RSC Publishing Faraday Discussions

Teaching Vibrational Spectra to Assign Themselves

Journal:	*Faraday Discussions*
Manuscript ID	FD-ART-04-2018-000075.R1
Article Type:	Paper
Date Submitted by the Author:	03-May-2018

Complete List of Authors:	Houston, Paul; Cornell University; Georgia Institute of Technology, Chemistry and Biochemistry Van Hoozen, Jr., Brian; Emory University, Chemistry Chen, Qu; Emory University, Chemistry Yu, Qi; Emory University, Chemistry Bowman, Joel; Emory University, Chemistry

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Teaching Vibrational Spectra to Assign Themselves†

Paul L. Houston,*[a,b] Brian L. Van Hoozen, Jr.,[c] Chen Qu,[c] Qi Yu,[c] and Joel M. Bowman[c]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

A new paradigm for assigning vibrational spectra is described. Instead of proceeding from potential to Hamiltonian to eigenvalues/eigenvectors/intensities to spectrum, the new method goes "backwards" directly from spectrum to eigenvectors. The eigenvectors then "assign" the spectrum, in that they identify the basis states that contribute to each eigenvalue. To start, we demonstrate an algorithm that can obtain useful estimates of the eigenvectors connecting a real, symmetric Hamiltonian to its eigenvalues even if the only available information about the Hamiltonian is its diagonal elements. When this algorithm is augmented with information about transition intensities, it can be used to assign a complex vibrational spectrum using only information about 1) eigenvalues (the peak centers of the spectrum) and 2) a harmonic basis set (taken to be the diagonal elements of the Hamiltonian). Examples will be discussed, including application to the vibrationally complex spectral region of the formic acid dimer.

1 Introduction

How parts of a molecule move relative to one another is an essential window for understanding what holds them together, how they interact with their environment, and how they might dissociate or react. Infrared and Raman spectroscopies are powerful and ubiquitous tools that provide this window. However, interpreting the spectrum currently involves a route illustrated by the red pathway in Fig. 1*a*. Typically, a potential, V, is determined from electronic structure calculations. That potential is then used with an accurate, preferably exact, representation of the kinetic energy operator, to determine the Hamiltonian operator. Then, the Hamiltonian is typically represented by a matrix expressed in a complete basis of orthonormal functions, e.g., eigenfunctions of some zero-order Hamiltonian, often the harmonic one. Next, the Hamiltonian matrix is diagonalized to give eigenvalues, which give the positions of peaks in the spectrum, and eigenvectors, which together with the basis give the eigenfunctions. Finally, additional information from the dipole moment and standard expressions for oscillator strengths of the vibrations are used with the eigenvalues to obtain the calculated spectrum for comparison to the experimental one. When a good match is found, one assumes that the potential correctly describes how the molecule vibrates and the eigenfunctions/eigenvectors are used to "assign" the various spectral bands.

The standard process from potential, V, and dipole moment to the spectrum, indicated in Fig. 1*a*, is certainly a correct, but cumbersome one, which for large molecules can literally require person-years to complete. There are numerous strategies to obtain and represent V, for example a quartic expansion about a minimum; a recent quartic expansion for CO_2 [1] is a good example. Another method, embodied in the program MULTIMODE[2], is the n-mode representation of V. MULTIMODE uses vibrational self-consistent and configuration interaction methods to calculate the eigenfunctions for vibrational eigenstates in terms of a primitive harmonic basis set. Assuming that an accurate potential is available, this approach requires the diagonalization of very large matrices, even for moderately sized molecules and complexes, and is quite time consuming. Indeed, all of these methods scale at least as the square of the number of vibrational modes and are thus limited to approximations for molecules with more than about ten atoms. Another method is to use a "spectroscopic" Hamiltonian. This method assumes a functional expansion for the Hamiltonian elements and then fits the coefficients by matching the spectrum. The spectroscopic Hamiltonian method was first introduced to the chemistry community from nuclear physics[3] by Iachello, Levine and colleagues[4–9]. Subsequent work has focused on effective model Hamiltonians[10–22]. Although spectroscopic Hamiltonians have been successfully applied to simple anharmonic systems,[11,13,15,17–19] they have not yet been used to describe complex spectral bands such as that of the formic acid dimer described below.

In this report, we describe success in exploring a new paradigm for spectral assignment that traverses the reverse, blue, route

[a] Department of Chemistry and Chemical Biology, Cornell Universtiy, Ithaca, NY 14853, USA. Fax: 01 607 255 4137; Tel: 01 607 592 2713; E-mail: plh2@cornell.edu
[b] School of Chemistry and Biochemistry, Georgia Inst. of Technology, Atlanta, GA 30332-0400, USA.
[c] Department of Chemistry and Cherry L. Emerson Center for Scientific Computation, Emory University Atlanta, Georgia 30322, USA.

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/b000000x/

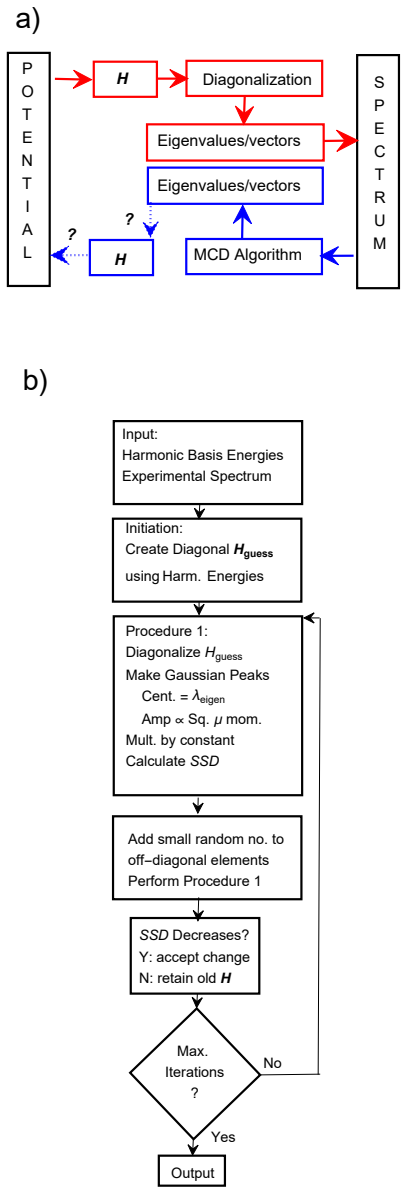

Fig. 1 *a)* Connections between the Potential Energy Function and the Spectrum. The red path is the traditional method of fitting a spectrum. The blue path, investigated here, provides an alternative way of assigning the spectrum with little recourse to the potential. *b)* MCD algorithm used for finding a Hamiltonian and its eigenvalues/vectors that is consistent with the spectrum and the harmonic basis energies.

in Fig. 1*a*. This route builds from a model description of the spectrum, specifically the text-book double-harmonic model. Although the double-harmonic model generally falls short of the level of accuracy needed to assign complex spectra, we show that it can be used as stepping stone to the goal of assigning complex spectra with no knowledge of the potential energy surface. We approach this goal by casting the problem as an Inverse Eigenvalue Problem,[23] to which we approximate the solution by using the underlying harmonic model and what we call a Monte Carlo Diagonalization (MCD) method, to be described below. We augment the method by making use of dipole derivatives (the second part of the double-harmonic approximation). We show that it is possible to fit a complex experimental spectral band extremely "quickly," with perhaps a day's effort. The assignment is then made using the eigenvectors determined from the spectrum. We believe that the spirit of the approach is general and could be applied to many problems in computational chemistry.

In what follows, we establish the credentials of the method with a brief description of the algorithm as used to determine eigenvectors for a two-mode representation of the CO_2 molecule. We then demonstrate the power of this new paradigm by applying it to the assignment of a complex spectral band of the formic acid dimer, $(HCOOH)_2$, (FAD) in the the 2500-3300 cm^{-1} spectral region. The FAD exhibits a C_{2h} structure containing two hydrogen bonds with tunneling between each over a D_{2h} barrier to give an equivalent C_{2h} structure.[24] It is a prototypical example of double hydrogen bond exchange in carboxylic acids. Early room-temperature spectra have been reported, but assignment has proven difficult due to the many hot bands.[25–29] More recent spectra of FAD in a cooled supersonic jet or in a gas matrix have been more informative.[30–38] Nydegger et al.[39] have recently reported a 2D IR investigation of the C-D stretching region of the deuterated analog. A recent room-temperature spectrum has been reported by Mackeprang et al.[40] Because of the large number of vibrational modes (24), most theoretical calculations have had to compromise either on the electronic structure method or on the method for calculating the Hamiltonian. Recent theoretical efforts have all found extensive mixing. These include Florio et al.,[41] who found coupling between the O-H stretch and the bending modes, Matanović and Došlić,[42] who found that the broadening of the spectrum involved Fermi resonances and coupling to low-frequency modes, Barnes and Sibert,[43] who found extensive state mixing, Pitsevich et al.,[44] who found agreement with some experimental frequencies, and Mackeprang et al.,[40] whose work reproduced the complexity of the spectrum but whose frequencies did not match the experimental ones. Despite the progress of these and other studies, the spectrum remains largely unassigned, and so it clearly is a major challenge for theory.

2 Methods

2.1 Diagonalization and Inverse Diagonalization

The goal of the numerical methods that follow is to generate from known eigenvalues and known Hamiltonian diagonal elements the eigenvectors that connect them. The H-matrix, as usual, is assumed to be real-symmetric. We generally assume that the diago-

nal elements of the Hamiltonian are the known basis energies. In one approach, the eigenvectors are obtained using a Monte Carlo search in which a trial H-matrix is diagonalized and, by a simple iterative procedure, convergence to the desired set of eigenvalues, Λ, is obtained. Once converged, the desired matrix of eigenvectors, **C**, where $\mathbf{C}\,\mathbf{H}\,\mathbf{C}^T = \Lambda$, can be obtained by standard techniques. Here, **C** is a matrix whose columns are the eigenvectors corresponding the the eigenvalues on the diagonal of Λ. This procedure, which does not yield unique results, is repeated a number of times to determine the dispersion of results. We term this method the Monte Carlo diagonalization algorithm (MCD).

Another stochastic approach randomly changes trial vectors, **C**, and from the equation $\mathbf{H}=\mathbf{C}^T\Lambda\mathbf{C}$, generates trial H-matrices. The search algorithm looks for a match between the calculated diagonals of **H** and the basis energies. We term this approach the Monte Carlo inverse diagonalization algorithm (MCID).

The details of the MCD and MCID algorithms are discussed in Section S1 of the ESI; a summary of the MCD mehtod will be provided below. We stress again that the goal in both cases is to obtain a small number of eigenvectors (relevent to, say, a large H-matrix) by seeking a connection between the eigenvalues of one or perhaps several much smaller H-matrices and the known, and thus constrained, Hamiltonian diagonal elements. We next comment on the relationship between the dimensionality of the problem and the number of unknown parameters.

2.2 Dimensionality and degree to which problem is underdetermined

At first glance, it doesn't seem likely that any method will be useful: if **H** is an $n_{dim} \times n_{dim}$ matrix, then the number of unique off-diagonal elements (unknowns) goes as $\frac{1}{2}n_{dim}(n_{dim}-1)$ whereas the number of diagonals (knowns) goes as n_{dim}. We could alternatively look at the number of eigenvector components that need to be determined, but we need to take into account the orthonormal constraints on the n_{dim}^2 eigenvector components. There are $\frac{1}{2}n_{dim}(n_{dim}-1)$ constraints from orthogonality and n_{dim} constraints from normalization. Simple algebra shows that the number of unconstrained eigenvector components is exactly equal to the number of unique off-diagonal elements of **H**, showing again that the approach of diagonalization of a guessed **H** and the inverse diagonalization using guessed eigenvector coefficients are equivalent problems. In either case there are $\frac{1}{2}n_{dim}(n_{dim}-1)$ unknowns. How many knowns are there? There are n_{dim} known basis energies and n_{dim} known eigenvalues, but these knowns are not independent: the sum of each group must be the same. We are thus left with $n_{dim}-1$ equations relating the knowns. The result is that the number of degrees of freedom goes as $\frac{1}{2}n_{dim}(n_{dim}-1)-(n_{dim}-1) = \frac{1}{2}(n_{dim}-2)(n_{dim}-1)$. The system will be under-determined for all $n_{dim} > 2$. Given this result, how can either algorithm for solution be useful?

An answer to this question is to focus on the how the equations relating the knowns constrain the range of unique off-diagonal elements (or, equivalently, the range of the eigenvector components). This information comes from the determinant equation.

2.3 The determinant equation

Only certain combinations of the unknown unique off-diagonal elements (or the unknown unconstrained eigenvector components) are allowed by the determinant equation, which connects the known eigenvalues and the known basis energies:

$$\sum_{i=0}^{n_{dim}} f_i \times (\lambda^{(j)})^i = 0, \qquad j = 1,\ldots n_{dim}. \tag{1}$$

Note that there are $(n_{dim}+1)$ coefficients f_i of the powers of $\lambda^{(j)}$ in the determinant equation; all but two of these involve the off-diagonal elements. In order for the determinant equation to have the appropriate eigenvalues, these $(n_{dim}+1)-2$ values of f_i must be invariant to any choices of the off-diagonal elements of the Hamiltonian. These resulting $n_{dim}-1$ equations are the constraints imposed by the known values referred to in the previous section.

A simple example is instructive. If $n_{dim} = 3$, there is only $\frac{1}{2}(n_{dim}-2)(n_{dim}-1) = 1$ degree of freedom, and there are $\frac{1}{2}n_{dim}(n_{dim}-1) = 3$ unique off-diagonal unknown elements. We thus expect the solutions for these unknowns to lie on a one-dimensional curve in the three-dimensional space of the unknown off-diagonal elements. Specifically, solution of the determinant for $n_{dim} = 3$ gives the constraints:

$$\begin{aligned} -E^{(1)}E^{(2)}E^{(3)} + E^{(3)}H_{12}^2 - 2H_{12}H_{13}H_{23} + E^{(1)}H_{23}^2 &= f_0, \\ E^{(1)}E^{(2)} + E^{(1)}E^{(3)} + E^{(2)}E^{(3)} - H_{12}^2 - H_{13}^2 - H_{23}^2 &= f_1, \end{aligned} \tag{2}$$

where $E^{(i)}$ are the three basis energies and H_{ij} are the three unique off-diagonal Hamiltonian elements.

The solution for the H_{ij} will be a family of points on a 1D curve that satisfies the total number of constraints. The curve for a particular example is shown in the top panel of Fig. 2. As shown in the middle panel of Fig. 2, the projections of this curve onto the H_{13} (red) or H_{23} (blue) vs H_{12} axes show the correlations that must hold in order for the determinant equation to be invariant. We see that for this example the range of H_{12} is limited approximately to 20-54 cm^{-1}, that of H_{13} to 0-70 cm^{-1}, and that of H_{23} to 0-50 cm^{-1}. Finally, we see from the bottom panel of Fig. 2 that the limit on H_{12} places a limit on component one of eigenvector one, as calculated using MCID; it must be between 0.90 and 0.945. Even though the solution provides only a range of answers, the finding that this eigenvector component is limited to somewhere near 0.92 provides the useful information that the first eigenvalue is composed primarily of the first basis function.

The situation becomes rapidly more complicated as n_{dim} increases. For $n_{dim} = 4$, there are 3 degrees of freedom and 6 unknown elements. The solution will be a family of points in a three dimensional volume located in the six-dimensional space of off-diagonal elements. In general, when $n_{dim} > 2$ there will be many solutions to the problem, but the range of possible off-diagonal element combinations, or the range of possible eigenvector components, will be limited by the coefficients f_i in the determinant equation. As found for the $n_{dim} = 3$ example, the limit on the possible eigenvalue components, particularly those

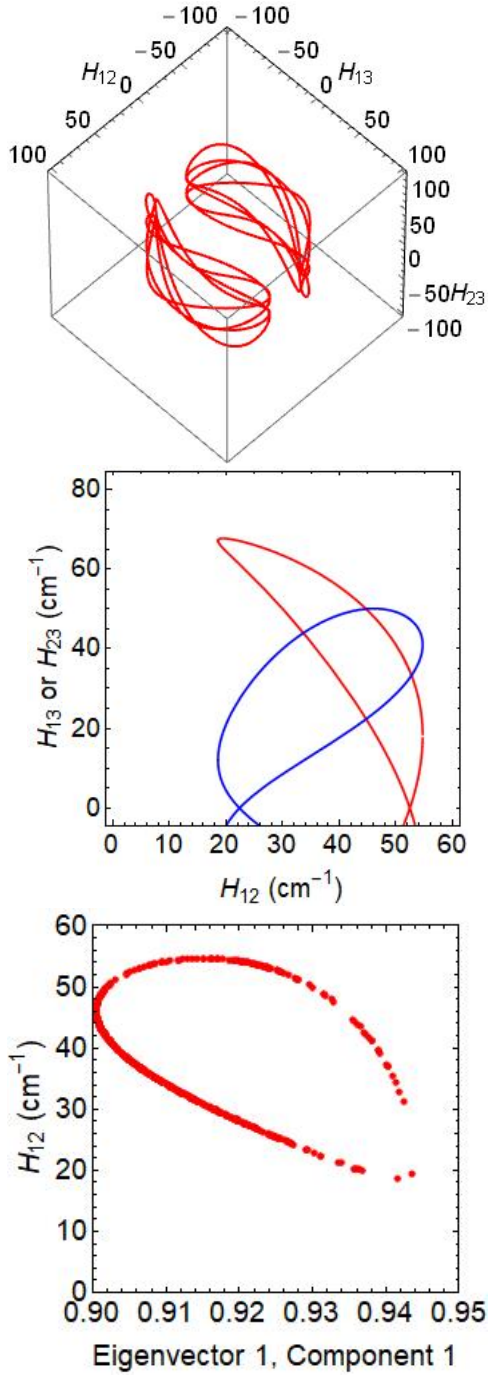

Fig. 2 Constraints imposed by the determinant equation for $n_{dim} = 3$. Top panel: The closed curves describing the solution to Eq. (2) in the 3D space of H_{12}, H_{13}, and H_{23}. Middle panel: projection of the curve onto the positive-positive quadrants of H_{13} vs H_{12} (red) and H_{23} vs H_{12} (blue), showing the limits of these variables. Bottom panel: The limit on the first component of the first eigenvector imposed by the limit on H_{12}.

that are large, is usually tighter than the limit on the possible off-diagonal elements. These components may provide very useful information. However, because the system becomes increasingly under-determined as n_{dim} increases, the limit on the range of eigenvector components becomes looser as n_{dim} increases.

In the algorithms that follow, we characterize the success of the algorithm by comparing the resultant off-diagonal Hamiltonian elements or eigenvalue coefficients to those of the known system used for the test. Each algorithm is run for many trials, using random starting values, and each algorithm gives a range of answers for the elements or coefficients. While there is no reason to assume that the frequency of values is subject to normal statistics, our observation is that characterizing each off-diagonal Hamiltonian element or eigenvalue coefficient by a mean and a standard deviation is more successful than, for example, characterizing it by the center and width of the range of results.

3 Algorithms

We briefly summarize the theoretical background and computational algorithm before proceeding to examples. There are two main aspects of the method. We first consider the solution to the following inverse eigenvalue problem appropriate to the assignment and analysis of vibrational spectra: given a set of model energies (diagonal elements of an otherwise unknown Hamiltonian matrix) and a set of eigenvalues (the spectrum), is it possible to determine off-diagonal Hamiltonian elements and ultimately the eigenvectors that can be used to assign the spectrum? The surprising answer to this question is that determination of the major components of the eigenvectors is possible. Second, we summarize how information about transition dipole moments can improve the determination of these eigenvectors.

Inverse Eigenvalue Problems are well-known in many applications,[23] but it does not appear that determining eigenvectors from only the diagonal elements of a matrix and its eigenvalues has been attempted in the literature. The reason, no doubt, is that for all $n_{dim} > 2$ such a system is increasingly under-determined as the matrix dimension, n_{dim}, increases. The problem looks hopeless. Nonetheless, we find that when the eigenvectors have only a few components that comprise most of the probability, it is possible to estimate the magnitude of these components accurately because, for such systems, the mapping between Hamiltonians and eigenvectors is such that a wide range of Hamiltonians map to a small range of eigenvectors. Thus, determination of the best Hamiltonians that fit the eigenvalues gives a reasonable estimate of the principal eigenvector components. We provide an application to CO_2 in the Results section below.

The second major component of our method involves the transition dipole moments, which determine the intensities in the spectrum. We briefly review the basic quantum theory of the IR spectrum. Recall that the intensity of an infrared transition at ν_f from the ground vibrational state to an excited state $|f\rangle$ is given rigorously by

$$I \propto \nu_f \left| \langle f | \vec{\mu} | 0 \rangle \right|^2, \tag{3}$$

where there is an equation such as (3) for each eigenstate (i), where ν_f is the frequency corresponding to the energy of eigen-

state (i), and $\vec{\mu}$ is the dipole moment, which depends on the molecular vibrational coordinates. In the double-harmonic approximation, the vibrational states are harmonic ones and $\vec{\mu}$ is approximated by a simple linear expansion

$$\vec{\mu} = \vec{\mu}_0 + \sum_{n=1}^{3N-6} \frac{\partial \vec{\mu}}{\partial Q^{(n)}} Q^{(n)} + ..., \quad (4)$$

$$I \propto \nu_f \frac{1}{2\omega_f} \left| \frac{\partial \vec{\mu}}{\partial Q^{(f)}} \right|^2, \quad (5)$$

where ω_f is the harmonic frequency and where, as a result of truncation of the expansion in Eq. (4), the only allowed transitions are fundamental ones.

In the rigorous approach, the exact wavefunctions are expanded in a complete orthonormal basis $\{\chi_k\}$, which here is the harmonic basis. Thus, $|f\rangle = \sum_k c_k^{(i)} \chi_k$ and the intensity is given by

$$I \propto \nu_f \sum_{\alpha=x,y,z} \sum_{k,k'} c_{k'}^{(i)} c_k^{(i)} \langle k' | \mu_\alpha | 0 \rangle \langle k | \mu_\alpha | 0 \rangle, \quad (6)$$

where for simplicity the ground state is (well-) approximated by the harmonic ground-state wavefunction. This equation is central to our approach.

Before proceeding, and at the risk of stating the obvious, it is important to note that if the eigenvectors $c_k^{(i)}$ are the exact ones, and the dipole moment function is also exact, then the above equation gives the exact spectrum, both in terms of the band positions and intensities. With this in mind, we can state the objective. The goal is determine these eigenvectors from experimental band positions and intensities. The "known" quantities are the harmonic frequencies and, if available, the dipole matrix elements above.

A search procedure is the basis of the approach we use here. Basing the search on the energies is known as an Inverse Eigenvalue Problem, which is very under determined. Adding information about intensities should in principle aid in the search. Two general physical insights also assist the search algorithm. The first is that the intensity of the peaks in most complex spectra is dominated by a few "bright" states, typically fundamentals, that mix with many "dark" states, typically combination states. This conjecture seems to be reasonable because the expansion of the dipole moment, given by Eq. (4) is usually truncated at the second term due to the fact that the second partial derivatives and cross derivatives are often very small relative to first derivatives. For example, in the case of FAD, Qu et al.[45] have found a two-order-of-magnitude difference between coefficients of the second and third terms of the expansion. If only the first two terms are included, combination bands have no intensity, so that all of the oscillator strength must be carried by the fundamentals. Other states will absorb only to the extent that they take on the character of the bright fundamentals via a Hamiltonian that mixes them.

A second physical insight is that interaction between states is usually appreciable only if the states are near one another in energy. For example, in a two-state system, the strength of the interaction varies as the reciprocal of the energy difference between the basis states. The resulting eigenstates have the same average energy as that of the basis states, but one is shifted to higher energy than the upper basis state, while the other is shifted to lower energy than the lower basis state.

Thus, our method assumes that we have a spectrum, usually from experiment, but it presupposes only a very limited knowledge of the potential, i.e., just the harmonic basis energies for the molecule. If, additionally, we have information about the relative oscillator strengths of the fundamental modes that lie in or near the region the spectrum, we will use this information, but it is not absolutely required. We note that the transition dipole moments and fundamental vibrational frequencies are often provided by even the most basic molecular structure calculations. In the case of FAD, more sophisticated calculations[45] have provided the squared dipole transition elements for the O-H, C=O, and C-H stretching modes (see Table S4), and we incorporate this information into the method.

The computational method used is what we call a Monte Carlo Diagonalization (MCD) method. As shown in Fig. 1*b*, initiation of the method requires a starting Hamiltonian, $\mathbf{H}_{guess}$, taken to be a matrix with zeros as the off-diagonal elements and with the harmonic basis energies as the diagonal elements. This Hamiltonian is used to compute a calculated spectrum using Procedure 1, which consists of four steps: a) diagonalizing $\mathbf{H}_{guess}$ to obtain its eigenvalues and eigenvectors; b) creating a Gaussian peak at each eigenvalue for which the amplitude is proportional to the eigenvalue times the weighted squared dipole matrix element from Eq. (6), c) multiplying the calculated spectrum by a constant chosen to best fit the experimental spectrum; and d) determining the variance from point-by-point comparison between the calculated and experimental spectra. This completes the initiation step. Further detail on steps b) and c) is included in the Dipole Transition Moment section (S2) of the ESI as well as in Table S4.[45]

The iterative step in the method is to vary the initial $\mathbf{H}_{guess}$ by adding a small random number to each off-diagonal element while maintaining the symmetry of the Hamiltonian. Procedure 1 is applied to the new Hamiltonian, and if the variance decreases, the new Hamiltonian is accepted; if not, the old one is retained. This step is then repeated for a large number of iterations (1500-10000) until there is no longer improvement in the fit to the experimental spectrum. In the CO_2 example, which we consider first in the next section, the iterations conclude when the variance in the eigenvalues reaches a minimum. Other methods and comments on algorithms can be found in Section S1 of the ESI.

4 Results

2-mode CO_2

The MCD procedure is verified for a 2-mode CO_2 example that exhibits classic Fermi resonances due to the near 2:1 ratio of the bend and symmetric stretch frequencies. Note, in our procedure, we do not assume the existence of this resonance.

Our calculations make use of a recent *ab initio*, coupled-cluster quartic force field (QFF) due to Rodriguez-Garcia et al.:[1]

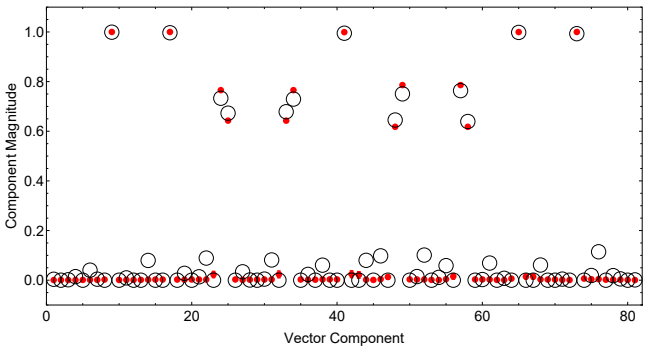

Fig. 3 Calculated and actual eigenvector magnitudes for a 9×9 model for CO_2 based on a quartic force field. The red points with (barely visible) error bars are the calculated results using the MCD algorithm with the harmonic diagonals, whereas the open circle centers are the actual values. The eigenvalue components are numbered 1-81, with 9 for each vector, starting with the vector corresponding to the lowest energy eigenvalue. Note that while five of the states have one large component near unity, four of the states are mixed, with two components each near 0.7. These are the states mixed by the Fermi resonance.

$$V^{QFF} = V_0 + \sum_{i=1}^{4} g_i Q_i + \frac{1}{2}\sum_{i=1}^{4} h_{ii} Q_i^2 + \frac{1}{6}\sum_{i=1}^{4} t_{iii} Q_i^3 + \frac{1}{24}\sum_{i=1}^{4} u_{iiii} Q_i^4 + \frac{1}{2}\sum_{i,j=1}^{4} t_{ijj} Q_i Q_j^2 + \frac{1}{4} u_{iijj} Q_i^2 Q_j^2. \quad (7)$$

They determined the coefficients from the potential calculations and then used vibrational self-consistent field and vibrational configuration-interaction methods to calculate the anharmonic energy levels, which were found to be within a mean absolute deviation of 3.5 cm^{-1} from the experimental values.

The test of the algorithm for determining eigenvectors from known eigenvalues and harmonic basis functions is restricted to the symmetric stretching and bending modes. Thus, only terms in the QFF involving these modes were considered. The benchmark, target eigenvalues come from a diagonalization of a 9×9 H matrix, built from a direct-product of $v = 0, 1, 2$ harmonic functions for each mode. The diagonal harmonic energies of the starting 9×9 H matrix are (1015.8, 1696.3, 2358.4, 2376.8, 3038.9, 3719.4, 3701.0, 4381.5, 5062.0), where all energies are in cm^{-1}. (The QFF matrix is given in detail in Table S5 of the SM.)

Figure 3 shows the magnitudes of the nine MCD trained (red dots) and "exact" (open circles) eigenvectors, which are given in increasing energy going from left to right on the x-axis. Note that vector numbers 3, 4 and 6, 7 have values around 0.7. These are two sets of Fermi resonance doublets. The 3, 4 pair correspond to the plus and minus combinations of $2\nu_b$ and ν_s and the 6, 7 pair to the plus and minus combinations of $\nu_s + 2\nu_b$ and $2\nu_s$. The MCD algorithm captures these very well, with, as noted above, no prior knowledge of them. It is simply the difference between the assumed-known harmonic energies and the exact ones that leads to the correct "training". The splitting between these pairs of doublets in the harmonic approximation is 18.4 cm^{-1} for both pairs, 78 and 107 cm^{-1} respectively for the "exact" and trained

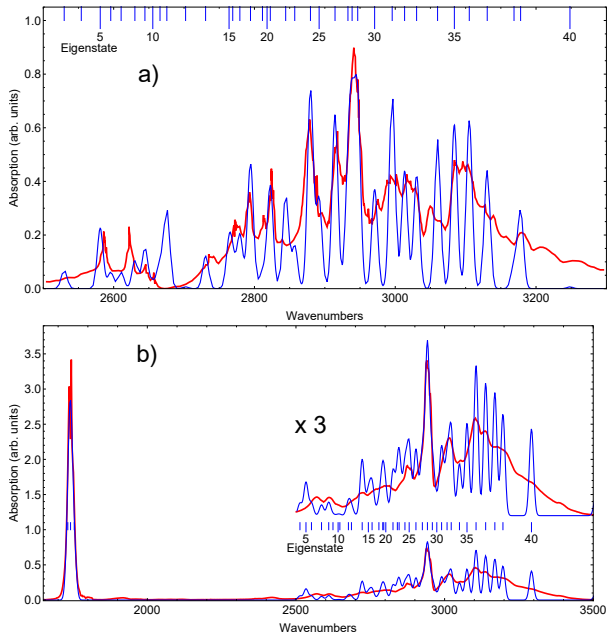

Fig. 4 *a)* Calculated (blue) and jet-cooled experimental[31] (red) spectra of the formic acid dimer. *b)* Calculated (blue) and room-temperature experimental[40] (red) spectra of the formic acid dimer.

doublets. Thus, the MCD algorithm succeeds in capturing the values of the vector component magnitudes and, in particular, providing an accurate analysis of the Fermi Resonance. Overall, the correlation between the eigenvector magnitudes from the original Hamiltonian and those determined from the calculation using only the eigenvalues and the basis functions has an R^2 value of 0.9882. Results for a few other systems, some more complicated, some less, are provided in the ESI section on Supplementary Results.

The Formic Acid Dimer

We now turn to the application of the methods described above to the spectrum of the formic acid dimer. A choice of basis set must be made from the harmonic basis energies. If we consider the range from 1600-3500 cm^{-1}, then even if we were to include only those fundamentals, overtones, and combination bands of the in-plane B_u modes that have a maximum of 4 fundamental components, we would need to consider more than 380 basis states. A much smaller but still very reasonable selection of 42 basis states has been made for the calculations reported below. These include the three fundamentals carrying oscillator strength: ν_{20}, the C=O stretch, ν_{22}, the C-H stretch, and ν_{24}, the O-H stretch. Most of the remaining states were chosen to be in the 2500-3300 cm^{-1} region, where a jet-cooled infrared spectrum has been reported.[31] Three were in the region of the C=O stretch near 1750 cm^{-1} and two, including ν_{24}, were above 3300 cm^{-1}. A list of the basis states and their harmonic energies is provided in the first three columns of Table S6.

Figure 3*a* shows a jet-cooled spectrum and a calculated fit for FAD in the 2500-3300 cm^{-1} region. The algorithm used to produce the blue calculated spectrum has been shown in Fig. 1*b* and

was used with the following modifications. At first we performed the fit as described above and in Fig. 1*b*, but we later found that more rapid convergence could be achieved by alternating on each iteration between a step in which all off-diagonal elements of $H_{i,j}$ were modified by a small random variation and a step in which only those elements involving the "bright" states, ν_{20}, ν_{22}, and ν_{24} were modified. To a first approximation, the first step lets the dark states interact with one another to give the correct peak positions, whereas the second step adjusts the amplitudes. In addition, if a substantial number of iterations occurred without improvement, we halved the range of the random changes. The FWHM used for the calculated peaks was set at the apparent experimental resolution, 11.8 cm^{-1}. The fits would be improved by increasing this broadening (32 cm^{-1} minimizes the variance), but it would then be more difficult to identify which calculated peaks and eigenstates belonged to the various spectral features.

As can be observed from the calculated (blue) and experimental (red) spectra in Fig. 4*a*, nearly every peak in the experimental spectrum corresponds to one (or more) peaks in the calculated spectrum. The identities of the peaks are given by their eigenstate number, shown in the comb at the top of the figure. For each eigenstate, the three basis function components with the largest absolute magnitudes are listed in Table S6; these comprise an assignment of the spectrum. A visual depiction of the assignments is provided in the top panel Fig. 5*a*, where the probabilities of each basis state are color coded and shown as a function of the energy. The basis states run from number 1 at the top row to number 42 at the bottom row; they are identified in the first three columns of Table S6. We note that the spectrum based on harmonic oscillators would have only three lines; these would be at 1780 cm^{-1} (ν_{20}, off-scale), 3097 cm^{-1} (ν_{22}), and 3326 cm^{-1} (ν_{24}, off-scale). Thus, the calculated spectrum is a very significant improvement over that predicted by the basis model.

In order to see if the fitting procedure provided proper relative intensities for different spectral regions, we also used the same procedure to fit the room-temperature spectrum.[40] Figure 4*b* shows the result for a FWHM broadening of 20 cm^{-1}. Of course, the lower resolution of this spectrum, and perhaps complications from rotational structure and hot bands, result in a less accurate assignment. Nonetheless, the relative amplitudes of the eigenstates near 1750 cm^{-1} and those in the region 2500-3500 cm^{-1} are correctly reproduced. In addition, there is considerable overlap between the positions and assignments for the spectra in Figs. 4*a* and 4*b*. The eigenvalues for the two are nearly the same, with a correlation R^2 value of 0.987; a correlation plot is shown in Fig. S3. Comparison for each eigenstate between the group of components with the largest three magnitudes showed a 48% match between the two fits. Table S7 provides the assignments for the room-temperature spectrum. In addition, Figs. S4 and S5 show array plots of the eigenvectors for the jet-cooled and room-temperature fits, respectively; Tables S8 and S9 provide the probability amplitude components for each bright state in every eigenvector for the jet-cooled and room-temperature fits, respectively; and Fig. S6 shows an array plot of the Hamiltonian matrix for the fit to the jet-cooled FAD spectrum shown in Fig. 4*a*.

It is interesting to note to what extent the bright state prob-

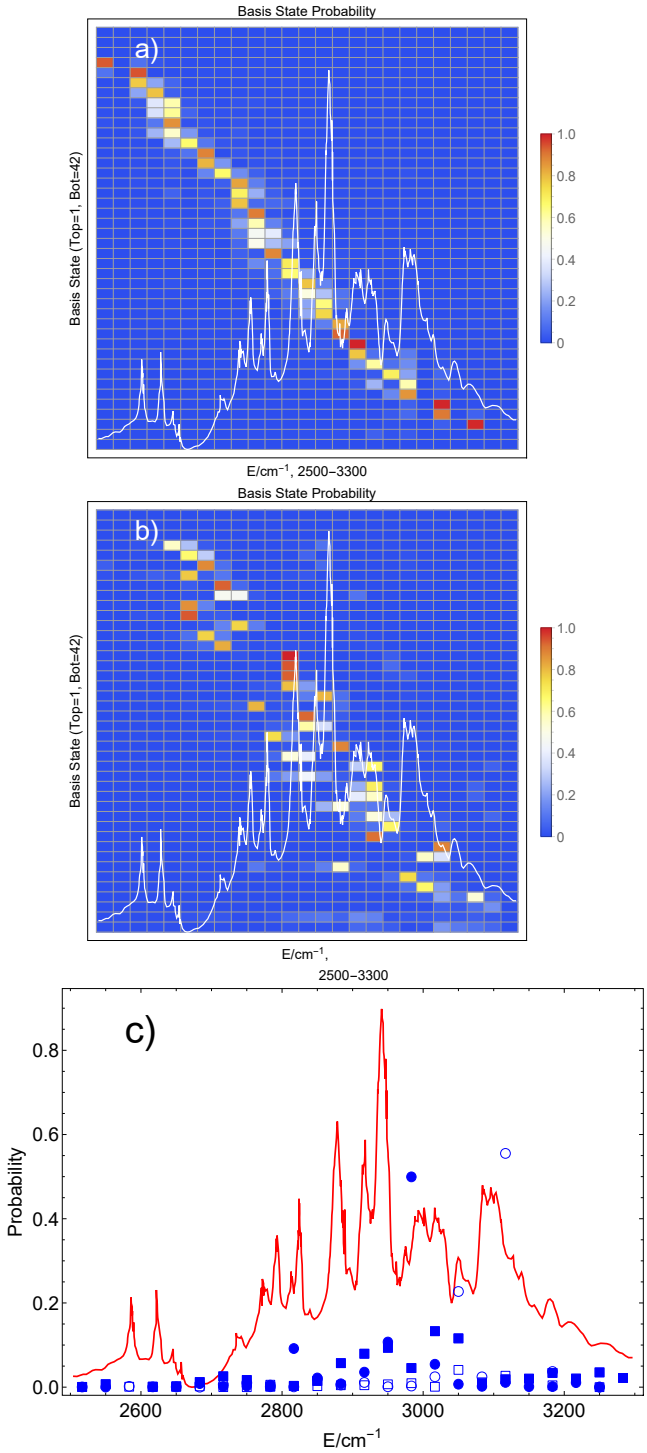

Fig. 5 Array plots giving the probability of basis state contributions to different spectral regions. *a)* MCD results. *b)* MM results. The probability is given by the color code provided by the bar at the right. In each panel, basis state numbers run from 1 (top) to 42 (bottom). The experimental spectrum[31] is in white. *c)* OH stretching (squares) and CH stretching (circles) probabilities for each eigenvector in the spectrum of FAD plotted as a function of eigenvalue. The solid data is from the MM calculation, whereas the open data is from the MCD calculation. The experimental spectrum[31] is in red.

abilities are distributed into the many dark states. Fig. 5*c* plots as a function of the the energy the probabilities for the OH and CH stretching modes (ν_{24} and ν_{22}) as open squares and circles, respectively. (The filled squares and circles will be discussed below.) The probability for these bright states is almost completely diffused into the dark states. With the exception of a few eigenstates, the probability of the bright states is quite small, even though they carry all of the intensity. Note that the position and intensity of the peaks at eigenvalue 42 and the position of eigenvalue 41 (ν_{24}), both of which are off the scale of the figure, are an artifact of the limited size of the basis set.

5 Discussion

To our knowledge, this assignment for the spectrum of the FAD in the 2500-3300 cm^{-1} region is the first one that so successfully reproduces the positions and amplitudes of the peaks observed from the jet-cooled and room-temperature spectra (Figs. 4*a* and 4*b*). Although the two spectra give assignments that do not coincide exactly, it is to be expected that the assignments from the lower-resolution room-temperature spectrum will not be as accurate as those from the jet-cooled spectrum, which has a much higher information content. It is interesting to note from the jet-cooled fit that the mixing between the dark states is mostly limited to a very few basis functions whose energies in each direction are closest to the state under consideration. This results in a predominantly diagonal array plot for the eigenvector component magnitudes, as can be seen in Fig. S4. Another interesting point from the fit is how the oscillator strength for each of the ν_{24}, ν_{22} and ν_{20} basis states becomes diluted into the dark states. The data, shown in tables S3-S4 and in Fig. 5*c*, give the probability amplitudes and probabilities, respectively, of these bright states for each eigenstate or spectral region. From this data it is clear that the zero-order states ν_{24} and ν_{22} are, with the exception of a few eigenstates, a minor contributor in terms of this amplitude, yet the major contributor to the intensity.

Generalizing and thinking about further applications of the method, one might ask about the accuracy of the assignments. We provide three answers to this question, each involving a particular test. The first test was to repeat the Monte Carlo analysis to determine the consistency of the results. Of four fits to the spectrum we find excellent agreement on the eigenvalues, as expected since we train on these, and quite satisfactory agreement on the eigenvectors. The correlation between any two sets of eigenvalues is characterized by an R^2 of at least 0.9996. When comparing any pair of calculations, there is generally a 62% match when determining the most important 1, 2, or 3 basis states; comparing only the most important component gave a 70% match. Similarly, for any pair of calculations a correlation plot for the complete set of vector component magnitudes has an average R^2 of 0.63.

As a second test, we start from a (red, see Fig. 1*a*) Hamiltonian that is physically reasonable in the sense that it can reproduce the experimental jet-cooled spectrum. We then change every off-diagonal element by a random amount, up to 150%, 100% or 50% of the original value. In each case, we diagonalize the Hamiltonian and make a spectrum, keeping track of the (red) eigenvalues and eigenvectors. We then perform the algorithm described in Fig. 1*b* on each of the three spectra, producing (blue) eigenvalues and eigenvectors, and we finally compare these to the (red) set produced from the modified Hamiltonian. For the 150%, 100% and 50% variations we find, respectively, that the R^2 values for the correlation of the (eigenvalues, eigenvector component magnitudes) are (0.9982, 0.578), (0.9997, 0.766), and (0.9994, 0.773). Perhaps more importantly, for 150%, 100% and 50% variations, comparisons of the most important 1, 2, or 3 eigenvector components give match rates of (63.1, 68.5, and 70.8)% . Having an automated method for assigning an unknown spectrum at even *ca.* 60-70% accuracy with little more than the harmonic basis functions as inputs should be a useful tool for all those trying to understand vibrational spectroscopy in molecules with many vibrational modes.

The third test is to compare the MCD results calculated here to more traditional results performed using a calculated potential energy surface[46] and MULTIMODE (MM) with a large basis set.[45] This comparison should be made with some caution, since the MCD calculation used a basis set of 42 states, whereas the MM calculation used one of 20660 states. The number of eigenstates per cm^{-1} in the 2500-3300 cm^{-1} region is about 0.05 in the MCD calculation, whereas it is about 1.6 in the MM calculation. In addition, the decomposition of MM eigenstates into the harmonic oscillator states is only approximately achieved, whereas it is directly achieved in the MCD calculation. With these caveats, one can compare the panels of Fig. 5*a* and 5*b*, both of which show the energies to which the basis states used the MCD calculation contribute. Not surprisingly, there is a bit more mixing and splitting of these states in the MM calculation, but the general trend of the data shows an overall similarity. In both cases, the basis states that mix to give contributions at a particular energy are those that are close to that energy. Generally, for both calculations there are only a few basis states that have large contributions to any energy bin. A second comparison can be made using Fig. 5*c*, which shows the dispersion of the ν_{24} and ν_{22} probabilities into different transition energy regions. There is some disagreement on the largest peaks due to the differences in basis set, but both calculations agree that the probability for both modes is spread out over a wide range of energies. Interestingly, although the probabilities at most of these energies are small, these small probabilities are responsible for the entire intensity of the spectrum.

The general conclusion from these tests is that the MCD gives a very good qualitative picture of what basis states are associated with each spectral feature. The quantitative agreement that can be expected for individual assignments or groups of assignments is somewhere between 40 and 70 %.

There are certainly some general properties of the potential that we might learn from analysis of the spectrum, even if determining the full potential from the spectrum (the blue path in Fig. *a*) is illusive. The principal components of the eigenvectors can be examined for resonances. For example, nearly symmetrical 2×2 or 3×3 blocks in array plots such as that in Fig. S4 may indicate a resonance, Fermi or otherwise. In this figure, for example, and from Table S6 we see that basis states 38 and 39 are nearly symmetrically mixed, suggesting a Fermi resonance. Indeed, the energy of the two basis states is nearly the same be-

cause there is a resonance between ν_{19} at 1715 cm^{-1} and $\nu_6 + \nu_{17}$ at 1723 cm^{-1}. If we assume that this resonance is the only factor that splits the two eigenstates, then we can deduce something about the coefficient of the cross anharmonicity in the potential that causes the mixing. As this example shows, even if we cannot generally go from the spectrum to the potential along the blue path, we can nonetheless obtain some insight about the potential from an assignment of the spectrum.

6 Conclusions

It is nearly certain that additional research along the lines suggested here could further improve the assignment method. Indeed, this problem seems to be one that might be amenable to machine learning. Knowledge of the relationships between the basis energies and the eigenvalues and eigenvectors controlling the spectrum learned from successful assignments might be analyzed and used for unknown assignments. Additionally, the procedures here could help to understand which terms in spectroscopic Hamiltonians might be most important, perhaps enabling that technique to extend to larger systems than, for examples, the very successful applications to HCCH, HCP, CH_3O, $CSCl_2$, CDBrClF, HO_2, etc.[11,13,15,17–19] The important message of the research presented here is that the (red) path from potential to spectrum is not the only method for spectral assignment. A new paradigm, the blue path, might equally well be exploited.

7 Acknowledgements

Support from the National Science Foundation (CHE-1463552) is gratefully acknowledged.

References

1 V. Rodriguez-Garcia, S. Hirata, K. Yagi, K. Hirao, T. Taketsugu, I. Schweigert, and M. Tasumi, *J. Chem. Phys*, 2007, **126**, 124303/1–5.
2 J. M. Bowman, S. Carter and X. Huang, *Int. Rev. Phys. Chem.*, 2003, **22**, 533–549.
3 F. Iachello, *Interacting Bosons in Nuclear Physics*, Plenum Press, New York, 1979.
4 C. E. Wulfman and R. D. Levine, *Chem. Phys. Lett.*, 1979, **60**, 372–376.
5 O. S. van Roosmalen, F. Iachello, R. D. Levine and A. E. L. Dieperink, *J. Chem. Phys.*, 1983, **79**, 2515–2536.
6 I. Benjamin and R. D. Levine, *Chem. Phys. Lett.*, 1985, **117**, 314–320.
7 A. Frank, R. Lemus and F. Iachello, *J. Chem. Phys.*, 1989, **91**, 29–41.
8 F. Iachello, S. Oss and R. Lemus, *J. Mol. Spectrosc.*, 1991, **146**, 58–78.
9 F. Iachello and R. D. Levine, *Algebraic Theory of Molecules*, Oxford University Press, New York, 1994.
10 L. E. Fried and G. S. Ezra, *J. Chem. Phys.*, 1987, **86**, 6270–6280.
11 M. E. Kellman, *Journal of Chemical Physics*, 1990, **93**, 6630–6635.
12 A. B. McCoy and E. L. Sibert, *J. Chem. Phys.*, 1991, **95**, 3476–3487.
13 H. Ishikawa, C. Nagao, N. Mikami and R. W. Field, *J. Chem. Phys.*, 1998, **109**, 492–503.
14 M. Mekhtiev and J. Hougen, *J. Mol Spec.*, 2000, **199**, 284–301.
15 C. Jung, H. S. Taylor and M. P. Jacobson, *Journal of Physical Chemistry A*, 2001, **105**, 681–693.
16 H. Waalkens, C. Jung and H. S. Taylor, *J. Phys. Chem. A*, 2002, **106**, 911–924.
17 C. Jung, C. Mejia-Monasterio and H. S. Taylor, *Journal of Chemical Physics*, 2004, **120**, 4194–4206.
18 C. Jung, H. S. Taylor and E. L. Sibert, *Journal of Physical Chemistry A*, 2006, **110**, 5317–5325.
19 G. L. Barnes and M. E. Kellman, *Journal of Chemical Physics*, 2011, **134**, 074108:1–11.
20 G. L. Barnes and M. E. Kellman, *J. Chem. Phys.*, 2012, **136**, 024114.
21 P. R. Franke, D. P. Tabor, C. P. Moradi, G. E. Douberly, J. Agarwal, H. F. SchaeferIII and I. Edwin L. Sibert, *J. Chem. Phys.*, 2016, **145**, 224304.
22 B. A. Johnson and E. L. Sibert, *J. Chem. Phys.*, 2017, **146**, 174112.
23 M. T.-C. Chu and G. H. Golub, *Inverse Eigenvalue Problems: Theory, Algorithms and Applications*, Oxford University Press, Oxford, England, 2005.
24 V. V. Vener, O. Kuhn and J. M. Bowman, *Chem. Phys. Lett.*, 2001, **349**, 562–570.
25 R. C. Millikan and K. S. Pitzer, *J. Am. Chem. Soc.*, 1958, **80**, 3515–3521.
26 J. E. Bertie and K. H. Michaelian, *J. Chem. Phys.*, 1982, **76**, 886.
27 J. E. Bertie, K. H. Michaelian, H. H. Eysel and D. Hager, *J. Chem. Phys.*, 1986, **85**, 4779.
28 T. Wachs, D. Borchardt and S. H. Bauer, *Spectrochim. Acta A*, 1987, **43**, 965.
29 Y. Maréchal, *J. Chem. Phys.*, 1987, 6344.
30 F. Ito and T. Nakanaga, *Chem. Phys.*, 2002, **277**, 163.
31 R. Georges, M. Freytes, D. Hurtmans, I. Kleiner, J. Vander Auwera and M. Herman, *Chem. Phys.*, 2004, **305**, 187.
32 F. Ito, *Chem. Phys. Lett.*, 2007, **447**, 202.
33 P. Zielke and M. A. Suhm, *Phys. Chem. Chem. Phys.*, 2007, **9**, 4528.
34 F. Ito, *J. Chem. Phys.*, 2008, **128**, 114310.
35 Y. H. Yoon, M. L. Hause, A. S. Case and F. F. Crim, *J. Chem. Phys.*, 2008, **128**, 084305.
36 A. Olbert-Majkut, J. Ahokas, J. Lundell and M. Pettersson, *Chem. Phys. Lett.*, 2009, **468**, 176.
37 F. Kollipost, R. W. Larsen, A. V. Domanskaya, M. Norenbert and M. A. Suhm, *J. Chem. Phys.*, 2012, **136**, 151101.
38 C. Emmeluth, M. A. Suhm and D. Luckhaus, *J. Chem. Phys.*, 2003, **118**, 2242–2255.
39 M. W. Nydegger, W. Rock and C. M. Cheatum, *Phys. Chem. Chem. Phys.*, 2011, **13**, 6098–6104.

40 K. Mackeprang, Z.-H. Xu, Z. Maroun, M. Meuwly and H. G. Kjaergaard, *Phys. Chem. Chem. Phys.*, 2016, **18**, 24654–24662.
41 G. M. Florio, T. S. Zwier, E. M. Myshakin and K. E. Jordan, *J. Chem. Phys.*, 2003, **118**, 1735–1746.
42 I. Matanović and N. Došlić, *Chem. Phys.*, 2007, **338**, 121.
43 G. L. Barnes and E. L. Sibert, *J. Mol. Spect.*, 2008, **249**, 78.
44 G. A. Pitsevich, A. E. Malevich, E. N. Kozlovskaya, I. Y. Doroshenko, V. Sablinskas, V. Pogorelov, D. Dovgal and V. Balevicius, *Vib. Spectrosc.*, 2015, **79**, 67.
45 C. Qu and J. M. Bowman, *submitted*, 2018, **xx**, yyy–yyy.
46 C. Qu and J. M. Bowman, *Phys. Chem. Chem. Phys.*, 2016, **18**, 24835.