory and application to NH_3. *Mol. Phys.*, 103:359–378, 2005.

[90] B. V. Noumerov. A method of extrapolation of perturbations. *MNRAS*, 84: 592–602, 1924.

[91] J. W. Cooley. An improved eigenvalue corrector formula for solving the Schrödinger equation for central fields. *Math. Comp.*, 15:363–374, 1961.

[92] S. N. Yurchenko, R. J. Barber, A. Yachmenev, W. Thiel, P. Jensen, and J. Tennyson. A variationally computed T=300 K line list for NH_3. *J. Phys. Chem. A*, 113:11845–11855, 2009.

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

[94] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide: Solution of Large-scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods (Software, Environments and Tools)*. Society for Industrial & Applied Mathematics, U.S., 1998. see http://www.caam.rice.edu/software/ARPACK/.

[95] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. *ScaLAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.

[96] G. Hose and H. S. Taylor. Quantum kolmogorov-arnol'd-moser-like theorem:

Fundamentals of localization in quantum theory. *Phys. Rev. Lett.*, 51:947–950, 1983.

[97] A. Yachmenev, S. N. Yurchenko, P. Jensen, and W. Thiel. A new "spectroscopic" potential energy surface for formaldehyde in its ground electronic state. *J. Chem. Phys.*, 134:11, 2011.

[98] S. N. Yurchenko, R. J. Barber, J. Tennyson, W. Thiel, and P. Jensen. Towards efficient refinement of molecular potential energy surfaces: Ammonia as a case study. *J. Mol. Spectrosc.*, 268:123–129, 2011.

[99] R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56:340–343, Aug 1939.

[100] S N Yurchenko, M. Carvajal, P. Jensen, F. Herregodts, and T R Huet. Potential parameters of PH_3 obtained by simultaneous fitting of ab initio data and experimental vibrational band origins. *Chem. Phys.*, 290:59–67, 2003.

[101] S. N. Yurchenko, R. J. Barber, and J. Tennyson. A variationally computed hot line list for NH_3. *MNRAS*, 413:1828–1834, 2011.

[102] D. S. Underwood, S. N. Yurchenko, J. Tennyson, and P. Jensen. Rotational spectrum of SO_3 and a theoretical evidence for the formation of rotational, energy level clusters in its vibrational ground state. *J. Chem. Phys.*, 140: 244316, 2014.

[103] A. J. Stone. input.f90 - a fortran90 module for parsing text input, 2005. See http://www-stone.ch.cam.ac.uk/programs/.

[104] L. S. Blackford, J. Demmel, J. Dongarra, I. Duff, S. Hammarling, G. Henry, M. Heroux, L. Kaufman, A. Lumsdaine, A. Petitet, R. Pozo, K. Remington, and R. C. Whaley. An updated set of basic linear algebra subprograms (blas). *ACM Transactions on Mathematical Software*, 28:135–151, 2001.

[105] A. F. Al-Refaie, S. N. Yurchenko, A. Yachmenev, and J. Tennyson. ExoMol line lists IX: A variationally computed line-list for hot formaldehyde. *MNRAS*, 448:1704–1714, 2015.

[106] nVidia. *CUBLAS Library User Guide*. nVidia, v5.0 edition, October 2012.

[107] E. Maciá. The role of phosphorus in chemical evolution. *Chem. Soc. Rev.*, 34:691–701, 2005.

[108] C. Sousa-Silva, S. N. Yurchenko, and J. Tennyson. A computed room temperature line list for phosphine. *J. Mol. Spectrosc.*, 288:28–37, 2013.

[109] R. Peer Wayne. *Chemistry of Atmospheres*. Oxford University Press, 2000.

[110] O. I. Korablev, M Ackerman, V. A Krasnopolsky, V. I. Moroz, C. Muller, A. V. Rodin, and S. K. Aterya. Tentative identification of formaldehyde in the martian atmosphere. *Planet Space Sci.*, 41:441–451, 1993.

[111] G.L. Villanueva, M.J. Mumma, R.E. Novak, Y.L. Radeva, H.U. Käufl, A. Smette, A. Tokunaga, A. Khayat, T. Encrenaz, and P. Hartogh. A sensitive search for organics (CH_4, CH_3OH, H_2CO, C_2H_6, C_2H_2, C_2H_4), hydroperoxyl (HO_2), nitrogen compounds (N_2O, NH_3, HCN) and chlorine species (HCL, CH_3CL) on mars using ground-based high-resolution infrared spectroscopy. *Icarus*, 223:11 – 27, 2013.

[112] B. Zuckerman, D. Buhl, P. Palmer, and L. E. Snyder. Observations of interstellar formaldehyde. *ApJ*, 160:485, 1970.

[113] W. D. Langer. Interstellar cloud evolution and the abundance of formaldehyde. *ApJ*, 210:328–333, 1976.

[114] B. Zuckerman, D. Buhl, P. Palmer, and L. E. Snyder. 12c/13c abundance ratios from observations of interstellar H_2CO. *ApJ*, 189:217–220, 1974.

[115] D. E. Woon. Modeling gas-grain chemistry with quantum chemical cluster calculations. i. heterogeneous hydrogenation of CO and H_2CO on icy grain mantles. *ApJ*, 569:541, 2002.

[116] W. A. Schutte. Production of organic molecules in interstellar ices. *Adv. Space Res.*, 30:1409 – 1417, 2002.

[117] J. M. Hollis, F. J. Lovas, and P. R. Jewell. Interstellar glycolaldehyde: The first sugar. *ApJ*, 540:L107–L110, 2000.

[118] D. Bockelee-Morvan and J. Crovisier. Formaldehyde in comets. ii - excitation of the rotational lines. *A&A*, 264:282–291, 1992.

[119] N. Dello Russo, R. J. Vervack, Jr., C. M. Lisse, H. A. Weaver, H. Kawakita, H. Kobayashi, A. L. Cochran, W. M. Harris, A. J. McKay, N. Biver, D. Bockelée-Morvan, and J. Crovisier. The volatile composition and activity of comet 103p/hartley 2 during the epoxi closest approach. *ApJ*, 734: L8, 2011.

[120] G. L. Villanueva, M. J. Mumma, M. A. Disanti, B. P. Bonev, E. L. Gibb, K. Magee-Sauer, G. A. Blake, and C. Salyk. The molecular composition of comet c/2007 w1 (boattini): Evidence of a peculiar outgassing and a rich chemistry. *Icarus*, 216:227–240, 2011.

[121] S. N. Milam, A. J. Remijan, M. Womack, L. Abrell, L. M. Ziurys, S. Wyckoff, A. J. Apponi, D. N. Friedel, L. E. Snyder, J. M. Veal, Patrick Palmer, L. M. Woodney, M. F. A'Hearn, J. R. Forster, M. C. H. Wright, I. de Pater, S. Choi, and M. Gesmundo. Formaldehyde in comets C/1995 O1 (Hale-Bopp), C/2002 T7 (linear), and C/2001 Q4 (neat): Investigating the cometary origin of H_2CO. *ApJ*, 649:1169, 2006.

[122] H. Cottin, M. C. Gazeau, Y. Benilan, and F. Raulin. Polyoxymethylene as parent molecule for the formaldehyde extended source in comet halley. *ApJ*, 556:417, 2001.

[123] K. I. Öberg, C. Qi, J. K. J. Fogel, E. A. Bergin, S. M. Andrews, C. Espaillat, T. A. van Kempen, D. J. Wilner, and I. Pascucci. The disk imaging survey of chemistry with sma. i. taurus protoplanetary disk data. *ApJ*, 720:480–493, 2010.

[124] G. Zasowski, F. Kemper, D. M. Watson, E. Furlan, C. J. Bohac, C. Hull, and J. D. Green. Spitzer infrared spectrograph observations of class i/ii objects in

taurus: Composition and thermal history of the circumstellar ices. *ApJ*, 694: 459–478, 2009.

[125] B. A. Sargent, W. Forrest, Dan M. Watson, P. D'Alessio, N. Calvet, E. Furlan, K. H. Kim, J. Green, K. Pontoppidan, I. Richter, and C. Tayrien. Emission from water vapor and absorption from other gases at 5-7.5 um in spitzer-irs spectra of protoplanetary disks. *The Astrophysical Journal*, 792(2):83, 2014.

[126] M. Neveu, H. -J. Kim, and S. A. Benner. The "strong" RNA world hypothesis: Fifty years old. *Astrobiology*, 13:391–403, 2013. ISSN 1531-1074.

[127] N. Goldman and I. Tamblyn. Prebiotic chemistry within a simple impacting icy mixture. *J. Phys. Chem. A*, 117:5124–5131, 2013. ISSN 1089-5639.

[128] J R Forster, W M Goss, T L Wilson, D Downes, and H R Dickel. A formaldehyde maser in NGC7538. *A&A*, 84:L1–L3, 1980.

[129] P Pratap, L E Snyder, and W Batrla. Are the NGC-7538 formaldehyde masers really unusual. *ApJ*, 387:241–247, 1992.

[130] J. G. Mangum, J. Darling, K. M. Menten, and C. Henkel. Formaldehyde densitometry of starburst galaxies. *ApJ*, 673:832–846, 2008.

[131] E. Araya, P. Hofner, M. Sewiło, H. Linz, S. Kurtz, L. Olmi, C. Watson, and E. Churchwell. First detection of an H_2CO 6 cm maser flare: A burst in iras 18566+0408. *ApJL*, 654:L95, 2007.

[132] J. W. C. Johns and A. R. W. McKellar. Stark spectroscopy with the CO laser: The dipole moment of H_2CO in the v2=2 state. *J. Chem. Phys.*, 63: 1682–1685, 1975.

[133] L R Brown, R H Hunt, and A S Pine. Wavenumbers, line strengths, and assignments in the doppler-limited spectrum of formaldehyde from 2700 cm^{-1} to 3000 cm^{-1}. *J. Mol. Spectrosc.*, 75(3):406–428, 1979.

[134] S. Nadler, S. J. Daunt, and D. C. Reuter. Tuable diode-laser measurements OF formaldehyde foreign-gas broadening parameters and line strengths in the 9-11 μm region. *Appl. Optics*, 26:1641–1646, 1987.

[135] D S Cline and P L Varghese. High-resolution spectral measurements in the ν_5 band of formaldehyde using a tunable IR diode-laser. *Appl. Optics*, 27 (15):3219–3224, AUG 1 1988.

[136] N. M. Poulin, M. J. Bramley, T. Carrington, H. G. Kjaergaard, and B. R. Henry. Calculation of vibrational (j=0) excitation energies and band intensities of formaldehyde using the recursive residue generation method. *J. Chem. Phys.*, 104:7807–7820, 1996.

[137] P. Theulé, A. Callegari, T. R. Rizzo, and J. S. Muenter. Fluorescence detected microwave stark effect measurements in excited vibrational states of H_2CO. *J. Chem. Phys.*, 119:8910–8915, 2003.

[138] D J Clouthier and D A Ramsay. The Spectroscopy of Formaldehyde and Thioformaldehyde. *Annu. Rev. Phys. Chem.*, 34:31–58, 1983.

[139] A. Perrin, D. Jacquemart, F. Kwabia Tchana, and N. Lacome. Absolute line intensities measurements and calculations for the 5.7 and 3.6 mu m bands of formaldehyde. *J. Quant. Spectrosc. Radiat. Transf.*, 110:700–716, 2009.

[140] A. Perrin, F. Keller, and J. M. Flaud. New analysis of the nu(2), nu(3), nu(4), and nu(6) bands of formaldehyde, (H_2CO)-c-12-o-16 line positions and intensities in the 5-10 mu m spectral region. *J. Mol. Spectrosc.*, 221: 192–198, 2003.

[141] F. K. Tchana, A. Perrin, and N. Lacome. New analysis of the nu(2) band of formaldehyde ($H_2^{12}C^{16}O$): Line positions for the nu(2), nu(3) nu(4) and nu(6) interacting bands. *J. Mol. Spectrosc.*, 245:141–144, 2007.

[142] A. Perrin, A. Valentin, and L. Daumont. New analysis of the 2(v4), v(4)+v(6), 2v(6), v(3)+v(4), v(3)+v(6), v(1), v(5), v(2)+v(4), 2v(3),

v(2)+v(6) and v(2)+v(3), bands of formaldehyde $H_2^{12}C^{16}O$: Line positions and intensities in the 3.5 μm spectral region. *J. Molec. Struct. (THEOCHEM)*, 780-81:28–44, 2006.

[143] F. Ito, T. Nakanaga, and H. Takeo. FTIR spectra of the 2-nu(4)(4), nu(4)+nu(6) and 2-nu(6) bands of formaldehyde. *Spectra Chimica Acta A*, 50:1397–1412, 1994.

[144] R. Perez, J. M. Brown, Y. Utkin, J. X. Han, and R. F. Curl. Observation of hot bands in the infrared spectrum of H_2CO. *J. Mol. Spectrosc.*, 236:151–157, 2006.

[145] L. Margules, A. Perrin, R. Janeckova, S. Bailleux, C. P. Endres, T. F. Giesen, and S. Schlemmer. Rotational transitions within the 2(1), 3(1), 4(1), and 6(1) states of formaldehyde $H_2^{12}C^{16}O$. *Can. J. Phys.*, 87:425–435, 2009.

[146] D Luckhaus, M J Coffey, M D Fritz, and F F Crim. Experimental and theoretical vibrational overtone spectra of v-ch=3,4,5, and 6 in formaldehyde (H_2CO). *J. Chem. Phys.*, 104:3472–3478, 1996. ISSN 0021-9606.

[147] S. Carter, A. R. Sharma, J. M. Bowman, P. Rosmus, and R. Tarroni. Calculations of rovibrational energies and dipole transition intensities for polyatomic molecules using multimode. *J. Chem. Phys.*, 131, 2009.

[148] S. N. Yurchenko. Electric dipole moments of small polyatomic molecules from first principles. In *Chemical Modelling: Volume 10*, volume 10, chapter 7, pages 183–228. The Royal Society of Chemistry, 2014.

[149] T. Nakagawa and Y. Morino. Coriolis interactions in the v4 and v6 bands of formaldehyde. *J. Mol. Spectrosc.*, 38:84 – 106, 1971.

[150] H Partridge and D W Schwenke. The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data. *J. Chem. Phys.*, 106:4618–4639, 1997.

[151] R. J. Bouwens, J. A. Hammerschmidt, M. M. Grzeskowiak, T. A. Stegink, P. M. Yorba, and W. F. Polik. Pure vibrational spectroscopy of s-0 formaldehyde by dispersed fluorescence. *J. Chem. Phys.*, 104:460–479, 1996.

[152] C Eckart. Some studies concerning rotating axes and polyatomic molecules. *Phys. Rev.*, 47:552–558, 1935.

[153] A. Yachmenev, I. Polyak, and W. Thiel. Theoretical rotation-vibration spectrum of thioformaldehyde. *J. Chem. Phys.*, 139(20), NOV 28 2013.

[154] B Fabricant, D Krieger, and J S Muenter. Molecular-beam electric resonance study of formaldehyde, thioformaldehyde, and ketene. *J. Chem. Phys.*, 67 (4):1576–1586, 1977.

[155] J. Tennyson, C. Hill, and S. N. Yurchenko. Data structures for ExoMol: Molecular line lists for exoplanet and other atmospheres. In *6^{th} international conference on atomic and molecular data and their applications ICAMDATA-2012*, volume 1545 of *AIP Conference Proceedings*, pages 186–195. AIP, New York, 2013.

[156] L. Neale and J. Tennyson. A high temperature partition function for H_3^+. *ApJ*, 454:L169–L173, 1995.

[157] H. Barry, L. Corner, G. Hancock, R. Peverall, and G. A. D. Ritchie. Cross sections in the 2 nu(5) band of formaldehyde studied by cavity enhanced absorption spectroscopy near 1.76 mu. *Phys. Chem. Chem. Phys.*, 4:445–450, 2002.

[158] I C Hisatsune and D F Eggers. Infrared intensities and bond moments in formaldehyde. *J. Chem. Phys.*, 23(3):487–491, 1955.

[159] SC Herndon, DD Nelson, YQ Li, and MS Zahniser. Determination of line strengths for selected transitions in the nu(2) band relative to the ν_1 and ν_5 bands of h_2co. *J. Quant. Spectrosc. Radiat. Transf.*, 90(2):207–216, JAN 15 2005.

[160] M. Gray, S. Parfenov and A. Sobolev. Private communications.

[161] D D Davis. Kinetics review of atmospheric reactions involving h_xo_y compounds. *Can. J. Chem.*, 52:1405–1414, 1974. ISSN 0008-4042.

[162] K V Chance, D G Johnson, W A Traub, and K W Jucks. Measurement of the stratospheric hydrogen-peroxide concentration profile using far infrared thermal emission-spectroscopy. *Geophys. Res. Lett.*, 18:1003–1006, 1991.

[163] N D C Allen, G G Abad, P F Bernath, and C D Boone. Satellite observations of the global distribution of hydrogen peroxide (H_2O_2) from ACE. *J. Quant. Spectrosc. Radiat. Transf.*, 115:66–77, 2013. ISSN 0008-4042.

[164] E. -L. Zins and L. Krim. Hydrogenation processes from hydrogen peroxide: an investigation in ne matrix for astrochemical purposes. *RSC Adv.*, 4:22172–22180, 2014.

[165] R T Clancy, B J Sandor, and G H Moriarty-Schieven. A measurement of the 362 ghz absorption line of mars atmospheric H_2O_2. *Icarus*, 168:116–121, 2004. ISSN 0019-1035.

[166] T Encrenaz, B Bezard, T K Greathouse, M J Richter, J H Lacy, S K Atreya, A S Wong, S Lebonnois, F Lefevre, and F Forget. Hydrogen peroxide on mars: Evidence for spatial and seasonal variations. *Icarus*, 170:424–429, 2004.

[167] T. Encrenaz, T. K. Greathouse, F. Lefevre, and S. K. Atreya. Hydrogen peroxide on mars: Observations, interpretation and future plans. *Planet Space Sci.*, 68:3–17, 2012.

[168] S. Aoki, M. Giuranna, Y. Kasaba, H. Nakagawa, G. Sindoni, A. Geminale, and V. Formisano. Search for hydrogen peroxide in the Martian atmosphere by the Planetary Fourier Spectrometer onboard Mars Express. *Icarus*, 245: 177–183, 2015.

[169] K. P. Hand and M. E. Brown. Keck ii observations of hemispherical differences in h2o2 on europa. *ApJL*, 766:L21, 2013.

[170] P. Bergman, B. Parise, R. Liseau, B. Larsson, H. Olofsson, K. M. Menten, and R. Güsten. Detection of interstellar hydrogen peroxide. *A&A*, 531:L8, 2011.

[171] F. Du, B. Parise, and P. Bergman. Production of interstellar hydrogen peroxide (H_2O_2) on the surface of dust grains. *A&A*, 538:A91, 2012.

[172] R. D. Bach, P. Y. Ayala, and H. B. Schlegel. A reassessment of the bond dissociation energies of peroxides. an ab initio study. *J. Am. Chem. Soc.*, 118:12758–12765, 1996.

[173] J T Hougen. Summary of group theoretical results for microwave and infrare studies of h_2o_2. *Can. J. Phys.*, 62:1392–1402, 1984.

[174] A Perrin, A Valentin, J M Flaud, C Camy-Peyret, L Schriver, A Schriver, and P Arcas. The 7.9-μm band of hydrogen peroxide – line positions and intensities. *J. Mol. Spectrosc.*, 171:358–373, 1995. ISSN 0022-2852.

[175] P. A. Giguere and T. K. K. Srinivasan. A raman study of h2o2 and d2o2 vapor. *J. Raman Spectrosc.*, 2:125–132, 1974.

[176] G. Rauhut, G. Knizia, and H.-J. Werner. Accurate calculation of vibrational frequencies using explicitly correlated coupled-cluster theory. *J. Chem. Phys.*, 130:054105, 2009.

[177] D Luckhaus. 6d vibrational quantum dynamics: Generalized coordinate discrete variable representation and (a)diabatic contraction. *J. Chem. Phys.*, 113: 1329–1347, 2000. ISSN 0021-9606.

[178] M Mladenovic. Discrete variable approaches to tetratomic molecules part ii: application to H_2O_2 and H_2CO. *Spectra Chimica Acta A*, 58:809–824, 2002. ISSN 1386-1425.

[179] H G Yu and J T Muckerman. A general variational algorithm to calculate vibrational energy levels of tetraatomic molecules. *J. Mol. Spectrosc.*, 214: 11–20, 2002. ISSN 0022-2852.

[180] S. Carter, N. C. Handy, and J. M. Bowman. High torsional vibrational energies of H_2O_2 and CH_3OH studied by multimode with a large amplitude motion coupled to two effective contraction schemes. *Mol. Phys.*, 107:727–737, 2009.

[181] D. E. Bacelo and R. C. Binning. Theoretical modeling of the peroxide stretch in H_2O_2, F_2O_2, and Fe_2O_4. *Intern. J. Quantum Chem.*, 105:740–749, 2005.

[182] R Q Chen, G B Ma, and H Guo. Six-dimensional quantum calculations of highly excited vibrational energy levels of hydrogen peroxide and its deuterated isotopomers. *J. Chem. Phys.*, 114:4763–4774, 2001. ISSN 0021-9606.

[183] S Y Lin and H Guo. Exact quantum mechanical calculations of rovibrational energy levels of hydrogen peroxide (HOOH). *J. Chem. Phys.*, 119:5867–5873, 2003. ISSN 0021-9606.

[184] J. Koput, S. Carter, and N. C. Handy. Potential energy surface and vibrational-rotational energy levels of hydrogen peroxide. *J. Phys. Chem. A*, 102:6325–6335, 1998.

[185] B Kuhn, T R Rizzo, D Luckhaus, M Quack, and M A Suhm. A new six-dimensional analytical potential up to chemically significant energies for the electronic ground state of hydrogen peroxide. *J. Chem. Phys.*, 111:2565–2587, 1999.

[186] S. Carter, A. R. Sharma, and J. M. Bowman. Multimode calculations of rovibrational energies and dipole transition intensities for polyatomic molecules with torsional motion: Application to H_2O_2. *J. Chem. Phys.*, 135:014308, 2011. ISSN 0021-9606.

[187] J M Bowman, S Carter, and X C Huang. MULTIMODE: a code to calculate rovibrational energies of polyatomic molecules. *Intern. J. Quantum Chem.*, 22:533–549, 2003.

[188] P. Małyszek and J. Koput. Accurate ab initio potential energy surface and vibration-rotation energy levels of hydrogen peroxide. *J. Comput. Chem.*, 34:337–345, 2013.

[189] L. A. Zumwalt and P. A. Giguere. The infra-red bands of hydrogen peroxide at $\lambda 9720$ and the structure and torsional oscillation of hydrogen peroxide. *J. Chem. Phys.*, 9:458–462, 1941.

[190] T. J. Johnson, R. L. Sams, S. D. Burton, and T. A. Blake. Absolute integrated intensities of vapor-phase hydrogen peroxide (h2o2) in the mid-infrared at atmospheric pressure. *Anal. Bioanal. Chem.*, 395:377–386, 2009.

[191] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, and M. Schütz. Molpro: a general-purpose quantum chemistry program package. *WIREs Comput. Mol. Sci.*, 2:242–253, 2012.

[192] E.A. Cohen and H.M. Pickett. The dipole moment of hydrogen peroxide. *J. Mol. Spectrosc.*, 87:582 – 583, 1981.

[193] S. Klee, M. Winnewisser, A. Perrin, and J.-M. Flaud. Absolute line intensities for the v6 band of h2o2. *J. Mol. Spectrosc.*, 195:154 – 161, 1999.

[194] P. A. Giguere. The infra-red spectrum of hydrogen peroxide. *J. Chem. Phys.*, 18:88–92, 1950.

[195] C. Hill, S. N. Yurchenko, and J. Tennyson. Temperature-dependent molecular absorption cross sections for exoplanets and other atmospheres. *Icarus*, 226: 1673–1677, 2013.

[196] A. F. Al-Refaie, O. L. Polyansky, R. I., Ovsyannikov, J. Tennyson, and S. N. Yurchenko. ExoMol line lists XV: A hot line-list for hydrogen peroxide. *MNRAS*, 2016.

[197] E. J. Barton, S. N. Yurchenko, J. Tennyson, S. Clausen, and A. Fateev. High-resolution absorption measurements of NH_3 at high temperatures: 500 - 2100 cm^{-1}. *J. Quant. Spectrosc. Radiat. Transf.*, 167:126–134, 2015.

[198] E. J. Barton, S. N. Yurchenko, J. Tennyson, S. Béguier, and A. Campargue. A near infrared line list for NH_3: Analysis of a Kitt Peak spectrum after 35 years. *J. Mol. Spectrosc.*, 325:7–12, 2016.

[199] M. Kuntz. A new implementation of the humlicek algorithm for the calculation of the voigt profile function. *J. Quant. Spectrosc. Radiat. Transf.*, 57(6): 819 – 824, 1997. ISSN 0022-4073.

[200] J. He and Q. Zhang. An exact calculation of the voigt spectral line profile in spectroscopy. *Journal of Optics A: Pure and Applied Optics*, 9(7):565, 2007.

[201] S. N. Yurchenko. ExoCross. *Private Communication*, https://github.com/Trovemaster/exocross, 2016.

[202] S. L. Grimm and K. Heng. HELIOS-K: An ultrafast, open-source opacity calculator for radiative transfer. *ApJ*, 808(2):182, 2015.

[203] J. Tennyson, S. N. Yurchenko, A. F. Al-Refaie, E. J. Barton, K. L. Chubb, P. A. Coles, S. Diamantopoulou, M. N. Gorman, C. Hill, A. Z. Lam, L. Lodi, L. K. McKemmish, Y. Na, A. Owens, O. L. Polyansky, C. Sousa-Silva, D. S. Underwood, A. Yachmenev, and E. Zak. The ExoMol database: molecular line lists for exoplanet and other hot atmospheres. *J. Mol. Spectrosc.*, 2016.

[204] D. S. Underwood, J. Tennyson, and S. N. Yurchenko. An ab initio variationally computed room-temperature line list for SO_3. *Phys. Chem. Chem. Phys.*, 15:10118–10125, 2013.

[205] K. Kohse-Höinghaus, P. Oßwald, T. A. Cool, T. Kasper, N. Hansen, F. Qi, C. K. Westbrook, and P. R. Westmoreland. Biofuel combustion chemistry: From ethanol to biodiesel. *Angewandte Chemie International Edition*, 49 (21):3572–3597, 2010. ISSN 1521-3773.

[206] R. J. H. Klein-Douwel, J. Luque, J. B. Jeffries, G. P. Smith, and D. R. Crosley. Laser-induced fluorescence of formaldehyde hot bands in flames. *Appl. Opt.*, 39(21):3712–3715, Jul 2000.

[207] J. E. Harrington and K. C. Smyth. Laser-induced fluorescence measurements of formaldehyde in a methane/air diffusion flame. *Chem. Phys. Lett.*, 202(3): 196 – 202, 1993. ISSN 0009-2614.

[208] L. A. N. Patrick, K. Julia and B. Andreas. Detection of formaldehyde in flames using uv and mir absorption spectroscopy. *Zeitschrift für Physikalische Chemie*, 229:483–494, 2015.

[209] A. Ray, F. Perez, A. M. Beltramini, M.Jakubowycz, P. Dimick, M. R. Jacobs, K. Roman, R. A. Bonomo, and R. A. Salata. Use of vaporized hydrogen peroxide decontamination during an outbreak of multidrug-resistant acinetobacter baumannii infection at a long-term acute care hospital. *Infection Control & Hospital Epidemiology*, 31:1236–1241, 2010.

[210] A. van der Vliet and Y. M.W. Janssen-Heininger. Hydrogen peroxide as a damage signal in tissue injury and inflammation: Murderer, mediator, or messenger? *Journal of Cellular Biochemistry*, 115(3):427–435, 2014. ISSN 1097-4644.

[211] Dorscheid D. Wadsworth S, Sin D. Clinical update on the use of biomarkers of airway inflammation in the management of asthma. *Journal of Asthma and Allergy*, 4:77–86, 2011.

[212] L. Dong, R. J. Griffin, N. P. Sanchez, Y. Yu and F. K. Tittel. Mid-ir laser-based sensor for hydrogen peroxide detection. *Society of Photographic Instrumentation Engineers*, 2016.

[213] A. I. Pavlyuchko, S. N. Yurchenko, and J. Tennyson. A hybrid variation-perturbation calculation of the ro-vibrational spectrum of nitric acid. *J. Chem. Phys.*, 142:094309, 2015.