TOWARD THE THEORETICAL PREDICTION OF MOLECULAR VIBRATIONAL FREQUENCIES WITH SPECTROSCOPIC ACCURACY

by

Avery Elizabeth Wiens

(under the direction of Henry Schaefer)

ABSTRACT

Accurate theoretical prediction of molecular vibrational frequencies is important for informing chemical experiments. Coupled-cluster theory with a second-order perturbative treatment of anharmonicity (VPT2) can yield accurate frequencies to within 10-20 cm^{-1}, at least for semi-rigid molecules without resonances. Outside of these cases, however, VPT2 can be rendered useless for some of the most challenging chemical systems. Additionally, spectroscopy accuracy (within 1 cm^{-1}) remains out of reach for the vast majority of vibrational spectra. Herein, I investigate the methylene amidogen radical (H_2CN) as a case study for current theoretical techniques that treat vibrational motion. I analyze the error sensitivity of our state-of-the-art theoretical treatment and find that its limits of accuracy lie in the VPT2 treatment of anharmonicity. A next step toward spectroscopic accuracy might be vibrational configuration interaction, but this relies on the availability of an accurate energy surface, which is often prohibitively expensive. Motivated by this need, I discuss a machine learning technique called autoregressive Gaussian process modeling, which can reduce the computational cost of obtaining training data by leveraging relationships between low- and high-accuracy models. I apply this approach to the prediction of a chemical energy surface for the first time. The initial benchmarks presented here suggest that it can significantly improve learning efficiency.

Index Words: Computational chemistry, wavefunction methods, coupled cluster theory, basis set extrapolation, machine learning, transfer learning, Gaussian process regression, potential energy surfaces

TOWARD THE THEORETICAL PREDICTION OF MOLECULAR VIBRATIONAL FREQUENCIES WITH SPECTROSCOPIC ACCURACY

by

Avery Elizabeth Wiens

B.S., University of Georgia, 2015

A Dissertation Submitted to the Graduate Faculty
of the University of Georgia in Partial Fulfillment
of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2019

TOWARD THE THEORETICAL PREDICTION OF MOLECULAR VIBRATIONAL FREQUENCIES WITH SPECTROSCOPIC ACCURACY

by

Avery Elizabeth Wiens

Major Professor:	Henry F. Schaefer III
Committee:	Wesley D. Allen
	Henning H. Meyer

Electronic version approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
December 2019

DEDICATION

Great are the works of the Lord, studied by all who delight in them.

Full of splendor and majesty is his work, and his righteousness endures forever.

Psalm 111:2-3

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Motivation and Background

> Imagine that you hold a magic wand in your hand that controls the motion of the nuclei in any molecule. You can move these nuclei, at will, even to the most unlikely mutual arrangements. You may direct your molecule to fold itself into any conformation you wish and to undergo the most exotic, real or hypothetical, reactions with other molecules. In a bright numerical display your magic wand also shows, among other properties, the energy content of the actual molecular arrangement and you may test the energetics of any real or hypothetical molecule, transition structure, or intermediate. Such a magic wand would be a chemist's dream; he could predict the existence of new molecules and new reactions. He could devise new drugs for medicine and more economical synthetic processes for chemical industry[1] (Paul Mezey, 1987).

The laws of quantum mechanics establish a framework for both understanding and predicting the motion and interactions of molecules. The differential Schrödinger equation, written $\hat{H}\Psi = E\Psi$, is the equation of motion for quantum particles. A wavefunction Ψ is a solution to the Schrödinger equation under a given set of conditions. Under the set of conditions that specify a chemical system, the molecular wavefunction contains all the knowable information about a molecule in that specified configuration. If we were able to compute the exact molecular wavefunction of all conceivable arrangements of a molecule (or even a dense grid of arrangements), this would yield all the imaginable information about a chemical system. A magic wand is a fitting analogy; access to this information about even a handful of chemical systems could be an incredibly powerful tool for the research and development of fuels, pharmaceuticals, materials, and more.

As it stands in 2019, a practical realization of the chemist's dream described by Mezey in 1987 takes the form of the potential energy hypersurface, often referred to as the potential energy surface or PES. A PES is the representation of molecular energetics as a function of the internal nuclear coordinates of the molecule. The PES relies on the assumption that the Schrödinger equation can be solved independently for

electronic and nuclear coordinates of a molecule; this is called the Born-Oppenheimer approximation and will be discussed further in the next section. Even within the Born-Oppenheimer approximation, computing the exact wavefunction of a chemical system is computationally intractable for systems of more than a few atoms. It is this computational expense that restricts us from having unrestrained access to the "magic wand" that is the quantum mechanical wavefunction.

In reality, we must approximate solutions to the electronic Schrödinger equation to discern the wavefunction at each desired configuration of a molecule. If we are able to execute this computation with reasonable accuracy, we obtain a reliable though not exact interpretation of the molecule's electronic structure at the geometry of interest. For example, the widely popular CCSD(T) approximation is currently considered the "gold standard" for many electronic structure calculations. However, the CCSD(T) algorithm has $\mathcal{O}(n^7)$ time complexity with respect to basis set size and with current computing power is too expensive to treat molecular systems of more than 100 atoms.

Additionally, the $\mathcal{O}(n^7)$ time complexity only addresses the computational cost of obtaining a solution at one nuclear configuration. If we wanted to acquire a full picture of the electronic structure of a chemical system across many nuclear geometries, this becomes intractable much more quickly. For an illustrative example, consider the engine fuel ethanol, C_2H_6O, which has 21 nuclear degrees of freedom. If we constructed a grid by sampling from each degree of freedom at 10 points, this would require solving the CCSD(T) equations at 10^21 molecular geometries. Even in the ideal world where a CCSD(T) computation requires only one second and we have our own supercomputer available for this singular purpose, this computation would not finish in my lifetime. Clearly, there is a need for fast approximations for both solving the Schrödinger equation and for predicting a PES from a small subset of points in the nuclear configuration space. These approximations are, at present, the only avenue to bringing within reach the power and knowledge that Mezey described in his idea of the chemist's magic wand. The following sections discuss the current popular approaches along both of these axes, and suggest a new combination of approaches that may be helpful in achieving this.

1.2 The Born-Oppenheimer Approximation

The energetics of a molecule with M nuclei and N electrons can, in principle, be obtained from quantum mechanics by solving the time-independent molecular Schrödinger equation.

$$\hat{H}(\mathbf{r}, \mathbf{R})\Psi = E(\mathbf{r}, \mathbf{R})\Psi \tag{1.1}$$

where $\mathbf{r}$ and $\mathbf{R}$ are vectors of electronic and nuclear coordinates, respectively. As relativistic quantum chemistry is outside the scope of this work, we can assume a nonrelativilistic Hamiltonian $\hat{H}$, defined as follows.

$$\hat{H}(\mathbf{r},\mathbf{R}) = -\tfrac{1}{2}\sum_{i}^{N}\nabla_i^2 - \frac{1}{2}\sum_{A}^{M}\nabla_A^2 - \sum_{i,A}^{N,M}\frac{Z_A}{r_{iA}} + \sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A>B}^{M}\frac{Z_A Z_B}{R_{AB}} \tag{1.2}$$

Here, interparticle distances are represented by

$$r_{ij} = |\mathbf{r}_i - \mathbf{r}_j| \qquad r_{iA} = |\mathbf{r}_i - \mathbf{R}_A| \qquad R_{AB} = |\mathbf{R}_A - \mathbf{R}_B| \tag{1.3}$$

where $\mathbf{r}_i$ represents the position of the i^{th} electron, $\mathbf{R}_A$ the position of the A^{th} nucleus, and Z_A its nuclear charge.[2] Thus Equation 1.1 is a partial differential equation in $3(M+N)$ variables. For concreteness, this means that the quantum mechanical energy of benzene, which has 12 nuclei and 42 electrons, is a function of 162 spatial coordinates.

The Born-Oppenheimer approximation, proposed in 1927, states that the motion of nuclei and electrons in a molecule can be treated on separate timescales. This central concept in chemical physics allows the wavefunction of a molecule to be separated into its electronic and nuclear components, as

$$\Psi_{\text{total}} = \Psi_{\text{e}} \otimes \Psi_{\text{N}} \tag{1.4}$$

where Ψ_{e} is the electronic and Ψ_{N} is the nuclear wavefunction, containing vibrational, translational, and rotational components. It has proven to be a highly effective lens for the study of molecular quantum mechanics, because electrons are considerably lighter than nuclei and have faster classical velocities, by three or more orders of magnitude. Therefore, it is practical to separate nuclear and electronic degrees of freedom adiabatically, and solve the electronic Schrödinger equation with the nuclei "frozen" at nuclear coordinates $\mathbf{R}_\alpha$.[3,4] In this framework, the nuclear kinetic energy (the second term in Equation 1.1) is zero and the nuclear repulsion energy (the final term in Equation 1.1) is a constant. The purely electronic Schrödinger equation is then given by

$$\hat{H}_{\text{e}}\Psi_{\text{e}}\left(\mathbf{r};\mathbf{R}_\alpha\right) = E_{\text{e}}\left(\mathbf{R}_\alpha\right)\Psi_{\text{e}}\left(\mathbf{r};\mathbf{R}_\alpha\right). \tag{1.5}$$

Here $\hat{H}_{\text{e}}$ is the electronic Hamiltonian,

$$\hat{H}_{\text{e}}(\mathbf{r};\mathbf{R}) = -\frac{1}{2}\sum_{i}\nabla_i^2 - \sum_{iA}\frac{Z_A}{r_{iA}} + \sum_{i>j}\frac{1}{r_{ij}} + \sum_{A>B}\frac{Z_A Z_B}{R_{AB}} \tag{1.6}$$

and it is clear that the parametric dependence of $\hat{H}_e$ on $\mathbf{R}_\alpha$ arises in the $\{r_{iA}\}$ terms. Thus in the Born-

Oppenheimer framework, we choose the nuclear coordinates $\mathbf{R}_\alpha$ and solve Equation 1.5 at that fixed geometry. Returning to the benzene example, the electronic wavefunction at a given nuclear position is is solution to a partial differential equation in 126 (rather than 162) variables and is sufficient for a quantitatively accurate treatment of the molecule's electronic structure.

Once the electronic problem has been solved, it is possible to solve for the motion of the nuclei under the same adiabatic assumptions used to solve the electronic problem. The nuclear Schrödinger equation describes the vibrations, rotations, and translations of the system and is given by

$$\hat{H}_N \Psi_N(\mathbf{R}) = E_N \Psi_N(\mathbf{R}). \tag{1.7}$$

From the point of view of the slow-moving nuclei, it is appropriate to replace the electronic coordinates with their average values over the electronic wavefunction. This generates a nuclear Hamiltonian for the motion of the nuclei in the average field of the electrons.

$$\hat{H}_N(\mathbf{R};\mathbf{r}) = -\sum_{i=1}^{M} \frac{1}{2M_A}\nabla_A^2 + \left\langle -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i,A}^{N,M}\frac{Z_A}{r_{iA}} + \sum_{j>i}^{N}\frac{1}{r_{ij}} \right\rangle + \sum_{A>B}^{M}\frac{Z_A Z_B}{R_{AB}} \tag{1.8}$$

$$= -\sum_{i=1}^{M} \frac{1}{2M_A}\nabla_A^2 + E_e(\mathbf{R}) \tag{1.9}$$

The first and second terms given above in this expression are the kinetic and potential energies for nuclear motion, respectively. Thus the nuclei in a Born-Oppenheimer system move on a potential energy hypersurface (PES), a continuous manifold of instantaneous electronic states represented as a function of the nuclear coordinates. A simple, one-dimensional schematic is shown in Figure 1.1. The potential energy surface provides an important foundation for the theoretical prediction of many chemical phenomena. The familiar concepts of molecular geometries, molecular energetics, conformational analysis, vibrational spectroscopy, photochemistry, reactivity, transition structure, reaction mechanisms and intermediates, activation energy, reaction dynamics, and kinetics are intimately related to the concept of a PES.

Naturally, the quality of a PES depends on the electronic structure method used to calculate E_e. In a perfect world, we would compute an energy hypersurface by solving Equation 1.5 exactly at a sufficiently dense grid of points in the space of possible nuclear configurations. However, an exact solution to the electronic Schrödinger equation, even within the nonrelativistic and Born-Oppenheimer approximations, is so expensive that even one calculation is intractable for all but the smallest molecules. In reality, we must take advantage of approximate solutions to the Schrödinger equation. Computational electronic structure theory has a convergent hierarchy of such approximations, and the challenge for a quantum chemist is to

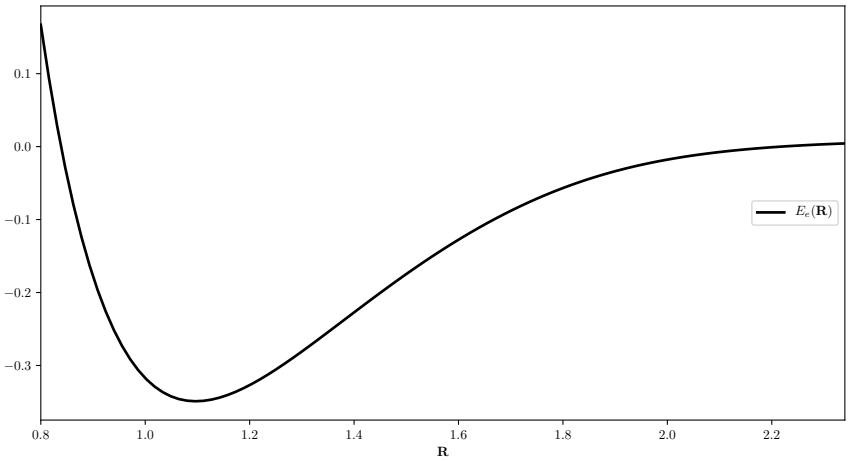

Figure 1.1: Schematic illustration of a potential energy surface for a diatomic molecule. For a simple one-dimensional system, this amounts to a dissociation curve.

select methods that provide the desired balance between accuracy and cost. The next section provides an overview of some of the commonly used techniques.

Before moving forward, it should be noted that the energy hypersurface model is not rigorously exact in its quantum mechanics, because it assumes that there is no coupling between electronic and nuclear motion. There are some complex cases where this is qualitatively inaccurate, including the famous cases of spin-orbit coupling, avoided crossings of energy surfaces, the Renner-Teller effect [5], charge-transfer reactions, and conical intersections. [6,7] Regardless, in most cases it is an extremely powerful tool and the foundation for our understanding of a wide variety of chemical phenomena. The scope of this work is limited to nonrelativistic Born-Oppenheimer problems. Methods for computing the electronic energy of Born-Oppenheimer systems at a specified configuration or variety of configurations are discussed in more formally in the next section.

1.3 Theoretical Treatments of Quantum Chemical Energies

1.3.1 Hartree-Fock Theory

A single N-particle Slater determinant is the simplest antisymmetric wavefunction that can be used to describe the ground state of an N-electron system. These N-particle determinants are made up of a set of one-electron functions $\{\chi_1, ..., \chi_N\}$ which are typically linear combinations of atom-centered Gaussian functions (atomic orbitals).

$$|\Phi\rangle = |\chi_1 \chi_2 \cdots \chi_N\rangle \tag{1.10}$$

The variational principle states that the best single Slater determinant wavefunction is the one that minimizes the following expectation value.

$$E_0 = \langle \Phi | \hat{H}_e | \Phi \rangle \tag{1.11}$$

The Hartree-Fock equations minimize the energy expectation value in the space spanned by the atomic orbitals $\{\chi_i\}$.

$$f(i)\chi(\mathbf{x}_i) = \varepsilon\chi(\mathbf{x}_i) \qquad\qquad f(i) = -\tfrac{1}{2}\nabla_i^2 - \sum_A^M \frac{Z_A}{r_{iA}} + v^{\mathrm{HF}}(i) \tag{1.12}$$

In a perfect world, we would calculate the "true" Hatree-Fock solution, *i.e.* the Slater determinant which minimizes the ground-state energy with respect to every possible choice of atomic orbitals. In other words, an ideal or "full" Hartree-Fock solution would be obtained by minimizing the energy in the complete (infinite) basis set of atomic orbitals. However, in practice we must use finite one-particle basis sets. This results in a different Hartree-Fock solution for every one-particle basis set. [2]

Even with a large one-particle basis set, the HF wavefunctions and resulting energies are not accurate enough to be used as anything more than a helpful starting point, because Hartree-Fock does not account for the correlation of electronic motion. In other words, it assumes that an electron will spend 100% of its time in the atomic orbital that minimizes the single-Slater determinant energy. In reality, electrons repel each other and sometimes excite into higher-lying unoccupied orbitals to avoid each other. This physical concept is called dynamic correlation, and while it accounts for a small fraction of the total electronic energy, it is a key contributor to the accurate description of chemical bonding and electronic structure. Regardless, many near-equilibrium systems can be at least qualitatively epproximated by Hartree-Fock theory. In these cases, the Hartree-Fock wavefunction can be used as a starting point for the treatment of dynamical correlation via one of many so-called "post-Hartree-Fock" theories.

1.3.2 The Full Configuration Interaction Limit

Although the Hartree-Fock wavefunction is not often sufficient on its own for achieving chemical accuracy, it is used as the starting point for building the fully correlated wavefunction. The full configuration interaction (FCI) wavefunction is composed of all possible Slater determinants generated from excitations out of the occupied molecular orbitals into all possible virtual orbitals. More formally, the FCI wavefunction is given by

$$|\Psi_{\mathrm{CI}}\rangle = \hat{C}|\Phi\rangle \qquad\qquad \hat{C} = \hat{C}_1 + \hat{C}_2 + ... + \hat{C}_N \tag{1.13}$$

where the operator $\hat{C}_i$ excites all combinations of i electrons from occupied HF orbitals to all combinations of virtual orbitals. Minimizing the energy with respect to $|\Psi_{\mathrm{CI}}\rangle$ yields the exact energy for a given one-particle basis set. Ideally, of course, we would use a complete one-particle basis and this would give us the exact FCI solution to Equation 1.5. As the errors resulting from the Born-Oppenheimer approximation and nonrelativistic Hamiltonian are typically very small, this technique would be sufficient to achieve sub-chemical accuracy in the majority of cases (with occasional first-order corrections to compensate for these effects). However, utilizing the full N-particle basis sets is impractical for all but the smallest molecules. The errors introduced through the use of finite (*i.e.* truncated) one- and N-electron basis sets are far more significant; this is a major limiting factor in the field of *ab initio* electronic structure theory.[2]

1.3.3 Møller-Plesset Perturbation Theory

Møller-Plesset Perturbation theory is one post-Hartree-Fock method for adding dynamic correlation into the wavefunction with a truncated N-particle basis set.[8] It is a special case of the Rayleigh-Schrödinger perturbation theory, a method from mathematical physics which splits the Hamiltonian into two parts. The first part is a "model" system, for which a solution is known (the HF solution) and a "perturbation", which is the more complex piece of the system and has an unknown solution. In this case, the perturbation is the difference between the FCI and HF solutions. The MPn level theory writes the perturbation as a Taylor expansion and truncates it at the n^{th} order. Unfortunately, this technique is not advisable for perturbative treatments greater than second-order (MP2), because it often oscillates or diverges rather than converging to the FCI limit.[9] For large molecules, however, MP2 is a good way to include some dynamical correlation at a very low additional cost on top of Hartree-Fock theory.

1.3.4 Coupled-Cluster Theory

As discussed in Subsection 1.3.2, when the full configuration interaction limit is not available, one must truncate the N-particle basis set. The most intuitive way to do this would be by simply truncating the FCI ansatz at the highest affordable excitation level, for example

$$|\Psi_{\mathrm{CISD}}\rangle = (\hat{C}_1 + \hat{C}_2)|\Phi\rangle. \tag{1.14}$$

However, this is not the most popular way of treating energies because it is not size-consistent. Instead, coupled-cluster (CC) theory uses an exponential ansatz to achieve robust size-consistency and converge more

quickly to the complete N-particle basis limit. The CC ansatz is defined as follows,

$$|\Psi_{\mathrm{CC}}\rangle = e^{\hat{T}}|\Phi\rangle \qquad \hat{T} = \hat{T}_1 + \hat{T}_2 + ... + \hat{T}_N \tag{1.15}$$

where $\hat{T}$ is the same as the $\hat{C}$ operator of CI theory; the two theories simply have different naming conventions for this operator.[10–12]

1.3.5 A Final Warning

The methods discussed herein assume that one single determinant can serve as a qualitative approximation to the wave-function. This is appropriate for a wide variety of chemical systems, particularly closed-shell and near-equilibrium molecules. One caveat is that this doesn't work where the ground electronic state can only be described a linear combination of two or more electronic configurations. In such a situtation, one single Slater determinant yields a qualitatively inaccurate set of orbitals, and thus it is inappropriate to use these as a starting point for treating higher-order correlation. This is called "strong correlation" and requires a multiconfigurational reference, which is outside the scope of this work.

1.4 An Overview of Surrogate Models in Quantum Chemistry

Quantum chemistry's first global surrogate model for a potential energy surface was Morse's diatomic potential, which was published in 1929 and still serves as the texbook model of diatomic bonding curves. The Morse potential requires three important training points to model a bonding curve. Building off of this idea, analytic functional forms were developed for certain classes of tri- and tetra-atomic systems, but progress in this direction had largely stalled until Bowman and coworkers developed their permutation invariant polynomial (PIP) expansion in 2003. Since then, global analytic surfaces have been successfully developed for systems of up to ten atoms.[13–15] However, as molecules get larger and more complex, it becomes difficult to develop a general functional form for large "classes" of molecules and one runs the risk of missing out on unexpected features.

In response to these challenges, surrogate models based on multivariate interpolation became popular. These methods connect observed energy points using piecewise functions rather than fitting them to one global functional form. Some of the most common examples include the spline[16], modified Shepard[17], and reproducing kernel Hilbert space methods.[18] The main trade-off between least-squares fitting and interpolation methods is in their flexibility. Interpolation methods require a higher density of training points to achieve qualitative accuracy, which is a disadvantage in quantum chemistry, but they improve systematically

with the size of the training set and can represent unusual surface topologies that might be glossed over by a physically motivated functional form.

In the last decade, there has been a surge of interest in machine learning surrogate models, which have been very successful in the technology sector, as an avenue for combining the strengths of these two classes of surrogate models. The bulk of these articles has centered around artificial neural network (ANN) potentials as put forth by Behler[19,20] and followed by many others.[21–23] Unlike the physically motivated analytic functional forms, ANNs have an extremely flexible functional form and can model complex, unique, and unexpected surface topologies. This is an advantage enabled by their inherent lack of structure; it is this same quality, however, that causes them to require a high density of training examples. Thus they are highly effective in most industry applications, where the cost of obtaining training data is often very small. With regard to high-accuracy energy surfaces, which are extremely expensive in comparison, their utility will be necessarily limited.

In contrast an alternative machine learning technique known as Gaussian process (GP) regression and kernel ridge regression has been comparatively underrepresented in the scientific literature. This method has received growing attention in the chemical physics in the past five years but is not commonly used for energy surface regression.[24–28] Preliminary benchmark studies have shown that GP regression can provide an order-of-magnitude reduction in training set size relative to ANN surrogate models.[29] Herein, a multi-fidelity Gaussian process regression technique known as nonlinear autogressive Gaussian process modeling is applied for the first time to the fitting of a high-accuracy chemical energy surface. This method and the following results are discussed in the second chapter of this work, which identifies it as a promising candidate for making energy surface regression more affordable.

CHAPTER 2

REINTERPRETING THE INFRARED SPECTRUM OF H + HCN: METHYLENE AMIDOGEN RADICAL AND ITS COPRODUCTS

[1]

[1]A. E. Wiens, E. C. Rossomme, G. J. R. Aroeira, O. M. Bernstein, A. V. Copan, H. F. Schaefer, and J. Agarwal. Accepted by the Journal of Chemical Physics. Reprinted here with permission of publisher.

2.1 Abstract

The methylene amidogen radical (H_2CN) plays a role in high-energy material combustion and extraterrestrial atmospheres. Recent theoretical work has struggled to match experimental assignments for its CN and antisymmetric CH_2 stretching frequencies (ν_2 and ν_5), which were reported to occur at 1725 and 3103 cm^{-1}. Herein, we compute the vibrational energy levels of this molecule by extrapolating quadruples-level coupled-cluster theory to the complete basis limit and adding corrections for vibrational anharmonicity. This level of theory predicts that ν_2 and ν_5 should occur at 1646 and 2892 cm^{-1}, at odds with the experimental assignments. To investigate the possibility of defects in our theoretical treatment, we analyze the sensitivity of our approach to each of its contributing approximations. Our analysis suggests that the observed deviation from experiment is too large to be explained as an accumulation of errors, leading us to conclude that these transitions were misassigned. To help resolve this discrepancy, we investigate possible byproducts of the H + HCN reaction, which was the source of H_2CN in the original experiment. In particular, we predict vibrational spectra for *cis*-HCNH, *trans*-HCNH, and H_2CNH using high-level coupled-cluster computations. Based on these results, we reassign the transition at 1725 cm^{-1} to ν_3 of *trans*-HCNH, yielding excellent agreement. Supporting this identification, we assign a known contaminant peak at 886 cm^{-1} to ν_5 of the same conformer. Our computations suggest that the peak observed at 3103 cm^{-1}, however, does not belong to any of the aforementioned species. To facilitate further investigation, we use structure and bonding arguments to narrow the range of possible candidates. These arguments lead us to tentatively put forth formaldazine [$(H_2CN)_2$] as a suggestion for further study, which we support with additional computations.

2.2 Introduction

Beginning with Herzberg's work in the 1920's,[30–32] advances in molecular spectroscopy have been attended by the increasingly precise measurement of small-molecule spectroscopic constants. Parallel developments in quantum chemistry have proven indispensable to this effort, due to the difficulty of interpreting complex spectra. The dissociation energy of H_2, equilibrium geometry of CH_2, and electronic structure of O_2 are prominent cases in which theory has spurred reinterpretations of experimental data.[33–36] In light of such challenges, rigorous spectroscopic assignment is best accomplished by attaining experimental and theoretical agreement within tight error bounds. This motivates the development of computational procedures which utilize theoretical tools to their full potential.

This work focuses on the methylene amidogen radical (H_2CN) as a case study for the refinement of theoretical methods. The H_2CN radical is an intermediate in the decomposition of nitramine propellants, a

class of high-energy molecules that includes the explosives HMX and RDX.[37–40] It has also been detected in the Taurus molecular cloud TMC-1[41] and is a likely precursor to the formation of HCN and HNC in several extraterrestrial atmospheres.[42,43] As a key player in these combustion and astrochemical processes, the H_2CN radical has been studied in numerous spectroscopic experiments[44–55] since it was first observed in 1962.[56]

In particular, three experimental studies have reported vibrational frequencies for the $\tilde{X}\ ^2B_2$ ground state of H_2CN.[53–55] In 1987, Jacox recorded an infrared spectrum of this radical in an argon matrix isolation experiment and assigned five peaks to ν_2, ν_3, ν_4, ν_5, and ν_6.[53] Four years later, Ellison and coworkers employed gas phase photoelectron spectroscopy with Franck-Condon simulations to observe transitions ν_1 and ν_3.[54] Lastly, a 1998 study by Räsänen and coworkers corroborated three of these previous assignments by detecting ν_3, ν_4, and ν_6 in xenon and krypton media.[55]

The vibrational spectrum of the methylene amidogen radical has also been studied using computational methods.[57–61] The most reliable theoretical work was carried out by Puzzarini and Barone, who presented coupled-cluster harmonic frequencies with B3LYP anharmonic corrections.[59,60] Even at this level of theory, they found significant differences from experiment for two of the fundamental transitions, ν_2 and ν_5, raising the possibility of misassignment. However, this finding was necessarily tentative, due to their use of a heavily spin contaminated ($\langle \hat{S}^2 \rangle > 0.95$) reference determinant, which can sometimes produce spurious results.[62–64]

Herein, we revisit the ground state fundamentals of the H_2CN radical with a view to identifying the source of the aforementioned discrepancies. Geometric parameters and vibrational frequencies are computed with a spin-restricted reference at the basis and correlation limits of coupled-cluster theory, using up to six-zeta basis sets and connected quadruple excitations. More importantly, we analyze the sensitivity of our methodology to each of its contributing approximations, in order to explore all possible sources of deviation from experiment. After concluding that the discrepancies observed for ν_2 and ν_5 cannot be explained as an unlucky error accumulation, we contribute additional analysis and computations investigating possible contaminants of the experimental reaction mixture.

2.3 Computational Methods and Software

Restricted open-shell Hartree-Fock (ROHF) reference determinants were employed for all computations. The frozen core approximation was used throughout, except where we explicitly denote that all electrons were correlated with AE. Single point energies were converged to within 10^{-10} E_h for all computations, and geometric parameters were converged to achieve a threshold of at least 10^{-7} E_h/a_0 for the RMS force. Coupled-cluster (CC) single point energies with up to perturbative triple excitations[10–12,65] [CCSD(T)]

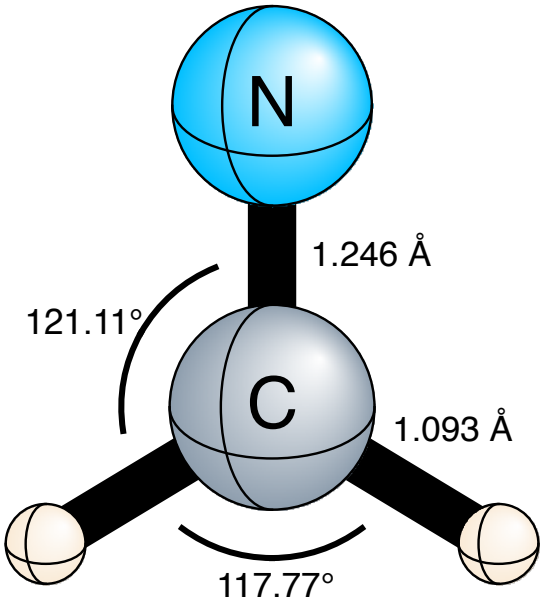

Figure 2.1: Planar equilibrium molecular structure for the ground ($\tilde{\mathrm{X}}\ ^2B_2$) electronic state of H_2CN, optimized at the CCSDT(Q)/CBS level of theory.

were computed in Molpro[66,67]. Higher order CC computations[68] were executed in MRCC.[68–72] Basis set extrapolated geometries and frequencies were obtained by interfacing the aforementioned energy routines with Psi4's optimization and finite difference modules[73]. Exact two component (X2C) relativistic corrections[74] were computed in Psi4[73]. Diagonal Born-Oppenheimer corrections[75,76] (DBOC) were performed in CFour.[77] CI computations[78,79] were carried out in Molpro. VPT2 anharmonic corrections[80,81] were computed with CFour and PyVPT2[82].

2.4 Basis extrapolations and additive corrections

The geometric parameters and harmonic frequencies in Figure 2.1 and Tables 2.2 and 2.3 are determined from composite energies using the focal point approach of Allen and coworkers.[83–85] These composite energy formulae involve additive energy corrections and basis set extrapolations. Additivity corrections derive from the assumption that the difference between two similar levels of theory converges rapidly in the basis set limit. These corrections have the following general form.

$$E^{\text{theory 2}}_{\text{basis 2}} \approx E^{\text{theory 1}}_{\text{basis 2}} + \left(E^{\text{theory 2}}_{\text{basis 1}} - E^{\text{theory 1}}_{\text{basis 1}} \right) \tag{2.1}$$

Since this formula is linear in the energies, additivity corrections have the same form for energy derivatives. For example, force constants can be approximated as follows,

$$\frac{\mathrm{d}E^{\text{theory 2}}_{\text{basis 2}}}{\mathrm{d}x_i \mathrm{d}x_j} \approx \left(\mathbf{H}^{\text{theory 1}}_{\text{basis 2}} + \mathbf{H}^{\text{theory 2}}_{\text{basis 1}} - \mathbf{H}^{\text{theory 1}}_{\text{basis 1}}\right)_{ij} \tag{2.2}$$

where $\mathbf{H}$ denotes the computed Hessian matrix. We have used this assumption to determine higher-order and core correlation corrections as well as relativistic and adiabatic corrections.

With respect to basis set extrapolation, FPA prescribes a three-point extrapolation for Hartree-Fock energies[86,87] and a two-point extrapolation for correlation energies.[88] Defining E_n as the energy computed at basis set cardinality n, the formulae are given by:

$$E^{\text{hf}}_{\infty} = E^{\text{hf}}_{n} - \frac{(E^{\text{hf}}_{n} - E^{\text{hf}}_{n-1})^2}{E^{\text{hf}}_{n} - 2E^{\text{hf}}_{n-1} + E^{\text{hf}}_{n-2}} \tag{2.3}$$

$$E^{\text{c}}_{\infty} = E^{\text{c}}_{n} - \frac{n^{-3}\left(E^{\text{c}}_{n} - E^{\text{c}}_{n-1}\right)}{n^{-3} - (n-1)^{-3}} \tag{2.4}$$

Differentiating these energy expressions allows us to apply the same extrapolation scheme to gradients and force constants. Thus, a correlated Hessian is extrapolated to the basis set limit via the second derivatives of Equation 2.4:

$$\frac{\mathrm{d}^2 E^{\text{c}}_{\infty}}{\mathrm{d}x_i \mathrm{d}x_j} = (\mathbf{H}^{\text{c}}_{n})_{ij} - \frac{n^{-3}\left[(\mathbf{H}^{\text{c}}_{n})_{ij} - (\mathbf{H}^{\text{c}}_{n-1})_{ij}\right]}{n^{-3} - (n-1)^{-3}} \tag{2.5}$$

Derivatives of Equation 2.3 are more complicated. Defining intermediates $\Delta_n = E^{\text{hf}}_{n} - E^{\text{hf}}_{n-1}$ and $\Delta'_n = \Delta_n - \Delta_{n-1}$, second derivatives of Equation 2.3 have the form

$$\begin{aligned}\frac{\mathrm{d}^2 E^{\text{hf}}_{\infty}}{\mathrm{d}x_i \mathrm{d}x_j} = (\mathbf{H}^{\text{hf}}_{n})_{ij} &- \frac{2\Delta_n}{\Delta'_n}\frac{\mathrm{d}^2\Delta_n}{\mathrm{d}x_i \mathrm{d}x_j} - \frac{2}{\Delta'_n}\frac{\mathrm{d}\Delta_n}{\mathrm{d}x_i}\frac{\mathrm{d}\Delta_n}{\mathrm{d}x_j} - \frac{2\Delta_n^2}{(\Delta'_n)^3}\frac{\mathrm{d}\Delta'_n}{\mathrm{d}x_i}\frac{\mathrm{d}\Delta'_n}{\mathrm{d}x_j} + \left(\frac{\Delta_n}{\Delta'_n}\right)^2 \frac{\mathrm{d}^2\Delta'_n}{\mathrm{d}x_i \mathrm{d}x_j} \\ &+ \frac{2\Delta_n}{(\Delta'_n)^2}\left(\frac{\mathrm{d}\Delta'_n}{\mathrm{d}x_j}\frac{\mathrm{d}\Delta_n}{\mathrm{d}x_i} + \frac{\mathrm{d}\Delta_n}{\mathrm{d}x_j}\frac{\mathrm{d}\Delta'_n}{\mathrm{d}x_i}\right)\end{aligned}$$

where the derivatives of Δ_n and Δ'_n are sums of computed gradient vectors or Hessian matrices.

2.5 Theoretical analysis of ν_2 and ν_5

In this section, we report vibrational frequencies for fundamentals ν_2 and ν_5 at the complete basis set (CBS) limit of coupled-cluster theory including up to connected quadruple excitations [CCSDT(Q)]. Additionally, we consider the magnitudes of error introduced by each of the assumptions in our theoretical approach–

namely, the use of (1) a single reference determinant, (2) a truncated excitation level, (3) a finite basis set, (4) the Born-Oppenheimer approximation, (5) the nonrelativistic Hamiltonian, and (6) the second-order perturbative treatment of anharmonicity. By investigating the sensitivity of this system to each of these assumptions, we are able to estimate error bounds on our methodology.

The accuracy of coupled-cluster theory rests on the simplifying assumption that a single determinant provides a qualitatively accurate zeroth order wavefunction. To assess the validity of this approach, we compare harmonic frequencies computed with a spin-restricted reference determinant to those computed with an optimized multiconfigurational reference. In both cases, dynamical correlation is treated using singles and doubles configuration interaction with a Davidson correction[89] (CISD+Q). These two approaches agree to within 25 cm^{-1} for ω_2 and 8 cm^{-1} for ω_5, as shown in Table 2.1. The small magnitude of these shifts demonstrate us that the system is relatively insensitive to multireference effects. Since the exponential ansatz of coupled-cluster theory is even less sensitive to the number of reference determinants than CI,[90] we seem to be on sure footing here.

To further quantify convergence to the limit of nonrelativistic Born-Oppenheimer theory, we have tabulated incremental changes in the harmonic frequencies using the focal point approach (FPA) of Allen and coworkers[83–85] and Dunning's correlation consistent polarized basis sets (cc-pVXZ, X = D, T, Q, 5, 6).[91] Table 2.2 shows the incremental convergence of ω_2 and ω_5 with respect to basis set and level of theory. These convergence tables demonstrate that the inclusion of connected quadruple excitations contributes 3.7 cm^{-1} to the final harmonic frequency for ω_2 and 1.3 cm^{-1} for ω_5. Furthermore, the basis set extrapolation contributes 1.4 cm^{-1} to the harmonic frequency for ω_2 and 0.1 cm^{-1} for ω_5. It is evident that convergence to the basis set limit has effectively been achieved with cc-pV6Z, and that combining this with the CCSDT(Q) treatment of electron correlation is sufficient to yield harmonic frequencies with accuracy on the order of 1 cm^{-1}.

Traditional coupled-cluster theory also relies on the nonrelativistic and Born-Oppenheimer approxima-

Table 2.1: Harmonic vibrational frequencies (ω) for the $\tilde{X}\ ^2B_2$ state of H_2CN, obtained using the internally contracted MR-CISD+Q method. Results are compared to single reference CISD+Q/cc-pVTZ

Mode	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
MRCI	3024.5	1667.5	1395.0	950.6	3095.2	985.9
CI	3031.9	1692.4	1401.6	994.4	3102.7	957.1

[a]Reference wave function computed with full valence CASSCF, using an active space including orbitals with natural orbital occupation number < 0.025.

Table 2.2: Second derivative convergence of ω_2 and ω_5, obtained using focal point analysis and Dunning's correlation consistent polarized basis sets. Values listed in brackets were obtained using the additivity assumption for force constants as detailed in Equation 2.2. Values listed in bold were extrapolated to the CBS limit as described in Section II.B. The rightmost column shows the incremental differences for total CCSDT(Q) frequencies with respect to basis set cardinality.

	SCF	δMP2	δD[a]	δ(T)[a]	δT[a]	δ(Q)[a]	
				ω_2			
DZ	1862.2	−91.1	17.0	−14.4	−0.2	−2.9	$\downarrow \delta$Z
TZ	1815.2	−110.2	17.0	−16.8	0.4	−3.7	−68.7
QZ	1809.0	−120.5	21.9	−17.8	0.7	[−3.7]	−12.2
5Z	1807.8	−121.5	21.0	−17.9	[0.7]	[−3.7]	−3.2
6Z	1807.6	−120.6	19.6	−18.2	[0.7]	[−3.7]	−1.1
CBS	**1807.5**	**−119.3**	**16.6**	**−17.9**	[**0.7**]	[**−3.7**]	−1.4
				ω_5			
DZ	3287.8	−18.3	−26.0	−8.9	−1.4	−1.1	$\downarrow \delta$Z
TZ	3172.7	−49.6	−30.4	−11.1	−1.4	−1.3	−153.2
QZ	3169.8	−53.2	−26.2	−11.4	−1.3	[−1.3]	−2.6
5Z	3169.9	−55.4	−26.0	−11.5	[−1.3]	[−1.3]	−2.0
6Z	3169.8	−55.9	−25.5	−11.5	[−1.3]	[−1.3]	−0.1
CBS	**3169.7**	**−57.4**	**−24.0**	**−11.6**	[**−1.4**]	[**−1.3**]	−0.1

[a] δn denotes a CCSD$\cdots n$ treatment of electron correlation
[b] Final frequencies are eigenvalues of the CCSDT(Q) Hessian, obtained from basis set extrapolations (Equations 2.3, 2.4) and additivity corrections (Equation 2.2).

$$\mathbf{H}_{\infty}^{(\mathrm{Q})} = \mathbf{H}_{\infty}^{\mathrm{hf}}(n{=}6) + \mathbf{H}_{\infty}^{\mathrm{c:(T)}}(n{=}6) + \mathbf{H}_4^{\mathrm{T}} - \mathbf{H}_4^{(\mathrm{T})} + \mathbf{H}_3^{(\mathrm{Q})} - \mathbf{H}_3^{\mathrm{T}}$$

tions. To investigate the validity of these assumptions, we have reported relativistic and adiabatic corrections for each of the harmonic vibrational frequencies of H_2CN in Table 2.3. The adiabatic corrections ($\delta\omega_{\mathrm{ad}}$), computed as the diagonal Born-Oppenheimer correction[75,76] at the ROHF/cc-pVTZ level of theory, are shown to be on the order of 0.1 cm^{-1} for both ω_2 and ω_5, justifying the use of the Born-Oppenheimer approximation. Relativistic corrections ($\delta\omega_{\mathrm{rel}}$), computed using the exact two-component approach[74] (X2C) at the CCSD/cc-pVTZ level of theory, are on the order of 0.01 - 0.1 cm^{-1}, justifying the use of a nonrelativistic Hamiltonian.

Table 2.3: Fundamental vibrational transitions (ν) for the $\tilde{X}\ ^2B_2$ state of the methylene amidogen radical. Geometric parameters (see Figure 2.1) and harmonic frequencies (ω) have been determined at the CCSDT(Q)/CBS level of theory using the focal point scheme in Table 2.2 with an additional core correction, computed as the difference of frozen core and all electron CCSD(T)/cc-pCVQZ gradients and force constants. Adiabatic ($\delta\omega_{ad}$) corrections were determined using the diagonal Born-Oppenheimer correction (DBOC) at the Hartree-Fock cc-pVTZ level of theory, and relativistic corrections ($\delta\omega_{rel}$) were determined using the exact two-component approach (X2C) at the CCSD/cc-pVTZ level of theory. Relative harmonic intensities (reported in parentheses) and VPT2 anharmonic corrections were determined at the AE-CCSD(T)/cc-pCVQZ level of theory.

Mode	Sym	Description	TED[a](%)	This Work					Ar[53]	gas[54]
				ω	$\delta\omega_{ad}$[b]	$\delta\omega_{rel}$[b]	$\delta\nu$	ν[c]	ν	ν
ν_1	a_1	CH_2 s-stretch	100%	2996.1	1.49	−0.01	−168.1	2829.5 (0.1)		2820
ν_2	a_1	CN stretch	91%	1677.5	0.12	0.34	−32.0	1646.0 (7)	1725	
ν_3	a_1	CH_2 scissor	90%	1379.1	−0.06	−0.44	−38.2	1340.4 (40)	1337	1337
ν_4	b_1	Out of plane	100%	973.8	−0.42	0.06	−19.3	954.1 (100)[d]	954	
ν_5	b_2	CH_2 a-stretch	100%	3068.0	1.44	0.09	−178.0	2891.5 (22)	3103	
ν_6	b_2	CH_2 rock	100%	937.4	−0.52	−0.04	−21.3	915.5 (27)	913	

[a]Total energy distributions[92] (TED), calculated at the CCSD(T)/cc-pVTZ level of theory in INTDER[93], were used to classify the normal modes.
[b]Additive energy corrections determined using Equation 1.
[c]$\nu = \omega + \delta\nu + \delta\omega_{ad} + \delta\omega_{rel}$
[d]Absolute intensity: 33 km mol^{-1}.

Finally, corrections for anharmonicity were computed with second-order vibrational perturbation theory (VPT2)[80,81] at the CCSD(T)/cc-pCVQZ level of theory with all electrons correlated (AE). These corrections ($\delta\nu$), shown in Table 2.3, redshift the fundamentals by 32 cm^{-1} (1.9%) for ω_2 and 178 cm^{-1} (5.8%) for ω_5. No anharmonic resonances were observed. Furthermore, the molecule has no floppy vibrational motions or significant vibrational coupling, making this a best-case scenario for the treatment of anharmonicity. In conjunction with coupled-cluster theory, VPT2 yields excellent results under these conditions, typically achieving accuracy to within 10 cm^{-1} or less.[85,94–96]

By assuming worst-case scenarios for each of the aforementioned sources of error, we propose 1% as a generous margin of uncertainty on our predictions for ν_2 and ν_5. On the experimental side, the dominant source of deviation from the gas phase frequencies is caused by argon matrix perturbation effects. Surveying a large database of neutral molecules, Jacox reported a root-mean-square (RMS) deviation of 0.9% between argon matrix and gas phase fundamentals.[97] Thus, typical argon matrix shifts for ν_2 and ν_5 might be about 15 cm^{-1} and 27 cm^{-1}, respectively. Even the most unfortunate constructive interference of errors cannot reconcile discrepancies of 79 cm^{-1} (4.8%) for ν_2 and 211 cm^{-1} (7.3%) for ν_5 at this level of theory. We conclude that the peaks observed by Jacox at 1725 and 3103 cm^{-1} have been incorrectly assigned.

Before moving on, we note as an aside that the misassignment of ν_5 is also supported by chemical intuition. Attributing ν_5 to Jacox's peak at 3103 cm^{-1} implies an unphysically large splitting of the symmetric and antisymmetric CH_2 stretching modes for an $H_2C{=}X$ system. For reference, experimental results for H_2CO and H_2CS show splittings of 61 and 54 cm^{-1}, respectively.[98,99] Theoretical calculations predict that H_2CP, the phosphoric congener of H_2CN, displays a 64 cm^{-1} splitting. These cases line up with our theoretical prediction of 62 cm^{-1} splitting and disagree with the experimental prediction of 283 cm^{-1} splitting.

2.6 Suggested reassignments

The most likely contaminants of the H + HCN reaction Jacox used to generate H_2CN are the *cis* and *trans* isomers of HCNH. Bair and Dunning's 1985 investigations of this reaction[100], summarized in Figure 2.2, indicate that while H_2CN is the major product under kinetic control, direct formation of *cis*-HCNH should be feasible at the experimental temperature of 14 K. From there, the forward barrier to *cis*-*trans* interconversion lies below even the initial barrier to H_2CN formation, suggesting that formation of *cis*-HCNH would likely be followed by conversion to the lower energy *trans* conformer.[100] Indeed, Jacox noted this possibility in her original study and attributed one of her observations to a CNH bend of HCNH, but was unable to distinguish between *cis* and *trans*. Jacox also suggested the possibility of observing H_2CNH as a contaminant species, but did not assign any transitions to this system. We have investigated the possibility that the absorptions at 1725 and 3103 cm^{-1} are due to any of these three contaminants by computing VPT2/AE-CCSD(T)/cc-pCVQZ fundamental vibrational frequencies for each of them. Our predictions are shown in Table 2.4, along with available experimental data. We note excellent agreement for H_2CNH (average 0.3% difference), for which gas phase data are available.

Absorption at 1725 cm^{-1}. Based on our computations, the closest match for the absorption at 1725 cm^{-1} is the CN-stretch (ν_3) of *trans*-HCNH. This differs by only 0.3% from our computed value of 1730 cm^{-1}, and the sign of the difference is consistent with the tendency of argon matrices to red-shift stretching frequencies.[97] The small magnitude of the predicted matrix perturbation is consistent with chemical intuition, since axial interactions of the CN-stretch with the argon matrix are buffered by terminal hydrogens. Furthermore, theory predicts that the CN-stretch of H_2CN has an intensity of about 2 km mol^{-1}, whereas theCN-stretch of *trans*-HCNH is predicted to be about 10 times more intense.

Additionally, we note that the contaminant peak observed by Jacox[53] at 886 cm^{-1} is in good agreement (0.5% difference) with our prediction of 891 cm^{-1} for the CNH bend (ν_5) of *trans*-HCNH. An impressive theoretical intensity of 218 km mol^{-1} for this vibration is also consistent with this assignment. Along with the thermodynamic arguments for the presence of *trans*-HCNH and consistency with theoretical intensities,

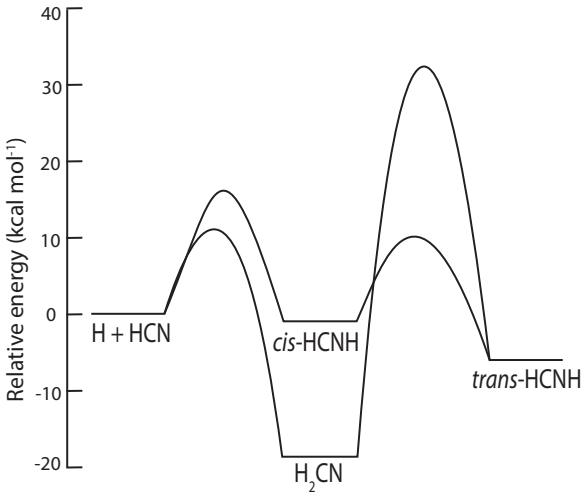

Figure 2.2: Qualitative energy profile for H + HCN and its reaction products, adapted from prior theoretical studies, Refs 60 and 100.

its ability to explain two of the prominent features in Jacox's spectrum to well within experimental and theoretical error bounds presents a strong case for reassigning the absorption at 1725 cm^{-1} to ν_3 of *trans*-HCNH.

Absorption at 3103 cm^{-1}. Of the candidates in Table 2.4, the closest numerical match for the absorption at 3103 cm^{-1} is the NH-stretch (ν_1) of *cis*-HCNH, 33 cm^{-1} away. However, this transition is ruled out by Jacox's isotopic substitutions, which show that the peak originates from a CH stretch. The closest CH stretch of the candidates in Table 2.4 is ν_2 of H_2CNH, which is 91 cm^{-1} away and is ruled out by a separate argon matrix study (Ref 101).

In other words, Table 2.4 seems to eliminate all of the obvious secondary products of the H + HCN reaction as possible sources for this feature. Chemical intuition points towards the same conclusion. Standard group frequency tables predict that sp^2 hybridized CH bonds should absorb around 3000-3100 cm^{-1}, but we can make more specific predictions using periodic trends. As shown in Figure 2.3, the CH_2 antisymmetric stretch frequency for an H_2CX system blue-shifts with decreasing X-electronegativity, starting from formaldehyde at 2843 cm^{-1} and monotonically approaching the limit of singlet methylene at 3190 cm^{-1}. Based on this trend, the peak at 3103 cm^{-1} should belong to an sp^2-hybridized CH bond attached to something roughly as electronegative as the methylene functional group.

Our analysis in Figure 2.3 leads us to consider other molecules which might land at the appropriate place in this trend. One candidate which has not been considered in previous studies is formaldezine [$(H_2CN)_2$], the product of H_2CN dimerization. The formation of an N−N bond between two H_2CN radicals could blue-shift the CH_2 antisymmetric stretching frequencies sufficiently to absorb in the 3100 cm^{-1} region. Theory

Table 2.4: Fundamental vibrational transitions (ν) for the ground state equilibrium geometries of *cis*-HCNH, *trans*-HCNH, and H_2CNH, computed using VPT2 at the AE-CCSD(T)/cc-pCVQZ level of theory. Relative harmonic intensities from the same level of theory are given in parentheses.

Mode	Sym	Description		
trans-HCNH				This Work
ν_1	a'	NH stretch		3277.3 (5)
ν_2	a'	CH stretch		2902.0 (9)
ν_3	a'	CN stretch		1730.6 (9)
ν_4	a'	HCN bend		1161.9 (5)
ν_5	a'	CNH bend		891.1 (100)[a]
ν_6	a''	HCNH torsion		962.7 (45)
cis-HCNH				This Work
ν_1	a'	NH stretch		3135.8 (2)
ν_2	a'	CH stretch		2844.7 (29)
ν_3	a'	CN stretch		1770.4 (16)
ν_4	a'	HCN bend		991.0 (100)[b]
ν_5	a'	CNH bend		843.6 (86)
ν_6	a''	HCNH torsion		890.1 (3)
H_2CNH			Gas[102–104]	This Work
ν_1	a'	NH stretch	3262.6	3283.2 (2)
ν_2	a'	CH stretch	3024.5	3017.2 (56)
ν_3	a'	CH stretch	2914.2	2903.8 (85)
ν_4	a'	CN stretch	1638.3	1644.4 (26)
ν_5	a'	CH_2 scissor	1452.0	1455.3 (13)
ν_6	a'	HCNH deform.	1344.3	1348.2 (91)
ν_7	a'	HCNH deform.	1058.2	1060.6 (81)
ν_8	a''	Torsion	1127.0	1132.1 (100)[c]
ν_9	a''	Out of plane	1063	1065.0 (35)

[a]Absolute intensity: 219 km mol^{-1}.
[b]Absolute intensity: 133 km mol^{-1}.
[c]Absolute intensity: 44 km mol^{-1}.

predicts that this dimerization is barrierless with a 53 kcal mol^{-1} binding energy[105], so it seems at least plausible that this secondary reaction could have occurred in the discharge or matrix. To investigate this possibility, we have computed VPT2 fundamental transitions for formaldazine with CCSD(T), employing a cc-pCVQZ basis set for harmonic frequencies and cc-pCVTZ basis for anharmonic corrections. This level of theory predicts that formaldazine has a CH-stretch at 3075 cm^{-1} which differs by 28 cm^{-1} (0.9%) from the peak observed in experiment. Since this matches Jacox's observation to within the standard argon matrix

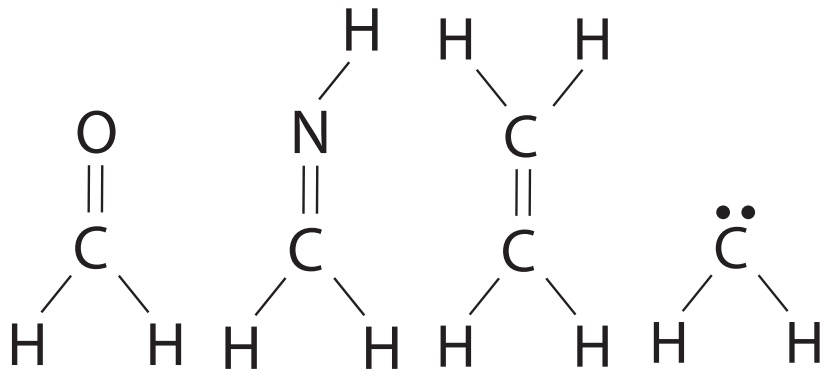

Figure 2.3: Comparison of CH_2-antisymmetric stretching vibrational frequencies for various sp^2-hybridized carbon atoms. The strength of CH bonds increases from left to right, causing a red-shift in the stretching frequencies. The CH_2 frequency listed at 3190 cm^{-1} corresponds to singlet methylene.

pertubration, we suggest that the CH-stretch of formaldazine is worth investigating in future experiments.

2.7 Remaining Fundamental Transitions

Having addressed ν_2 and ν_5 in detail, we now briefly turn to our predictions for the remaining vibrational fundamentals. Our final predictions are detailed in Table 2.3, computed at the same CCSDT(Q)/CBS level of theory with anharmonic and other corrections. They are compared to Jacox's matrix isolation study and, where possible, the more recent gas-phase photoelectron study by Ellison and coworkers.

Our final values show excellent agreement with the two gas phase fundamentals measured by Ellison and coworkers for H_2CN. In their experiment, they assigned transitions at 2820 and 1337 cm^{-1} to the CH_2 symmetric stretching vibration (ν_1) and the CH_2 scissor bend (ν_3). Our computational predictions differ from these values by 10 cm^{-1} and 3 cm^{-1}, each yielding a 0.3% difference. We also find quantitative agreement with the peaks observed in Jacox's study at 954 and 913 cm^{-1}, corresponding to the out of plane bend (ν_4) and the CH_2 rock (ν_6) of H_2CN. Our predictions for these transitions differ from her measurements by 0.1 cm^{-1} (0.01%) for ν_4 and 2.5 cm^{-1} (0.3%) for ν_6. These differences are well within one RMS deviation (0.9%) for argon matrix perturbations.

2.8 Conclusions

We have reported vibrational frequencies for the H_2CN radical at the basis and correlation limits of non-relativistic Born-Oppenheimer theory. These computations are based on a spin restricted reference and include a detailed analysis of the sensitivity of our approach to multireference effects, high-order dynamical correlation, basis set size, the Born-Oppenheimer approximation, relativistic effects, and vibrational anharmonicity. From these results we have concluded that, for ν_2 and ν_5, theory converges to a result which is irreconcilable with current experimental assignments. We have proposed that the observed peaks are instead due to contaminant species. For the peak at 1725 cm^{-1}, originally assigned to ν_2 of H_2CN, we propose a reassignment to the CN stretch (ν_3) of *trans*-HCNH. This is supported by our second assignment, which attributes a contaminant feature observed by Jacox at 886 cm^{-1} to the CNH bend (ν_5) of this conformer. We also present evidence that the peak at 3103 cm^{-1}, originally attributed to ν_5 of H_2CN, is not due to any of the obvious byproducts of the $H + HCN$ reaction, and point to formaldazine as a promising candidate for future investigation.

In addition to revising the assignment of the vibrational spectrum of H_2CN, this work has sought to develop a theoretical approach which takes full advantage of the present state-of-the-art in quantum chemical software. To conclude, we briefly note some ways in which our methodology could be improved in the future. The dominant error in our predictions comes from our second-order perturbative treatment of vibrational anharmonicity. This could be improved through the use of fourth-order vibrational perturbation theory, or better yet, a variational treatment combined with efficient surface-fitting techniques.[108] Additional improvements to the electronic structure treatment may be necessary to tighten the error bars even further for challenging systems. This could include Dirac-MRCI[109,110] for relativistic effects, Köppel-Domcke-Cederbaum vibronic coupling[111,112] for nonadiabatic effects, and internally contracted multireference coupled-cluster for static correlation.[90] These are avenues we hope to explore in future work.

2.9 Supporting Information

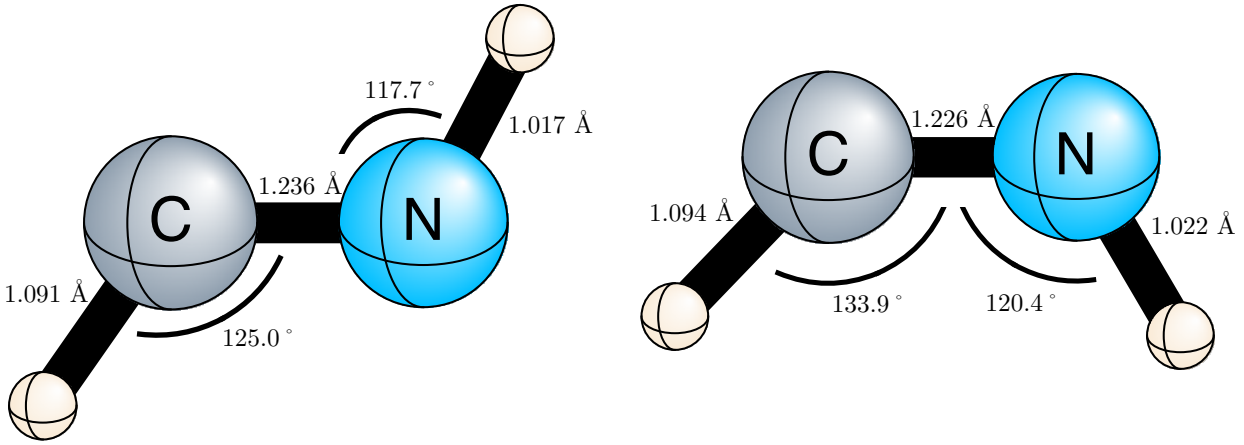

Figure 2.4: Planar equilibrium molecular structures for the ground electronic states of *trans*-HCNH and *cis*-HCNH, both optimized at the AE-CCSD(T)/cc-pCVQZ level of theory.

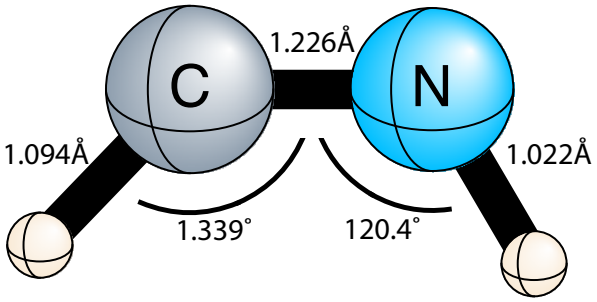

Figure 2.5: Planar equilibrium molecular structure for the ground electronic state of *cis*-HCNH, optimized at the AE-CCSD(T)/cc-pCVQZ level of theory.

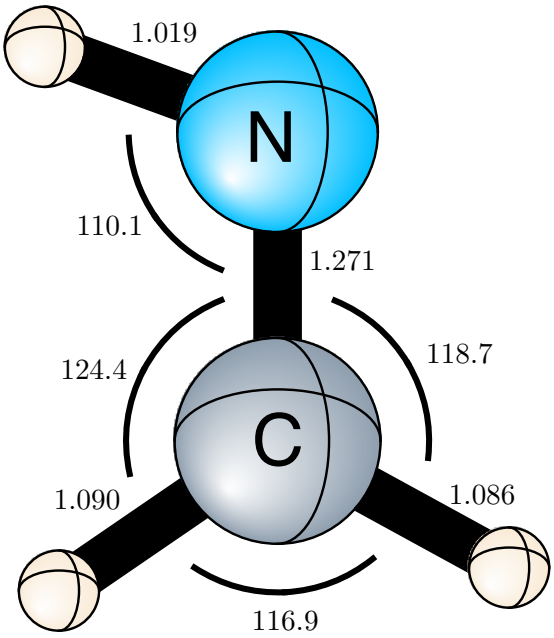

Figure 2.6: Planar equilibrium molecular structure for the ground electronic state of H_2CNH, optimized at the AE-CCSD(T)/cc-pCVQZ level of theory.

Table 2.5: Fundamental vibrational frequency table for the ground state of *trans*-HCNH. Harmonic frequencies (ω) and VPT2 anharmonic corrections were computed using AE-CCSD(T)/cc-pCVQZ. Fundamentals are given by $\nu = \omega + \delta\nu$. Harmonic intensities are reported in km mol^{-1}.

Mode	Sym	ω	$\delta\nu$	ν	Int.
1	a'	3471.8	-194.5	3277.3	10.0
2	a'	3064.6	-162.6	2902.0	19.5
3	a'	1759.4	-28.8	1730.6	20.6
4	a'	1202.1	-40.2	1161.9	11.5
5	a'	916.0	-24.8	891.1	218.7
6	a"	981.2	-18.5	962.7	99.4

Table 2.6: Fundamental vibrational frequency table for the ground state of *cis*-HCNH. Harmonic frequencies (ω) and VPT2 anharmonic corrections ($\delta\nu$) were computed using AE-CCSD(T)/cc-pCVQZ. Fundamentals are given by $\nu = \omega + \delta\nu$. Harmonic intensities are reported in km mol^{-1}.

Mode	Sym	ω	$\delta\nu$	ν	Intensity
1	a'	3359.4	3135.8	-223.5	2.1
2	a'	3017.7	2844.7	-173.1	32.8
3	a'	1809.8	1793.4	-16.4[a]	20.9
4	a'	1031.2	991.0	-40.3	132.6
5	a'	861.2	843.6	-17.6	114.1
6	a"	903.8	890.1	-13.6	3.9

[a] Treated for anharmonic resonance with the overtone of ν_6.

Table 2.7: Fundamental vibrational frequency table for the ground electronic state of H2CNH. Harmonic frequencies (ω) and VPT2 anharmonic corrections ($\delta\nu$) were computed using AE-CCSD(T)/cc-pCVQZ. Fundamentals are given by $\nu = \omega + \delta\nu$. Harmonic intensities are reported in km mol^{-1}.

Mode	Sym	ω	$\delta\nu$	ν	Int.
1	a'	3458.2	-175.0	3283.2	0.7
2	a'	3162.1	-145.0	3017.2[a]	24.3
3	a'	3059.7	-156.0	2903.8[b]	37.0
4	a'	1683.6	-39.2	1644.4	11.4
5	a'	1486.2	-31.0	1455.3	5.7
6	a'	1384.7	-36.5	1348.2	39.7
7	a'	1075.9	-15.3	1060.6	35.3
8	a"	1162.5	-30.4	1132.1	43.6
9	a"	1084.5	-19.5	1065.0	15.1

[a] Treated for resonance with the combination band of ν_4 and ν_5.
[b] Treated for resonance with the combination band of ν_4 and ν_6.

Table 2.8: Anharmonicity constants (χ_{ij}) for the ground electronic states of H_2CN, *trans*-HCNH, and *cis*-HCNH, computed with VPT2 at the AE-CCSD(T)/cc-pCVQZ level of theory.

i	j	χij (cm^{-1})		
		trans-HCNH	*cis*-HCNH	H_2CN
6	6	-3.70	-10.25	-0.32
6	5	12.53	-0.20	-10.03
6	4	-3.46	8.72	5.86
6	3	-5.06	29.46	-8.29
6	2	-16.34	-16.45	-14.95
6	1	-9.91	-7.80	-14.02
5	5	-3.47	-1.73	-40.31
5	4	-27.15	-7.95	-20.04
5	3	-4.98	-3.60	-25.68
5	2	-10.99	-12.36	5.14
5	1	-5.14	-4.15	-144.09
4	4	-6.75	-14.20	-1.03
4	3	-7.88	-14.90	-2.87
4	2	-4.57	-4.44	-6.51
4	1	-10.42	-5.17	-10.98
3	3	-10.35	-11.75	-3.79
3	2	7.52	9.81	2.86
3	1	-5.86	-6.64	-27.36
2	2	-76.63	-83.89	-9.81
2	1	5.73	12.81	-11.40
1	1	-90.86	-109.02	-32.10

Table 2.9: Anharmonicity constants (χ_{ij}) for the ground electronic states of H_2CNH, computed with VPT2 at the AE-CCSD(T)/cc-pCVQZ level of theory.

i	j	χij (cm^{-1})	i	j	χij (cm^{-1})
9	9	-1.1167	7	1	-2.0304
9	8	-4.4715	6	6	-6.2557
9	7	3.9326	6	5	-3.6376
9	6	2.7866	6	4	10.1371
9	5	-2.0469	6	3	-36.0385
9	4	-7.5830	6	2	3.6734
9	3	-9.7173	6	1	-6.7582
9	2	-14.7232	5	5	-4.1042
9	1	-2.7936	5	4	13.1518
8	8	-4.8501	5	3	-12.3976
8	7	1.3843	5	2	-33.7414
8	6	-1.7876	5	1	-2.4987
8	5	-1.7824	4	4	-5.6131
8	4	-5.1672	4	3	-45.3047
8	3	-6.1022	4	2	-8.0782
8	2	-7.5608	4	1	-1.0174
8	1	-15.8775	3	3	-41.6409
7	7	1.9458	3	2	-80.7605
7	6	-16.3387	3	1	2.5677
7	5	-2.5806	2	2	-43.8433
7	4	-12.1141	2	1	-1.4146
7	3	-6.0776	1	1	-80.0210
7	2	-4.5709			

CHAPTER 3

MULTI-FIDELITY GAUSSIAN PROCESS REGRESSION FOR CHEMICAL ENERGY SURFACES

[1]

[2] A. E. Wiens, A. V. Copan, and H. F. Schaefer. Accepted by Chemical Physics Letters. Reprinted here with permission of publisher.

3.1 Abstract

Modeling high-accuracy energy surfaces remains a challenge, due in large part to the cost of *ab initio* energy computations. Recently, there has been growing interest in Gaussian process (GP) regression, a machine learning technique that yields remarkably accurate fits with small training sets. We discuss an extension of GP modeling known as autoregressive Gaussian process (ARGP) modeling which has been shown by Perdikaris *et al.* [Proc. R. Soc. A **473** (2017)] to improve learning efficiency in fluid dynamics simulations but has never been applied to chemical energy surfaces. ARGP regression makes use of a cost-effective approximation to the target function to learn its structure before fitting the training data. This approach seems ideal for high-accuracy quantum chemistry, where the next-best approximation is often much cheaper than the desired level of theory. Our sample calculations demonstrate that ARGP modeling improves the prediction error of a five-point GP regression by two orders of magnitude for an N_2 dissociation curve. For the potential energy surface of H_2O, ARGP regression approaches quantitative accuracy with just 25 training points on the target surface, whereas ordinary GP regression requires at least 75 training points to achieve similar accuracy.

3.2 Introduction

Reliable modeling of chemical phenomena often requires the availability of an accurate energy surface. Since computing points on a quantum mechanical energy surface is expensive, it is typically cost-prohibitive to obtain these point values exclusively from *ab initio* data. Instead, some points must be inferred using a surrogate model, *i.e.* a method for predicting the outcome of an expensive process based on a limited set of sample calculations. Historically, surrogate models in quantum chemistry have relied on either interpolation[16–18] or least-squares fitting techniques,[13–15] but in the past decade there has been increasing interest in surrogate models based on modern machine learning. Much of this research has centered around feedforward neural networks,[19,20,22,23] which have proven useful in the technology sector. Neural network methods, however, are best suited to the "big data" regime, as they require a large number of training points to achieve an accurate fit.

Gaussian process (GP) regression is an alternative machine learning technique which is optimal for small- and medium-sized data sets. Having been widely used for decades in statistics and geostatistics applications, GP regression began to receive more widespread attention in the early 2000s for other applications in engineering.[113] In the recent years, GP methods have been increasingly used for quantum chemical applications,[114–123] including energy surface regression. A 2018 benchmark study shows that GP regression

achieves accurate fits with substantially smaller training sets than neural networks for modeling an accurate potential energy surface of formaldehyde (CH_2O).[29]

Another quickly developing research area in the machine learning community is *transfer learning*, which can also enhance learning efficiency when training data is expensive.[124] Transfer learning uses knowledge gained from a relevant auxiliary task to facilitate the learning of a target task. The most common approach to transfer learning is multi-task learning, which trains multiple related tasks in parallel, using extra tasks as an inductive bias to inform predictions of the target. Multi-fidelity modeling is a subset of multi-task transfer learning that exploits correlations between low- and high-accuracy data to enhance learning efficiency. This concept has also recently received some attention in the physical chemistry community.[125–127] In particular, Cui and Krems recently made use of a multi-fidelity technique to transfer knowledge between classical and quantum computations in a collision dynamics study.[128]

In the present work, we seek to draw attention to a multi-fidelity GP technique known as autoregressive Gaussian process (ARGP) modeling, which uses transfer learning to achieve further improvements over ordinary GP regression. This method is not original to us,[129–131] but to our knowledge it has never been applied to the modeling of high-accuracy chemical energy surfaces. We point out that it seems particularly well-suited to electronic structure theory, where energy points are expensive but a large variety of approximations is available. ARGP regression was recently developed in a general, nonlinear form by Perdikaris *et al.* to study mixed convection flows in fluid dynamics.[131] We demonstrate the utility of this method by presenting sample calculations for an N_2 dissociation curve, a potential energy surface of H_2O, and a near-equilibrium potential energy surface of CH_2O. Our preliminary benchmarks suggest that it can yield substantial gains in learning efficiency.

3.3 Multi-Fidelity Gaussian Process Regression

A Gaussian process (GP) is a continuous collection of random variables, any finite number of which have a joint Gaussian distribution. More formally, a GP, which we will denote by Y, is a distribution over functions whose values on an arbitrary point $\mathrm{X} \equiv \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ have a multivariate normal probability density of the form

$$\phi_{Y^{\mathrm{X}}}(y^{\mathrm{X}}) = \det(2\pi\Sigma^{\mathrm{X,X}})^{-\frac{1}{2}} \cdot \exp\left(-\tfrac{1}{2}(\Delta y^{\mathrm{X}})^{\dagger}(\Sigma^{\mathrm{X,X}})^{-1}\Delta y^{\mathrm{X}}\right) \tag{3.1}$$

$$\mu^{\mathrm{X}} \equiv \begin{pmatrix} \mu(\mathbf{x}_1) & \cdots & \mu(\mathbf{x}_n) \end{pmatrix}^{\mathrm{T}} \qquad \Sigma^{\mathrm{X,X}} \equiv \begin{pmatrix} \Sigma(\mathbf{x}_1, \mathbf{x}_1) & \cdots & \Sigma(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & \ddots & \vdots \\ \Sigma(\mathbf{x}_n, \mathbf{x}_1) & \cdots & \Sigma(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} \tag{3.2}$$

where $\Delta y^{\mathrm{X}} = y^{\mathrm{X}} - \mu^{\mathrm{X}}$, and $\mu(\mathbf{x})$ and $\Sigma(\mathbf{x}, \mathbf{x}')$ are called the mean and covariance functions, since they define the GP's mean vector and covariance matrix on any given set of points.

Gaussian process regression determines its predictions from a conditional GP, which contains the subset of functions in Y passing through the observed training points.

$$\tilde{Y}^{\mathrm{X}} \equiv \left(Y^{\mathrm{X}} \,\middle|\, Y^{\mathrm{T}} = \hat{y}^{\mathrm{T}}\right) \tag{3.3}$$

Here $\hat{y}$ denotes the target function, and $\hat{y}^{\mathrm{T}}$ are its known values on the training set

$$\mathrm{T} = \{\mathbf{t}_1, ..., \mathbf{t}_N\}\,. \tag{3.4}$$

In the language of Bayesian inference, Y is the prior distribution and $\tilde{Y}$ is the posterior distribution which reflects our knowledge of the training data. Its probability density is given by the following.

$$\phi_{\tilde{Y}^{\mathrm{X}}}(y^{\mathrm{X}}) \equiv \phi_{Y^{\mathrm{X}}|Y^{\mathrm{T}}}(y^{\mathrm{X}}|\hat{y}^{\mathrm{T}}) = \frac{\phi_{Y^{\mathrm{X}},Y^{\mathrm{T}}}(y^{\mathrm{X}}, \hat{y}^{\mathrm{T}})}{\phi_{Y^{\mathrm{T}}}(\hat{y}^{\mathrm{T}})} \tag{3.5}$$

The joint distribution of Y^{X} and Y^{T} is simply the GP prior evaluated over the set of all points in X and T. The mean values of the posterior distribution predict the values of the target function on these points, and the diagonal elements of its covariance matrix are variances describing the uncertainty at each point. These quantities are given by the following. [113]

$$\tilde{\mu}^{\mathrm{X}} = \mu^{\mathrm{X}} + \Sigma^{\mathrm{X,T}}(\Sigma^{\mathrm{T,T}})^{-1}(\hat{y}^{\mathrm{T}} - \mu^{\mathrm{T}}) \tag{3.6}$$

$$\tilde{\Sigma}^{\mathrm{X,X}} = \Sigma^{\mathrm{X,X}} - \Sigma^{\mathrm{X,T}}(\Sigma^{\mathrm{T,T}})^{-1}\Sigma^{\mathrm{T,X}} \tag{3.7}$$

The shape of the prior covariance function, $\Sigma(\mathbf{x}, \mathbf{x}')$, is controlled by a set of *hyperparameters*. A common covariance structure is the so-called radial basis function (RBF) kernel [132]

$$(\Sigma^{\mathrm{X,X}})_{ij} = \sigma^2 \exp\left(-\tfrac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^{\dagger} \boldsymbol{\lambda}^{-2} \left(\mathbf{x}_i - \mathbf{x}_j\right)\right) \tag{3.8}$$

whose hyperparameters are the output variance, σ, and the "lengthscales", $\boldsymbol{\lambda} \equiv \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. These are

determined to maximize the prior probability of the observed training data, a technique known as maximum likelihood estimation (MLE).[113] This procedure scales as $\mathcal{O}(N^3)$, a well-known limitation of GP regression that has been addressed elsewhere.[133,134]

Autoregressive Gaussian process (ARGP) modeling extends the procedure just described to learn an expensive-to-evaluate target function $\hat{y}_\nu$ alongside a series of auxiliary functions $\hat{y}_0$, $\hat{y}_1$,..., $\hat{y}_{\nu-1}$, which approximate the target with increasing accuracy. Each level of theory is expressed as a transformation of the previous (lower) level of theory plus a correction term, which is modeled as a GP.

One simple choice for this transformation is multiplication by a scalar, in which case the prior distribution for the t^{th} level of theory is formally described as[129]

$$Y_t(\mathrm{X}) = s_t Y_{t-1}(\mathrm{X}) + Z_t(\mathrm{X}) \tag{3.9}$$

where the Z_t is the Gaussian process, and scaling parameters $\{s_1, ..., s_t\}$ are found by maximum likelihood estimation along with the covariance hyperparameters. Unfortunately, direct Bayesian inference on this model is quite expensive, because the different approximation levels are copuled. In particular, predicting the mean of the posterior distribution requires the inversion of a $\Sigma_t|\mathrm{T}_t| \times \Sigma_t|\mathrm{T}_t|$ covariance matrix, where T_t is the training set for y_t and $|\mathrm{T}_t|$ is its cardinality. This cost problem in linear ARGP modeling was solved in 2014 by Le Gratiet and Garnier,[130] who showed that replacing the prior Gaussian process on the right-hand side of Equation 3.9 with its posterior GP $\tilde{Y}_{t-1}$ yields a decoupled scheme.

$$Y_t(\mathrm{X}) = s_t \tilde{Y}_{t-1}(\mathrm{X}) + Z_t(\mathrm{X}) \tag{3.10}$$

These modified linear ARGP regression equations yield equivalent predictions for the target y_t, and a given level t can be solved with successive $|\mathrm{T}_t| \times |\mathrm{T}_t|$ covariance matrix inversions. The base-level model $Y_0 = Z_0$ is solved by ordinary GP regression. The higher-level predictions are derived by applying the ordinary GP regression to Z_t and backsolving Equations 3.6 and 3.7 for the posterior mean and variances of of $\tilde{Y}_t$:

$$\tilde{y}_t(\mathrm{X}) = s_t\, \tilde{y}_{t-1}(\mathrm{X}) + \Sigma_t(\mathrm{X}, \mathrm{T}_t)\Sigma_t^{-1}(\mathrm{T}_t, \mathrm{T}_t)\ \cdot (y_t(\mathrm{T}_t) - s_t \tilde{y}_{t-1}(\mathrm{T}_t)) \tag{3.11}$$

$$\tilde{\sigma}_t^2(\mathrm{X}) = s_t^2\, \tilde{\sigma}_{t-1}^2(\mathrm{X}) + \Sigma_t(\mathrm{X}, \mathrm{X})\ - \Sigma_t(\mathrm{X}, \mathrm{T}_t)\Sigma_t^{-1}(\mathrm{T}_t, \mathrm{T}_t)\Sigma_t(\mathrm{T}_t, \mathrm{X})\ \cdot \tag{3.12}$$

In general, the training sets needed to achieve a given accuracy in the fit decrease in size from T_0 to T_t. When the training data for the lower levels of theory are much less expensive, this scheme can significantly reduce the cost of obtaining training data for an expensive system. This linear ARGP model is well-established in

the engineering literature and is an effective choice when the relationship between an approximate model and its target is nearly linear.[135–138]

A more general ARGP regression scheme was recently put forth by Perdikaris and coworkers.[131] In this case, the transformation of the previous level of theory is itself modeled as a GP over functions of $\mathbf{x}$ and $y_{t-1}^{\mathbf{x}}$, a placeholder variable for the value of the posterior distribution of the lower-level theory. Formally, we write

$$Y_t^{\mathrm{X}} = S_t^{\mathrm{X};\tilde{Y}_{t-1}^{\mathrm{X}}} + Z_t^{\mathrm{X}} \tag{3.13}$$

where S_t and Z_t combine to form a single zero-mean GP, which is given the following covariance structure.

$$\Sigma_t^{\mathrm{X,X}} \equiv \Sigma_{\mathrm{s},t}^{\mathrm{X,X}} \odot \Lambda_{\mathrm{y},t}^{\mathrm{X,X}} + \Sigma_{\mathrm{z},t}^{\mathrm{X,X}} \qquad\qquad \Lambda_{\mathrm{y},t}(\mathbf{x},\mathbf{x}') \equiv \Sigma_{\mathrm{y},t}(y_{t-1}^{\mathbf{x}}, y_{t-1}^{\mathbf{x}'}) \tag{3.14}$$

The symbol $\odot$ denotes element-wise multiplication. Equations 3.1 and 3.5 then yield a posterior density which depends parametrically on the placeholder variable. Following the usual procedure in Bayesian regression, this extra parameter is finally "marginalized out" with respect to the posterior distribution of the lower-level theory in order to determine the next posterior distribution.

$$\phi_{\tilde{Y}_t^{\mathbf{x}}}(y_t^{\mathbf{x}}) = \int dy_{t-1}^{\mathbf{x}}\, \phi_{\tilde{Y}_t^{\mathbf{x}}}(y_t^{\mathbf{x}}; y_{t-1}^{\mathbf{x}}) \cdot \phi_{\tilde{Y}_{t-1}^{\mathbf{x}}}(y_{t-1}^{\mathbf{x}}) \tag{3.15}$$

This approach has the advantage of fully propagating the uncertainty through to the final level of theory. Hyperparameters for the covariance functions in equation 3.14 are determined by MLE, with $y_{t-1}^{\mathrm{T}_t}$ set to the posterior mean, $\tilde{\mu}_{t-1}^{\mathrm{T}_t}$. This step also scales as $\mathcal{O}(N_t^3)$, but we note that as the fidelity of the data is increased, N_t will decrease. Because this step still corresponds to an ordinary GP regression problem, any scalable procedure for training GPs, such as the aforementioned References 133 and 134 can be applied. For a more detailed discussion of these regression schemes see Ref 131.

3.4 Generation of Surfaces and Training Sets

Reference values for the dissociation curve of N_2 were computed at one hundred evenly spaced points in the space of $r_{\mathrm{NN}} \in \{0.8, 2.35\}$Å. The bond lengths ($r$) were represented as Morse variables (e^{-r}) for each GP regression. Energies were computed using Davidson-corrected multireference configuration interaction theory, including single and double exictations (MRCISD+Q) with the aug-cc-pCV5Z basis set.[89,139,140] An approximate dissociation curve was computed with MRCISD+Q/cc-pCVTZ. A cheaper approximation was computed with CASSCF and the 6-31G* basis set.[141–144] Active spaces included 6 electrons in 6 orbitals

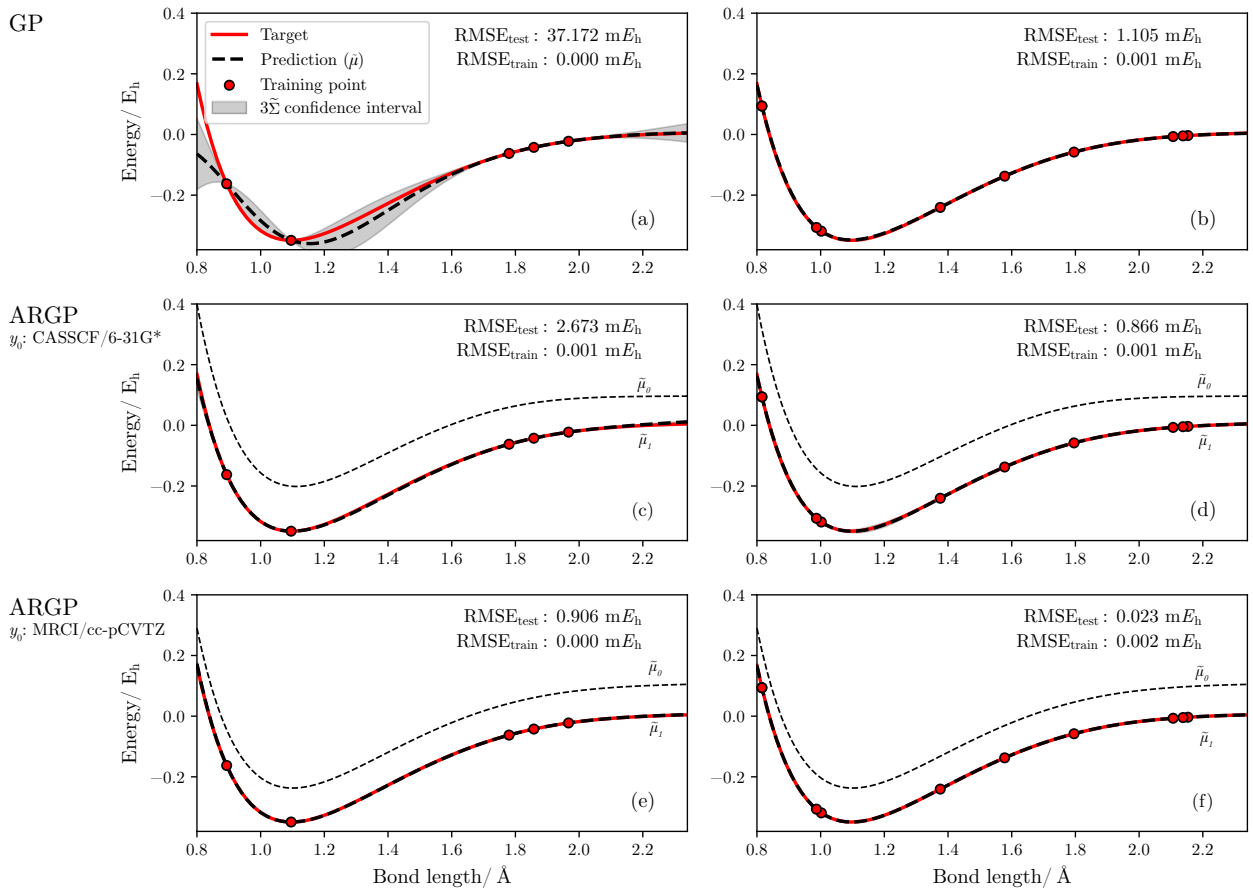

Figure 3.1: Comparison of GP and nonlinear ARGP models for N_2 dissociation at the Davidson-corrected multireference configuration interaction singles and doubles (MRCISD+Q) level of theory with the aug-cc-pCV5Z basis set.[89,139,140] Plots (c) and (d) employed complete active space self-consistent field theory[141,142] (CASSCF) with the 6-31G* basis set[143,144] as the base model for ARGP regression. Plots (e) and (f) employed MRCISD+Q/cc-pCVTZ as a base model. The ARGP base models were trained using ordinary GP regression on 30 points from the approximate surface; the energy curves illustrating the base models have been shifted upwards for clarity. Prediction errors for are reported as root-mean-square errors (RMSEs) from the training and test sets, *i.e.* all the points on the curve not used for training.

spanning the σ and π space of the triple bond. All of these computations were performed in Molpro[66,67] with all electrons correlated. Training sets were selected by uniform random sampling.

Reference values for the *ab inito* surface of H_2O were computed on the coordinate ranges $r_{\text{OH}} \in \{0.75, 1.25\}$ Å and $a_{\text{HOH}} \in \{80°, 160°\}$, at a uniform grid of 15 points, resulting in a total of 3375 points on the PES. All H_2O energies were computed using coupled cluster theory with single, double, and connected triple excitations[10,65] [CCSD(T)] in Molpro[66,67] with core electrons frozen. The base model surface was computed with a cc-pVDZ basis set, and the target surface was computed with cc-pV5Z.[91] Training sets were chosen from five-hundred uniform random samples by using a χ^2 test to pick a training set whose energetic distribution best represented the test set distribution.

RBF kernels with dimension-specific lengthscales (see equation 3.8) were used for the prior covariances. Their hyperparameters were determined by MLE using the Python-language Gaussian process library GPy.[145] Noise in the data was treated as Gaussian-distributed: $\varepsilon \approx \mathcal{N}(0, \alpha)$, where α is the Gaussian noise variance. This term was also optimized by maximum likelihood estimation in GPy and accounted for in the diagonal elements of the covariance matrix over the training set as follows.

$$\left(\Sigma^{\mathrm{T,T}}\right)_{ii} = \Sigma\left(\mathbf{t}_i, \mathbf{t}_i; \sigma^2, \lambda\right) + \alpha \tag{3.16}$$

The nonlinear ARGP formula in Equation 3.15 was solved by Monte Carlo integration in Python. We note that this requires sampling from the posterior distribution of the base model ($\tilde{Y}_0$), and propagating each output as an input to the next recursive level. This step scales linearly with the data, and all operations can be vectorized across multiple test points. The predictions were evaluated by their root-mean-square errors from the test set, *i.e.* the set of unseen points on the target surface used to measure the model's performance. For both applications considered here, the test set is the entire set of points not used for training on the *ab initio* surface.

3.5 Results

Figure 3.1 compares three Gaussian process (GP) regression schemes using the MRCISD+Q/aug-cc-pCV5Z energy surface of N_2 as a benchmark. The results are plotted and compared quantitatively via their root-mean-square errors (RMSEs) relative to the test set. RMSEs relative to the training set, introduced by noise in the model, are also reported.

Table 3.1: Comparison of ordinary GP and nonlinear ARGP prediction errors for the CCSD(T)/cc-pV5Z[10,65,91] potential energy surface of H_2O, with different training set sizes (N = 25, 50, 75, 100). For the ARGP regressions, the CCSD(T)/cc-pVDZ surface served as the base model, which was predicted using ordinary GP regression with N_0 = 150, 200, 250, and 300 training points. Prediction errors are listed in mE_h as root-mean-square errors from the training set and test set, *i.e.* all the *ab initio* points not used for training.

	$\mathrm{RMSE}_{\mathrm{test}}$ (mE_h)		$\mathrm{RMSE}_{\mathrm{train}}$ (mE_h)	
N	GP	ARGP	GP	ARGP
25	28.26	0.86	12.92	0.00
50	1.86	0.31	0.01	0.00
75	0.31	0.13	0.03	0.00
100	0.20	0.03	0.02	0.00

Plots (a) and (b) show ordinary GP regressions with five and ten training points, respectively. With five training points, the usual GP approach fails to capture the correct bonding curve within a 3σ (99.7%) confidence interval. In this case, the training data does not span the range of energy values on the curve. The MLE procedure therefore yields a misinformed prior distribution, resulting in an underestimated confidence interval and a large root-mean-square error of 37.2 mE_h from the target surface. With ten training points, ordinary GP regression yields a qualitatively accurate result with an error of 1.1 mE_h.

Plots (c) and (d) show the results of ARGP regression predicting the MRCISD+Q/aug-cc-pCV5Z target surface with CASSCF/6-31G* as the base model. With five training points, ARGP regression correctly captures the target surface within a 3σ confidence interval and achieves chemical accuracy with a prediction RMSE of 2.7 mE_h. It is remarkable that such a cheap and highly approximate level of theory as CASSCF with a small basis set is able to improve the prediction of the target by an order of magnitude relative to ordinary GP, and at almost neglible additional cost. With ten training points, however, the same ARGP method yields an 0.87 mE_h prediction error, only a marginal improvement over ordinary GP. For reference, more than one thousand CASSCF/STO-3G single points could have been computed in the time required to execute one target single point computation.

Plot (e) and (f) show that, when the more accurate MRCISD+Q/cc-pCVTZ theory is used as a base model, ARGP regression achieves sub-chemical accuracy in the prediction for both five and ten training points (0.91 and 0.02 mE_h RMSE, respectively). In both cases, this is an improvement of two orders of magnitude in the prediction error relative to ordinary GP. Although this is a more expensive base model, still one hundred MRCI/cc-pCVTZ single points could have been computed for every target point, and transferring knowledge from this surface yields significant gains in accuracy.

In summary, the training set of five points demonstrates that, where ordinary GP regression fails qualitatively, ARGP regression can yield chemical accuracy with a crude base model and sub-chemical accuracy with the intermediate base model of MRCISD+Q/cc-pCVTZ. The ten-point training set shows that, where ordinary GP regression succeeds qualitatively, ARGP regression can improve the prediction by up to two orders of magnitude using the MRCISD+Q/cc-pCVTZ base model.

Table 3.1 compares ordinary GP and nonlinear ARGP regression using the CCSD(T)/cc-pV5Z energy surface of H_2O as a benchmark. With 25 training points, ordinary GP regression yields a prediction error of 28.26 mE_h. ARGP regression outperforms this result by two orders of magnitude, achieving sub-chemical accuracy and a prediction error of 0.86 mE_h. With 50 training points, the ordinary GP prediction improves drastically to a root-mean-square error of 1.86 mE_h. However, ARGP regression still achieves an order-of-magnitude better (0.31 mE_h). With 75 training points, ordinary GP regression achives sub-chemical accuracy (0.31 mE_h RMSE), and ARGP modeling begins to converge to its limit of accuracy, which for this

application appears to be < 0.1 mE_h. With 100 training points, ordinary GP regression improves slightly, yielding prediction RMSE of (0.20 mE_h RMSE), and ARGP achieves an impressive prediction error of just 0.03 mE_h. We note that for both this and the N_2 application, ARGP modeling seems to have the desired property of offering the greatest gains in accuracy for small training sets.

Table 3.2: Comparison of ordinary GP and nonlinear ARGP prediction errors for the CCSD(T)/aug-cc-pV5Z[10,65,91] potential energy surface of CH_2O, with different training set sizes (N = 30, 50, 70, 90). For the ARGP regressions, the CCSD(T)/cc-pVDZ surface served as the base model, which was predicted using ordinary GP regression with N_0 = 100, 120, 140, and 160 training points. Prediction errors are listed in mE_h as root-mean-square errors from the training set and test set, *i.e.* all the *ab initio* points not used for training.

	$\mathrm{RMSE}_\mathrm{test}$ (mE_h)		$\mathrm{RMSE}_\mathrm{train}$ (mE_h)	
N	GP	ARGP	GP	ARGP
30	5.90	0.13	2.27	0.00
50	0.37	0.08	0.00	0.00
70	0.21	0.04	0.00	0.00
90	0.10	0.03	0.00	0.00

3.6 Conclusions

The nonlinear ARGP regression scheme discussed here is able to improve the prediction error of the MRCISD+Q/aug-cc-pCV5Z dissociation curve of N_2 by more than a factor of 40 with five randomly selected training points. For the three-dimensional energy surface of H_2O, ARGP regression comfortably achieves sub-chemical accuracy with 25 training points on the target *ab initio* surface. These initial benchmarks suggest that using ARGP regression to leverage the relationships between different levels of theory may substantially improve learning efficiency in chemical energy surface regression. Furthermore, this approach is especially well-suited to high-accuracy quantum chemistry, where the cost differential between the desired level of theory and its nearest approximations can be very large. For example, the time it takes to solve the widely-used CCSD(T) approximation scales with the fourth power of basis set size for a given system, which in turn scales quadratically with basis set cardinality.

Additional benchmark studies of polyatomic chemical systems will be necessary to further quantify the improvements in learning efficiency that autoregressive GP regression might yield over ordinary GP regression. Optimal sampling algorithms for this method will also need to be considered. Finally, an effective multi-fidelity model will take advantage of convergent hierarchies in basis set and level of theory to determine the best series of approximations. These are avenues that we hope to explore in future research.

CHAPTER 4

CONCLUSION

4.1 Final conclusions

The nonlinear ARGP regression scheme discussed here is able to improve the prediction error of the MRCISD+Q/aug-cc-pCV5Z dissociation curve of N_2 by more than a factor of 40 with five randomly selected training points. For the three-dimensional energy surface of H_2O, ARGP regression comfortably achieves sub-chemical accuracy with 25 training points on the target *ab initio* surface. These initial benchmarks suggest that using ARGP regression to leverage the relationships between different levels of theory may substantially improve learning efficiency in chemical energy surface regression. Furthermore, this approach is especially well-suited to high-accuracy quantum chemistry, where the cost differential between the desired level of theory and its nearest approximations can be very large. For example, the time it takes to solve the widely-used CCSD(T) approximation scales with the fourth power of basis set size for a given system, which in turn scales quadratically with basis set cardinality.

Additional benchmark studies of polyatomic chemical systems will be necessary to further quantify the improvements in learning efficiency that autoregressive GP regression might yield over ordinary GP regression. Optimal sampling algorithms for this method will also need to be considered. Finally, an effective multi-fidelity model will take advantage of convergent hierarchies in basis set and level of theory to determine the best series of approximations. These are avenues that we hope to explore in future research.

Bibliography

[1] Mezey, P. G. *Potential Energy Hypersurfaces*; Elsevier Science Publishers B. V.: Amsterdam, The Netherlands, 1987.

[2] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover Publications, Inc., 1982.

[3] Born, M.; Oppenheimer, R. *Annalen Der Physik* **1927**, 30.

[4] Diercksen, G. H. F.; Sutcliffe, B. T.; Veillard, A. *Computational Techniques in Quantum Chemistry*; Reidel, 1975; p 1.

[5] Lee, T. J.; Fox, D. J.; Schaefer, H. F.; Pitzer, R. M. *J. Chem. Phys.* **1984**, *81*, 336.

[6] Yarkony, D. R. *J. Phys. Chem. A* **2001**, *105*, 6277.

[7] Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **1986**, *84*, 2212.

[8] Moller, C.; Plesset, M. S. *Phys. Rev. 46*, 618.

[9] Leininger, M. L.; Allen, W. D.; Schaefer, H. F.; Sherrill, C. D. *J. Chem. Phys.* **2000**, *112*, 9213.

[10] Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.

[11] Deegan, M. J. O.; Knowles, P. J. *Chem. Phys. Lett.* **1994**, *227*, 321.

[12] Stanton, J. F. *Chem. Phys. Lett.* **1997**, *281*, 130.

[13] Aguado, A.; Paniagua, M. *J. Chem. Phys.* **1992**, *96*, 1265.

[14] Brown, A.; Braams, B. J.; Christoffel, K.; Jin, Z.; Bowman, J. M. *J. Chem. Phys.* **2003**, *119*, 8790.

[15] Braams, B. J.; Bowman, J. M. *Int. Rev. Phys. Chem.* **2009**, *28*, 577.

[16] Schatz, G. C. *Rev. Mod. Phys.* **1989**, *61*, 669.

[17] Ischtwan, J.; Collins, M. A. *J. Chem. Phys.* **1994**, *100*, 8080.

[18] Hollebeek, T.; Ho, T.-S.; Rabitz, H. *Annu. Rev. Phys. Chem.* **1999**, *50*, 537.

[19] Behler, J.; Parrinello, M. *Phys. Rev. Lett.* **2007**, *98*, 146401.

[20] Behler, J. *Int. J. Quantum Chem.* **2015**, *115*, 1032.

[21] Handley, C. M.; Popelier, P. L. A. *J. Chem. Phys.* **2010**, *114*, 3371.

[22] Manzhos, S.; Dawes, R.; Carrington, T. *Int. J. Quantum Chem.* **2015**, *115*, 1012.

[23] Jiang, B.; Li, J.; Guo, H. *Int. Rev. Phys. Chem.* **2016**, *35*, 479.

[24] Rupp, M.; von Lilienfeld, A. O.; Burke, K. *J. Chem. Phys.* **108**, *148*, 241401.

[25] Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. *Nat. Commun.* **2017**, *8*, 872.

[26] Li, L.; Snyder, J. C.; Pelaschier, I. M.; Huang, J.; Niranjan, U. N.; Duncan, P.; Rupp, M.; Müller, K.-R.; Burke, K. *Int. J. Quantum Chem.* **2016**, *116*, 819.

[27] Li, L.; Baker, T. E.; White, S. R.; Burke, K. *Phys. Rev. B* **2016**, *94*, 245129.

[28] Snyder, J. C.; Rupp, M.; Hansen, K.; Blooston, L.; Müller, K.-R.; Burke, K. *J. Chem. Phys.* **2013**, *139*, 224104.

[29] Kamath, A.; Vargas-Hernández, R. A.; Krems, R. V.; Carrington, T.; Manzhos, S. *J. Chem. Phys.* **2018**, *148*, 241702.

[30] Herzberg, G. *Molecular spectra and molecular structure: Spectra of diatomic molecules*; D. Van Nostrand Company: Princeton, New Jersey, 1939; Vol. 1.

[31] Herzberg, G. *Molecular spectra and molecular structure: Infrared and raman spectra of polyatomic molecules*; D. Van Nostrand Company: Princeton, New Jersey, 1945; Vol. 2.

[32] Herzberg, G. *Molecular Spectra and Molecular Structure: Electronic Spectra of Polyatomic Molecules*; D. Van Nostrand Company: Princeton, New Jersey, 1966; Vol. 3.

[33] Herzberg, G. *Phys. Rev. Lett.* **1969**, *23*, 1081.

[34] Bender, C. F.; Schaefer, H. F. *J. Am. Chem. Soc.* **1970**, *92*, 4984.

[35] Schaefer, H. F. *Science* **1986**, *231*, 1100.

[36] Julienne, P. S.; Krauss, M. *J. Mol. Spectrosc.* **1975**, *56*, 270.

[37] Alexander, M. H.; Dagdigian, P. J.; Jacox, M. E.; Kolb, C. E.; Melius, C. F.; Rabitz, H.; Smooke, M. D.; Tsang, W. *Prog. Energ. Combust.* **1991**, *17*, 263.

[38] Morgan, C. U.; Beyer, R. A. *Combust. Flame* **1979**, *36*, 99.

[39] Adams, G. F.; Shaw, R. W. *Annu. Rev. Phys. Chem.* **1992**, *43*, 311.

[40] Chakraborty, D.; Muller, R. P.; Dasgupta, S.; Goddard, W. A. *J. Comput.-Aided Mater.* **2001**, *8*, 203.

[41] Ohishi, M.; McGonagle, D.; Irvine, W. M.; Yamamoto, S.; Saito, S. *Astrophys. J.* **1994**, *427*, L51 – L54.

[42] Kaye, J. A.; Strobel, D. F. *Icarus* **1983**, *54*, 417–433.

[43] Imanaka, H.; Smith, M. A. *P. Natl. Acad. Sci. USA* **2010**, *107*, 12423.

[44] Yamamoto, S.; Saito, S. *J. Chem. Phys.* **1992**, *96*, 4157.

[45] Dagdigian, P. J.; Anderson, W. R.; Sausa, R. C.; Miziolek, A. W. *J. Phys. Chem.* **1989**, *93*, 6059.

[46] Nizamov, B.; Dagdigian, P. J. *J. Phys. Chem. A* **2003**, *107*, 2256–2263.

[47] Ogilvie, J. F.; Horne, D. G. *J. Chem. Phys.* **1968**, *48*, 2248.

[48] Teslja, A.; Dagdigian, P. J.; Banck, M.; Eisfeld, W. *J. Phys. Chem. A* **2006**, *110*, 7826–7834.

[49] Bernard, E. J.; Strazisar, B. R.; Davis, H. F. *Chem. Phys. Lett.* **1999**, *313*, 461.

[50] McManus, H. J.; Fessenden, R. W.; Chipman, D. M. *J. Phys. Chem.* **1988**, *92*, 3781.

[51] Nesbitt, F. L.; Marston, G.; Stief, L. J.; Wickramaaratchi, M. A.; Tao, W.; Klemm, R. B. *J. Phys. Chem.* **1991**, *95*, 7613.

[52] Kamachi, M.; Kuwata, K.; Murahashi, S. *J. Phys. Chem.* **1971**, *75*, 164.

[53] Jacox, M. E. *J. Phys. Chem.* **1987**, *91*, 6595–6600.

[54] Cowles, D. C.; Travers, M. J.; Frueh, J. L.; Ellison, G. B. *J. Chem. Phys.* **1991**, *94*, 3517 – 3528.

[55] Petterson, M.; Lundell, J.; Khriachtchev, L.; Räsänen, M. *J. Chem. Phys.* **1998**, *109*, 618.

[56] Cochran, E. L.; Adrian, F. J.; Bowers, V. A. *J. Chem. Phys.* **1962**, *36*, 1938–1942.

[57] Brinkmann, N. R.; Wesolowski, S. S.; Schaefer, H. F. *J. Chem. Phys.* **2001**, *114*, 3055.

[58] Eisfeld, W. *J. Chem. Phys.* **2004**, *120*, 6056.

[59] Puzzarini, C.; Barone, V. *Chem. Phys. Lett.* **2009**, *467*, 276.

[60] Puzzarini, C. *Int. J. Quantum Chem.* **2010**, *110*, 2483.

[61] Barone, V.; Carbonniere, P.; Pouchan, C. *J. Chem. Phys.* **2005**, *122*, 224308.

[62] Jensen, F. *Chem. Phys. Lett.* **1990**, *169*, 519.

[63] Parkinson, C. J.; Mayer, P. M.; Radom, L. *Theor. Chem. Acc.* **1999**, *102*, 92.

[64] Szalay, P. G.; Gauss, J. *J. Chem. Phys.* **1997**, *107*, 9028.

[65] Bartlett, R. J.; Watts, J. D.; Kucharski, S. A.; Noga, J. *Chem. Phys. Lett.* **1990**, *165*, 513.

[66] Werner, H.-J. et al. MOLPRO, version 2015.1, a package of ab initio programs. 2015; see http://www.molpro.net.

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

[68] Kállay, M.; Surján, P. R. *J. Chem. Phys.* **2001**, *115*, 2945.

[69] MRCC, a quantum chemical program suite written by M. Kállay, Z. Rolik, J. Csontos, I. Ladjánszki, L. Szegedy, B. Ladóczki, G. Samu, K. Petrov, M. Farkas, P. Nagy, D. Mester, and B. Hgely. See also Z. Rolik, L. Szegedy, I. Ladjánszki, B. Ladóczki, and M. Kállay, J. Chem. Phys. 139, 094105 (2013), as well as: www.mrcc.hu.

[70] Kállay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 214105.

[71] Bomble, Y. J.; Stanton, J. F.; Kállay, M.; Gauss, J. *J. Chem. Phys.* **2005**, *123*, 054101.

[72] Kállay, M.; Gauss, J. *J. Chem. Phys.* **2008**, *129*, 144101.

[73] Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability, R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince III, E. G. Hohenstein, U. Bozkaya, A. Yu. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer III, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, and C. D. Sherrill, J. Chem. Theory Comput., 13(7) 3185.

[74] Cheng, L.; Gauss, J. *J. Chem. Phys.* **2011**, *135*, 084114.

[75] Handy, N. C.; Yamaguchi, Y.; Schaefer, H. F. *J. Chem. Phys.* **1986**, *84*, 4481.

[76] Sellers, H.; Pulay, P. *Chem. Phys. Lett.* **1984**, *103*, 463.

[77] CFOUR, a quantum chemical program package written by J.F. Stanton, J. Gauss, L. Cheng, M.E. Harding, D.A. Matthews, P.G. Szalay with contributions from A.A. Auer, R.J. Bartlett, U. Benedikt, C. Berger, D.E. Bernholdt, Y.J. Bomble, O. Christiansen, F. Engel, R. Faber, M. Heckert, O. Heun, C. Huber, T.-C. Jagau, D. Jonsson, J. Jusélius, K. Klein, W.J. Lauderdale, F. Lipparini, T. Metzroth, L.A. Mück, D.P. O'Neill, D.R. Price, E. Prochnow, C. Puzzarini, K. Ruud, F. Schiffmann, W. Schwalbach, C. Simmons, S. Stopkowicz, A. Tajti, J. Vázquez, F. Wang, J.D. Watts and the integral packages *MOLECULE* (J. Almlöf and P.R. Taylor), *PROPS* (P.R. Taylor), *ABACUS* (T. Helgaker, H.J. Aa. Jensen, P. Jørgensen, and J. Olsen), and ECP routines by A. V. Mitin and C. van Wüllen. For the current version, see http://www.cfour.de.

[78] Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1988**, *89*, 5803.

[79] Szalay, P. G.; Müller, T.; Gidofalvi, G.; Lischka, H.; Shepard, R. *Chem. Rev.* **2012**, *112*, 108.

[80] Clabo, D. A.; Allen, W. D.; Remington, R. B.; Yamaguchi, Y.; Schaefer, H. F. *Chem. Phys.* **1988**, *123*, 187.

[81] Nielsen, H. H. *Rev. Mod. Phys.* **1951**, *23*, 90.

[82] PyVPT2 is a vibrational anharmonicity program written in Python by J. Agarwal. Center for Computational Quantum Chemistry, University of Georgia, Athens, GA.

[83] East, A. L. L.; Allen, W. D. *J. Chem. Phys.* **1993**, *99*, 4638–4650.

[84] Császár, A. G.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **1998**, *108*, 9751.

[85] Schuurman, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2004**, *120*, 11586.

[86] Feller, D. *J. Chem. Phys.* **1992**, *96*, 6104.

[87] Feller, D. *J. Chem. Phys.* **1993**, *98*, 7059.

[88] Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.

[89] Langhoff, S. R.; Davidson, E. R. *Int. J. Quantum Chem.* **1974**, *8*, 61.

[90] Hanauer, M.; Köhn, A. *J. Chem. Phys.* **2012**, *136*, 204107.

[91] Dunning, T. J. *J. Chem. Phys.* **1989**, *90*, 1007.

[92] Allen, W. D.; Császár, A. G.; Horner, D. A. *J. Am. Chem. Soc.* **1992**, *114*, 6834.

[93] INTDER2005 is a general program developed by Wesley D. Allen and co-workers that performs various vibrational analyses and higher-order nonlinear transformations among force field representations.

[94] Vázquez, J. A.; Stanton, J. F. *J. Phys. Chem. A* **2002**, *106*, 4429.

[95] Konen, I. M.; Pollack, I. B.; Li, E. X. J.; Lester, M. I. *J. Chem. Phys.* **2005**, *122*, 094320.

[96] Copan, A. V.; Schaefer, H. F.; Agarwal, J. *Mol. Phys.* **2015**, *113*, 2992.

[97] Jacox, M. E. *Chem. Soc. Rev.* **2002**, *31*, 108.

[98] Nakanaga, T.; Kondo, S.; Saëki, S. *J. Chem. Phys.* **1982**, *76*, 3860.

[99] Johns, J. W. C.; Olson, W. B. *J. Mol. Spectrosc.* **1971**, *39*, 479.

[100] Bair, R. A.; Dunning, T. H. *J. Chem. Phys.* **1985**, *82*, 2280.

[101] Jacox, M. E.; Milligan, D. E. *J. Mol. Spectrosc.* **1975**, *56*, 333.

[102] Halonen, L.; Duxbury, G. *Chem. Phys. Lett.* **1985**, *118*, 246.

[103] Halonen, L.; Duxbury, G. *J. Chem. Phys.* **1985**, *83*, 2091.

[104] Hamada, Y.; Hashiguchi, K.; Tsuboi, M.; Koga, Y.; Kondo, S. *J. Mol. Spectrosc.* **1984**, *105*, 70.

[105] Luo, C.; Duan, X.; Liu, J.; Li, Z. *J. Phys. Chem. A* **2008**, *112*, 8979.

[106] Feldman, T.; Romanko, J.; Welsh, H. L. *Can. J. Phys.* **1956**, *34*, 737.

[107] Petek, H.; Nesbitt, D. J.; Darwin, D. C.; Ogilby, P. R.; Moore, C. B.; Ramsay, D. A. *J. Chem. Phys.* **1989**, *91*, 6566.

[108] Kolb, B.; Marshall, P.; Zhao, B.; Jiang, B.; Guo, H. *J. Phys. Chem. A* **2017**, *121*, 2552.

[109] Shiozaki, T.; Mizukami, W. *J. Chem. Theory Comput.* **2015**, *11*, 4733.

[110] BAGEL, Brilliantly Advanced General Electronic-structure Library. http://www.nubakery.org under the GNU General Public License.

[111] Domcke, W.; Köppel, H.; Cederbaum, L. S. *Mol. Phys.* **1981**, *43*, 851.

[112] Rabidoux, S. M.; Eijkhout, V.; Stanton, J. F. *J. Phys. Chem. A* **2014**, *118*, 12059.

[113] Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, Massachusetts, 2006.

[114] Cui, J.; Li, Z.; Krems, R. V. *J. Chem. Phys.* **2015**, *143*, 154101.

[115] Cui, J.; Krems, R. V. *J. Phys. B: At. Mol. Opt. Phys.* **2016**, *49*, 224001.

[116] Alborzpour, J. P.; Tew, D. P.; Habershon, S. *J. Chem. Phys.* **2016**, *145*, 174112.

[117] Kolb, B.; Marshall, P.; Zhao, B.; Jiang, B.; Guo, H. *J. Phys. Chem. A* **2017**, *121*, 2252.

[118] Uteva, E.; Graham, R. S.; Wilkinson, R. D.; Wheatley, R. J. *J. Chem. Phys.* **2017**, *147*, 161706.

[119] John, S. T.; Csányi, G. *J. Phys. Chem. B* **2017**, *121*, 10934.

[120] Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. *Science Advances* **2017**, *3*, e1701816.

[121] Qu, C.; Yu, Q.; Van Hoozen, B. L.; Bowman, J. M.; Vargas-Hernández, R. A. *J. Chem. Theory Comput.* **2018**, *14*, 3381.

[122] Denzel, A.; Kästner, J. *J. Chem. Phys.* **2018**, *148*, 094114.

[123] Schmitz, G.; Christiansen, O. *J. Chem. Phys.* **2018**, *148*, 241704.

[124] Pan, S. J.; Yang, Q. *IEEE T. Knowl. Data En.* **2010**, *22*, 1345–1359.

[125] Miller, R. L.; Harding, L. B.; Davis, M. J.; Gray, S. K. *J. Chem. Phys.* **2012**, *136*, 074102.

[126] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. *New J. Phys.* **2013**, *15*, 095003.

[127] Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. *ACS Cent. Sci.* **2019**, *5*, 57.

[128] Cui, J.; Krems, R. V. *Phys. Rev. Lett.* **2015**, *115*, 073202.

[129] Kennedy, M. C.; O'Hagan, A. *Biometrika* **2000**, *87*, 1.

[130] Le Gratiet, L.; Garnier, J. *Int. J. Uncertain. Quan.* **2014**, *4*, 365.

[131] Perdikaris, P.; Raissi, M.; Damianou, A.; Lawrence, N. D.; Karniadakis, G. E. *Proc. R. Soc. A* **2017**, *473*, 1.

[132] Vert, J. P.; Tsuda, K.; Schölkopf, B. *Kernel Methods in Computational Biology*; MIT Press: Cambridge, MA, USA, 2004; pp 35–70.

[133] Hensman, J.; Fusi, N.; Lawrence, N. D. *arXiv e-prints* **2013**, arXiv:1309.6835.

[134] Snelson, E.; Ghahramani, Z. Sparse Gaussian Processes Using Pseudo-inputs. Proceedings of the 18th International Conference on Neural Information Processing Systems. Cambridge, MA, USA, 2005; pp 1257–1264.

[135] Koziel, S.; Bekasiewicz, A.; Couckuyt, I.; Dhaene, T. *IEEE T. Antenn. Propag.* **2014**, *62*, 5900.

[136] Xu, T.; Valocchi, A. J.; Ye, M.; Liang, F. *Water Resour. Res.* **2017**, *53*, 4084.

[137] Thenon, A.; Gervais, V.; Ravalec, M. L. *Computat. Geosci.* **2016**, *20*, 1231.

[138] Babaee, H.; Perdikaris, P.; Chryssostomidis, C.; Karniadakis, G. E. *J. Fluid Mech.* **2016**, *809*, 895.

[139] Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1995**, *103*, 4572.

[140] Peterson, K. A.; Dunning, T. H. *J. Chem. Phys.* **2002**, *117*, 10548.

[141] Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 8052.

[142] Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259.

[143] Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213.

[144] Francl, M. M.; Pietro, W. J.; Hehre, W. J. *J. Chem. Phys.* **1982**, *77*, 3654.

[145] GPy, GPy: A Gaussian process framework in Python. http://github.com/SheffieldML/GPy, since 2012.