Balancing accuracy and efficiency in selecting vibrational configuration interaction basis states using vibrational perturbation theory

Marat Sibaev and Deborah L. Crittenden
Department of Chemistry, University of Canterbury, Christchurch, New Zealand

(Dated: 29 January 2017)

This work describes the benchmarking of a VCI algorithm that combines the favourable computational scaling of VPT2 with the algorithmic robustness of VCI, in which VCI basis states are selected according to the magnitude of their contribution to the VPT2 energy, for the ground state and fundamental excited states. Particularly novel aspects of this work include: expanding the potential to 6th order in normal mode coordinates, using a double-iterative procedure in which configuration selection and VCI wavefunction updates are performed iteratively (micro-iterations) over a range of screening threshold values (macro-iterations), and characterisation of computational resource requirements as a function of molecular size. Computational costs may be further reduced by a priori truncation of the VCI wavefunction according to maximum extent of mode coupling, along with discarding negligible force constants and VCI matrix elements, and formulating the wavefunction in a harmonic oscillator product basis to enable efficient evaluation of VCI matrix elements. Combining these strategies, we define a series of screening procedures that scale as \( O(N_{\text{mode}}^6) - O(N_{\text{mode}}^9) \) in run time and \( O(N_{\text{mode}}^6) - O(N_{\text{mode}}^7) \) in memory, depending on the desired level of accuracy. Our open-source code is freely available for download from http://www.sourceforge.net/projects/pyvci-vpt2.

I. INTRODUCTION

Advances in electronic structure methods and computer technology enable increasingly accurate modelling of electronic properties for ever larger and more complex chemical systems. In contrast to the sustained and ongoing progress in the development and application of methods for solving the electronic Schrödinger equation, harmonic normal mode analysis remains the predominantly used method for solving the nuclear vibrational Schrödinger equation.

This can largely be attributed to the difficulty and computational cost associated with modelling multidimensional anharmonic potential energy surfaces (PES). We have recently implemented a coordinate transformation procedure to help address this problem,\(^1\) based upon obtaining concise, low order PES expansions in internal coordinates and then transforming to higher order expansions in normal mode coordinates,\(^2-4\) to facilitate solving the nuclear vibrational problem. This minimizes the number of computationally intensive \textit{ab initio} calculations required to accurately describe the PES.

Therefore, we now turn our attention to investigating general, scalable and accurate methods for solving the nuclear vibrational problem. Any two of these constraints can be readily satisfied by standard methods; harmonic normal mode analysis\(^5\) and vibrational self-consistent field theory (VSCF)\(^6-13\) are general and scalable, but not particularly accurate, vibrational configuration interaction (VCI)\(^13-15\) and vibrational coupled cluster (VCC)\(^13,16-20\) methods are accurate and general, but scale badly with molecular size. Methods based upon expressing both the kinetic energy operator and potential energy surface in internal coordinates\(^21,22\) are highly accurate and potentially scalable, but lack generality. Normal mode coordinates are the only practical choice for a black box nuclear vibrational structure model.\(^23\)

Second-order vibrational perturbation theory (VPT2)\(^24-34\) displays a promising balance of generality, scalability and accuracy, but naive implementations suffer from numerical instability due to accidental near-degeneracies leading to divergence of the VPT2 energy expression. More sophisticated implementations avoid this problem by constructing and diagonalizing VCI sub-matrices for sets of resonant frequencies.\(^24-32,35,36\) However, the highly efficient analytical implementations that rely on pre-determining sets of resonant states do not always capture all relevant states and their couplings.\(^28,35,37-38\) Even more rigorous methods, based upon van Vleck perturbation theory\(^27,28,37-40\) with correct treatment of resonances, still rely on accurate identification of resonant terms.\(^41,42\) This is likely to become increasingly problematic for larger systems with a greater number of, and more extensively coupled, resonant states.

An alternative, although less popular, way of combining VCI and VPT2 is to first perform a VCI calculation and then apply a VPT2 correction.\(^43,44\) This is conceptually appealing, as VCI naturally captures strong resonances, while VPT2 is designed to capture small perturbations from a reference state. However, this procedure is only cost-effective if the VCI matrix contains exclusively or predominantly strongly coupled basis states, with the remaining basis states are only weakly coupled to the resultant VCI states. This may be achieved by using VPT2-based screening procedures to select strongly-coupled basis states for inclusion in the VCI matrix.\(^45-46\)

Even though VPT2-screening procedures reduce the number of elements explicitly included in the VCI matrix, previous studies have found that memory require-
ments can still become prohibitive, even when performing case studies on small molecules and capturing weak coupling using a post-VCI VPT2 correction.\textsuperscript{43, 44} A range of strategies have been explored to selectively identify only strongly-coupled VCI basis states, including; using a state-specific approach that involves performing separate VCI calculations for each vibrational state of interest,\textsuperscript{44–46} restricting the maximum excitation level, below the convergence limit,\textsuperscript{45–47} restricting the extent of mode-coupling in excited states, particularly for larger molecules,\textsuperscript{43, 44} and mode-by-mode customization of the VCI basis.\textsuperscript{48} Disadvantages associated with these strategies include increasing run-time, decreasing accuracy or forfeiting accessibility by requiring expert user input.

Here, we address the VCI matrix construction, diagonalization and storage memory bottleneck problem differently, using a modified double-iterative VPT2 screening procedure to minimize the number of configurations selected, storing only non-negligible matrix elements in sparse matrix format, and using sparse matrix diagonalization routines.

Previous studies have also indicated that time requirements can also become substantial.\textsuperscript{43–48} To reduce overall run time, the most commonly employed strategy is to decrease the time required to generate and evaluate the potential energy surface, either by limiting the dimensionality of the potential energy surface and/or using a truncated series expansion to represent the PES, typically a quartic force field.\textsuperscript{43–48} However, not only does this decrease accuracy, but it can destabilize the VCI wavefunction at higher excitation orders, causing divergence in calculated frequencies.\textsuperscript{49} Therefore, in the present work, we employ sextic force fields throughout,\textsuperscript{50} and explicitly quantify the effect of using reduced mode representations timings and accuracy of results. We also save time by constructing a single VCI matrix to describe the vibrational ground state and all fundamental excited states simultaneously, rather than taking a state-specific approach.

The overall aim of the present work is to choose optimum representations of the PES and the basis set wavefunctions for achieving most accurate results and computational efficiency. Systematic truncation of both potential energy surface and wavefunction expansions makes it computationally feasible to calculate anharmonic vibrational spectra for larger molecules than achievable previously. However, the larger the molecule, the greater the extent of truncation that will be required. To this end, we also explore convergence behaviour with respect to each user-adjustable screening parameter and/or convergence threshold, to derive informed error estimates.

\section{II. THEORY AND ALGORITHMS}

\subsection{A. VPT2}

The vibrational Hamiltonian must be partitioned into a primary analytically solvable component and a secondary weak perturbation. We follow the usual approach of separating out the kinetic energy and harmonic potential terms, leaving the anharmonic potential correction and, optionally, the Coriolis coupling operator.

\begin{equation}
\hat{H} = \hat{H}_0 + \Delta \hat{H}
\end{equation}

\begin{equation}
\hat{H}_0 = -\frac{1}{2}\sum_{i=1}^{M} \frac{\partial^2}{\partial Q_i^2} + \frac{1}{2!}\sum_{i=1}^{M} F_i Q_i^2
\end{equation}

\begin{equation}
\Delta \hat{H} = \frac{1}{3!} \sum_{i,j,k} F_{ijk} Q_i Q_j Q_k + \frac{1}{4!} \sum_{i,j,k,l} F_{ijkl} Q_i Q_j Q_k Q_l + \frac{1}{5!} \sum_{i,j,k,l,m} F_{ijklm} Q_i Q_j Q_k Q_l Q_m + \frac{1}{6!} \sum_{i,j,k,l,m,n} F_{ijklmn} Q_i Q_j Q_k Q_l Q_m Q_n + \left(\hat{H}_{\text{Cor}}\right)
\end{equation}

\begin{equation}
\hat{H}_{\text{Cor}} = -\sum_{\alpha} B_{\alpha} \sum_{i<j} \sum_{k<l} c_{ij}^\alpha c_{kl}^\alpha \left( Q_i \frac{\partial}{\partial Q_j^*} - Q_j \frac{\partial}{\partial Q_i^*} \right) (Q_k \frac{\partial}{\partial Q_l^*} - Q_l \frac{\partial}{\partial Q_k})
\end{equation}

For computational expediency, basis states are constructed as Hartree products of harmonic oscillator basis functions:

\begin{equation}
\Phi_n(Q_1, \ldots, Q_M) = \prod_{i=1}^{M} \phi_n(Q_i)
\end{equation}

where \( n \) is a string of quantum numbers \( n_1, \ldots, n_i, \ldots, n_M \), specifying the vibrational state across all \( M \) modes. The strings that define the complete set of possible VPT2 and VCI basis states are generated by specifying a maximum value for the sum of the vibrational quantum numbers, which will henceforth be referred to as the excitation level, with its value denoted in round brackets, \( e.g. \) VCI(8) matrix is indexed by configurations with a sum of vibrational quantum numbers of 8 or less. The VCI matrix may be further restricted by the number of modes excited simultaneously, denoted VCI(n,mMC), where is the excitation level and \( m \) is the mode-coupling level.

The VCI wavefunction is expressed as a linear combination of basis states:

\begin{equation}
\Psi_n(Q_1, \ldots, Q_M) = \sum_{n'} c_{nn'} \Phi_{n'}(Q_1, \ldots, Q_M)
\end{equation}
with associated energies $\epsilon_n$. We have previously described in detail how VCI matrix elements are evaluated and coefficients obtained.

The harmonic energy levels for each Hartree product basis state $\Phi_n(Q_1, \ldots, Q_M)$ are denoted by $\epsilon_{n,0}$. These are calculated by the usual harmonic normal mode analysis procedure.

It is also helpful to partition the set of all excited Hartree product basis states defined by a given maximum excitation level into mutually exclusive VCI and VPT2 expansion sets, indexed by $n$ and $n'$, respectively.

The general expression for the second order perturbation theory energy correction to a given VCI state, $n$, is then:

$$\Delta \epsilon_n = \sum_{n'} \frac{|\langle \Psi_n(Q_1, \ldots, Q_M) | \Delta \hat{H} | \Phi_{n'}(Q_1, \ldots, Q_M) \rangle|^2}{\epsilon_n - \epsilon_{n',0}}$$  

(7)

The sum runs over all excited states in the VPT2 expansion set but not the VCI set. This reduces to the usual non-degenerate VPT2 energy expression if the unperturbed wavefunction is harmonic i.e. $\Psi_n = \Phi_n$, $\epsilon_n = \epsilon_{n,0}$ and the VPT2 expansion set contains all harmonic excited states except $n$.

For future reference, we define the VPT2 pair energy, $\xi_{n,n'}$, as the contribution each VPT2 basis state, $n'$, makes to the overall VPT2 energy correction for a given VCI state, $n$:

$$\xi_{n,n'} = \frac{|\langle \Psi_n(Q_1, \ldots, Q_M) | \Delta \hat{H} | \Phi_{n'}(Q_1, \ldots, Q_M) \rangle|^2}{\epsilon_n - \epsilon_{n',0}}$$  

(8)

B. VPT2 based screening

The primary aim of our screening algorithm is to select a VCI sub-matrix from a larger “parent” VCI matrix that contains only the configurations required to describe the ground and fundamental excited states of the system i.e. configurations that couple strongly to each of these states, or are responsible for strong coupling between them. In the process, we also determine VPT2 corrections for each state, for each VCI sub-matrix generated.

Our iterative VPT2 based screening algorithm proceeds as described below. In the illustrations, each grid point represents an individual VCI matrix element, dark grey shading denotes elements explicitly selected or generated for inclusion in VCI sub-matrix, light grey points represent elements calculated during VPT2 screening which may be subsequently also included in the VCI sub-matrix, and a thick black line denotes a VCI sub-matrix to be diagonalized.

1. Generate the upper triangular elements of the square and symmetric VCI(1) sub-matrix, of leading dimension $N_{inc} = M + 1$, and diagonalize to obtain the initial VCI wavefunction for the ground and fundamental excited states.

2. Generate the remaining unique elements in the top $N_{inc} \times N_{total}$ block of the full VCI matrix, where $N_{total}$ = leading order of “parent” VCI matrix, according to the maximum excitation level = sum of quantum numbers of included basis states.

3. Select configurations to include in the VCI sub-matrix to be diagonalized based upon the VPT2 pair energy (8) exceeding a pre-set threshold for the ground state, fundamental excited states and states strongly resonant with fundamentals, noting that the full VCI matrix elements are also intermediates in the evaluation of the VPT2 pair energy, as $\langle \Psi_n | \Delta \hat{H} | \Phi_{n'} \rangle = \langle \Psi_n | \hat{H} | \Phi_{n'} \rangle$ if $n \neq n'$.

4. Add new configurations to VCI sub-matrix, $N_{inc} = N_{inc} + N_{selected}$, and diagonalize to update VCI wavefunction.

5. Repeat steps 3 and 4 until no new configurations are selected (micro-iterations).

6. Repeat steps 3, 4 and 5, decreasing the screening threshold, until fundamental VPT2-corrected VCI frequencies are converged (macro-iterations).
fundamentals in its VCI wavefunction larger than a pre-set threshold, here set to 0.15. When strong resonances occur, assignment of the fundamental becomes unclear and inclusion of resonant states in the VCI matrix during each VPT2 step ensures that all relevant basis functions are selected.

Sequentially lowering the VPT2 screening threshold ensures that the VCI wavefunction is iteratively improved so that it provides an appropriately accurate reference state for selecting new configurations for inclusion at each VPT2 step. Similarly, multiple configuration selection iterations are carried out at each VPT2 screening threshold, again to ensure that all relevant strongly coupled configurations - and no irrelevant ones - are included in the VCI matrix. This double iterative procedure is designed to construct the most concise VCI matrix that accurately describes the states of interest, hence minimizing memory requirements.

### III. METHODS

Our test set comprises 40 molecules containing between 4 and 10 atoms for which sextic force fields (SFFs) in normal mode coordinates are available\(^1\),\(^5\),\(^6\). These molecules are illustrated in supplemental Table S1.\(^2\) For smaller molecules (4-6 atoms), accurate SFFs of at least CCSD(T)/cc-pVTZ quality have been compiled from the literature\(^5\),\(^6\). For larger molecules (6-10 atoms), HF/6-311G** quality SFFs in normal mode coordinates have been generated via a coordinate transformation procedure from quartic force fields (QFFs) constructed in curvilinear internal coordinates.\(^1\) All force fields may be obtained through the extensible PyPES library.\(^3\)

Benchmark VCI results are generated by including in the VCI expansion all configurations with a specified sum of vibrational quantum numbers to a maximum of 6, excluding the Coriolis coupling term from the Hamiltonian. We have previously shown that VCI(6) predicts fundamental frequencies to well within 1 cm\(^{-1}\) of benchmark values, on average, across a chemically diverse data set.\(^4\) For 9 and 10 atom molecules, the time and memory requirements for the naïve VCI(6) algorithm become prohibitive, so we do not attempt to externally benchmark these systems.

With benchmark data for 4-8 atom molecules in hand, we turn our attention to exploring strategies for reducing computational cost.

The first, and simplest, strategy is to restrict the extent of mode coupling, i.e. by limiting the number of vibrational modes that may be concurrently excited when forming the VCI basis states. VPT2-based screening is then carried out using the algorithm described above, iterating over different choices of screening threshold from \(1 \times 10^{-3}\) to \(1 \times 10^{-7}\) E\(_h\), as listed in Table I.

To minimize memory requirements, all non-negligible (\(> 1 \times 10^{-8}\) E\(_h\)) VCI matrix elements comprising vibrational polyads\(^5\) are stored in sparse matrix format and the VCI sub-matrix is diagonalized using the implicitly restarted Lanczos diagonalization procedure implemented in ARPACK and accessed via wrapper routines implemented within SciPy. The amount of memory required also depends on the requested number of lowest eigenstates to be computed. This is estimated as the number of harmonic frequencies between 0 and the highest fundamental frequency plus a 200 cm\(^{-1}\) buffer. This approach ensures that the desired fundamental states are included, but limits the number of additional eigenvalues and eigenvectors.

To further reduce run time, a reduced mode representation of the potential may be used. In previous work, we have demonstrated that omitting force constants with more than 4 unique indices from the SFF does not substantially change the calculated fundamental frequencies, introducing errors of less than 1 cm\(^{-1}\), on average. We also use a threshold of 1 cm\(^{-1}\) to screen the force constants, introducing a further error of < 0.2 cm\(^{-1}\).\(^1\)

Although errors due to truncating the wavefunction expansion (VPT2 screening, reduced mode-coupling) and PES expansion (force constant screening) are expected to be only weakly correlated, we test this explicitly before recommending an overall screening procedure.

Computational efficiency is assessed by quantifying the number of states included in the VCI sub-matrix (\(N\_\text{inc}\)) and the number of eigenvalues calculated during diagonalization (\(N\_\text{eigs}\)), which determine CPU time and RAM requirements.

### IV. RESULTS & DISCUSSION

#### A. Reduced mode coupling in VCI expansion

Errors due to reducing the extent of mode coupling in the VCI wavefunction are calculated as:

\[
\Delta \nu = \nu_{\text{VCI}(6)} - \nu_{\text{VCI}(6,m\text{MC})}
\]

where \(m\) = maximum extent of mode coupling (MC) in the VCI(6) wavefunction. Errors are illustrated in Figure 1 and summarized in Table II. As expected, errors decrease as the extent of mode coupling increases.

Inspection of Figure 1 reveals that the error distribution also changes.

At lower mode coupling levels, the majority of errors are < 0, with approximate frequencies larger than the

<table>
<thead>
<tr>
<th>Table I: Screening threshold values (E(_h))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1 \times 10^{-3})</td>
</tr>
<tr>
<td>(5 \times 10^{-4})</td>
</tr>
<tr>
<td>(1 \times 10^{-4})</td>
</tr>
<tr>
<td>(8 \times 10^{-5})</td>
</tr>
<tr>
<td>(6 \times 10^{-5})</td>
</tr>
<tr>
<td>(4 \times 10^{-5})</td>
</tr>
</tbody>
</table>
**FIG. 1:** Errors in VCI(6) fundamental frequencies due to reduced mode-coupling in the wavefunction: a) 3-mode coupling; b) 4-mode coupling and c) 5-mode coupling, marked according to molecular size: 4 atoms (×), 5 atoms (△), 6 atoms (○), 7 atoms (+), 8 atoms (○).

**TABLE II:** Mean absolute errors (cm$^{-1}$) in fundamental vibrational frequencies as a function of molecule size, due to reduced mode coupling (MC) in the VCI(6) wavefunction.

<table>
<thead>
<tr>
<th>n$_{\text{atom}}$</th>
<th>3MC</th>
<th>4MC</th>
<th>5MC</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.5</td>
<td>0.04</td>
<td>0.00004</td>
</tr>
<tr>
<td>5</td>
<td>5.3</td>
<td>0.5</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>16.8</td>
<td>2.1</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>19.9</td>
<td>2.9</td>
<td>0.1</td>
</tr>
<tr>
<td>8</td>
<td>32.9</td>
<td>4.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

reference values. This indicates that the anharmonicity in each fundamental frequency has not been completely captured.

At 5 mode coupling, only the stretching frequencies are significantly in error. These errors are randomly distributed about 0, which suggests that they arise due to neglect of indirect coupling. This is confirmed by closer inspection of the coefficient vectors for these modes, which reveals that they are coupled to resonant states, i.e. states that are themselves coupled to one another. This results in an overall highly coupled state that is difficult to capture correctly within the reduced mode coupling approximation.

Larger molecules have larger errors, particularly at lower mode coupling levels. 3MC errors, in particular, are both large and strongly correlated with molecular size. Therefore, despite the significant reduction in computational cost that truncating the VCI wavefunction at 3 mode coupling affords, the deteriorating accuracy of results as a function of molecule size makes this approach inappropriate for larger molecules.

On the other hand, 5 mode coupling truncation of the wavefunction affords highly accurate fundamental frequencies with errors that do not grow with molecule size beyond $n_{\text{atom}} = 6$. However, the efficiency gains from using 5MC are modest, so do not significantly extend the limits of applicability beyond the molecules that may otherwise be modelled using full VCI(6).

Therefore, we overall recommend 4 mode coupling pre-screening of the VCI wavefunction for an acceptable balance between accuracy loss and efficiency gain. However, we do note that this can introduce up to 15 cm$^{-1}$ errors in fundamental frequencies in the worse case scenario of highly anharmonic, strongly coupled and resonant modes. This approximation also routinely introduces errors around 5 cm$^{-1}$, particularly for the larger molecules in our test set.

**B. VPT2 screening**

Mean absolute errors in VPT2-screened VCI frequencies, with an optional post-VCI VPT2 correction, are presented as a function of final screening threshold in Table
In both cases, errors in calculated fundamental frequencies decrease as the screening threshold decreases and more configurations are included in the VCI submatrix.

TABLE III: Mean absolute errors (cm\(^{-1}\)) in VPT2-screened VCI(6) fundamental frequencies, at a range of final screening thresholds, both with (column 3) and without (column 2) a subsequent VPT2 correction.

<table>
<thead>
<tr>
<th>Screening threshold (E(_h))</th>
<th>VCI(_{scr})(6) + VPT2</th>
<th>VCI(_{scr})(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1 \times 10^{-4}) E(_h)</td>
<td>65.5</td>
<td>10.1</td>
</tr>
<tr>
<td>(1 \times 10^{-5}) E(_h)</td>
<td>24.9</td>
<td>3.6</td>
</tr>
<tr>
<td>(1 \times 10^{-6}) E(_h)</td>
<td>10.2</td>
<td>1.3</td>
</tr>
<tr>
<td>(1 \times 10^{-7}) E(_h)</td>
<td>4.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

As the VPT2 correction provides substantially more accurate results for a minimal increase in computational cost, we recommend that it is always applied. Therefore, all subsequent VPT2-screened VCI results presented in this work will include the subsequent VPT2 correction. We note, in passing, that the VPT2 correction does not alter the rate of convergence with respect to screening threshold, but rather the associated pre-factor.

TABLE IV: Breakdown of mean absolute errors (cm\(^{-1}\)) in fundamental frequencies by molecule size, at VCI\(_{scr}\)(6)+VPT2, with a range of final screening threshold values.

<table>
<thead>
<tr>
<th>(n_{atom})</th>
<th>Screening threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(10^{-4}) E(_h)</td>
</tr>
<tr>
<td>4</td>
<td>2.7</td>
</tr>
<tr>
<td>5</td>
<td>3.4</td>
</tr>
<tr>
<td>6</td>
<td>3.8</td>
</tr>
<tr>
<td>7</td>
<td>3.8</td>
</tr>
<tr>
<td>8</td>
<td>3.7</td>
</tr>
</tbody>
</table>

The relationship between error and molecule size is one of the factors that will determine the scalability of our VPT2 screening approach. This may be assessed on the basis of the data illustrated in Figure 2, and associated summary statistics in Table IV.

Unlike the errors associated with a priori reduction in mode coupling, VPT2 screening errors are randomly distributed for all choices of screening threshold value. Further, for larger molecules (\(\geq 6\) atoms) errors are only weakly related to molecular size, particularly at the highest acceptable screening threshold of \(10^{-5}\) E\(_h\). This makes VPT2 screening a particularly attractive way of reducing both the time and memory demands of a full VCI(6) calculation.

However, it should be noted that maximum errors associated with applying the \(10^{-5}\) E\(_h\) screening threshold would be needed to achieve a given level of accuracy.
can be significant, with outliers in Figure 2 at up to 30 cm\(^{-1}\), which are due to states with strong resonances. It would be safer to use a screening threshold of \(10^{-6} E_h\), which predicts fundamental frequencies to within 15 cm\(^{-1}\) in all cases, in line with the maximum error associated with applying the 4 mode coupling approximation.

C. Reduced mode representation of the potential

We have previously shown that reducing the mode representation (MR) of potential energy surface expansions is an effective method for decreasing computational run time without significantly sacrificing accuracy. Here, we investigate the extent of coupling between wavefunction truncation and PES truncation errors.

TABLE V: Mean absolute errors (cm\(^{-1}\)) in fundamental frequencies as a function of molecule size, using a selection of truncated VCI(6) wavefunction expansions, with both full (6MR) and reduced (4MR) mode-representations of the sextic force field potential.

<table>
<thead>
<tr>
<th>(n_{\text{atom}})</th>
<th>VCI(_{\text{scr}}(6)+\text{VPT2, 10}^{-5} E_h \text{ threshold})</th>
<th>VCI(_{\text{scr}}(6)+\text{VPT2, 10}^{-6} E_h \text{ threshold})</th>
<th>VCI(6,4MC)</th>
<th>VCI(6,4MC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.7</td>
<td>2.7</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>3.2</td>
<td>3.4</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>3.9</td>
<td>3.8</td>
<td>1.4</td>
<td>1.5</td>
</tr>
<tr>
<td>7</td>
<td>3.2</td>
<td>3.8</td>
<td>1.3</td>
<td>1.4</td>
</tr>
<tr>
<td>8</td>
<td>3.9</td>
<td>3.7</td>
<td>1.9</td>
<td>1.9</td>
</tr>
</tbody>
</table>

The data presented in Table V confirm that the 4MR SFF provides an excellent approximation to the full SFF at much reduced computational cost. Further, we confirm that errors due to truncation of the force field expansion are not amplified by also truncating the wavefunction expansion.

D. Combined screening algorithm accuracy

The data presented in sections B and C allow us to make some clear and unambiguous choices that reduce computational cost with minimal accuracy loss. Throughout the remaining results and discussion, potential energy surfaces will always be represented as 4MR sextic force fields and post-hoc VPT2 corrections will always be applied to VPT2-screened VCI wavefunctions.

Larger errors arise due to omitting VCI matrix elements, whether via iterative VPT2 screening or reduced mode-coupling pre-screening. However, it is advantageous to employ both strategies to reduce memory and run time. Therefore, it is necessary to weigh up accuracy loss against efficiency gain.

Here, we quantify the errors associated with using both reduced mode coupling and VPT2 screening in tandem and, in the next section, establish the scaling laws that apply to computational time and memory requirements.

Errors in fundamental frequencies are illustrated in Figure 3 and summarized in columns 2 and 4 of Table VI. For clarity, reference statistics without pre-truncation of the VCI wavefunction (columns 3,5) or VPT2 screening (column 6) are reproduced in this Table.

TABLE VI: Mean absolute errors (cm\(^{-1}\)) in fundamental frequencies as a function of molecule size, at VCI\(_{\text{scr}}(6,6MC)+\text{VPT2, with both full (}m=6MC\) and reduced (\(m=4MC\)) mode-coupling in the wavefunction, using final screening thresholds of \(10^{-5}\) and \(10^{-6} E_h\). Unscreened VCI(6,4MC) data is also provided for reference.

<table>
<thead>
<tr>
<th>(n_{\text{atom}})</th>
<th>(10^{-5} E_h \text{ threshold})</th>
<th>(10^{-6} E_h \text{ threshold})</th>
<th>Unscreened</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.7</td>
<td>2.7</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>3.6</td>
<td>3.2</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>5.1</td>
<td>3.9</td>
<td>2.7</td>
</tr>
<tr>
<td>7</td>
<td>5.9</td>
<td>3.2</td>
<td>3.7</td>
</tr>
<tr>
<td>8</td>
<td>6.0</td>
<td>3.9</td>
<td>4.7</td>
</tr>
</tbody>
</table>

At a final VPT2 screening threshold of \(10^{-5} E_h\) with 4 mode coupling, VPT2 screening errors dominate for smaller molecules (up to 6 atoms), while errors due to VPT2 screening and reduced mode-coupling become cumulative for larger molecules. Overall, the error grows as a function of molecule size, reaching an average of 6 cm\(^{-1}\) across the fundamental frequencies of all 8 atom molecules. Inspection of Figure 3 reveals maximum errors of up to 30 cm\(^{-1}\), primarily arising from VPT2 screening.

Decreasing the screening threshold to \(10^{-6} E_h\) changes the error distribution substantially. At this screening threshold, errors in calculated fundamental frequencies are primarily due to the 4 mode coupling approximation rather than the VPT2 screening, as evident by comparing columns 4 and 6 of Table VI.

In both cases, the 4MC approximation is responsible for the component of the error that increases as a function of molecule size. Avoiding reduced mode coupling pre-screening of the VCI(6) wavefunction - columns 3 and 5 of Table VI - affords results that are substantially more accurate, with errors that are largely independent of molecular size.

Therefore, in the next section, we additionally quantify the computational cost associated with performing VCI\(_{\text{scr}}(6)+\text{VPT2 calculations without any pre-}

truncation of the wavefunction expansion, using a tight final VPT2 screening threshold of \(10^{-7}\).
E. Computational resource requirements

Empirical scaling laws for the number of configurations selected for inclusion in the VCI sub-matrix \((N_{inc})\) during the VPT2 screening procedure and number of eigenvalues calculated during diagonalization \((N_{eigs})\), both with and without reduced mode coupling pre-screening, are given in Table VII. The run time and memory required are derived from these fundamental quantities.

The run time scales with the total number of matrix elements to be evaluated,

\[
O(N_{inc}N_{states})
\]

where \(N_{states}\) is the total number of VCI basis states, which scales as \(O(N_{mode}^6)\) for VCI(6) and \(O(N_{mode}^4)\) for VCI(6,4MC).

The memory required for the implicitly restarted Lanczos diagonalization procedure, as implemented in ARPACK, scales as

\[
\max [O(N_{inc}N_{eigs}), O(N_{eigs}^2)]
\]

The data from which the scaling laws are derived are provided in supplemental Table S2.52

Table VII: Empirical scaling laws for the number of configurations selected for inclusion in the VCI sub-matrix \((N_{inc})\), number of eigenvalues calculated during diagonalization \((N_{eigs})\), run time, and memory required during diagonalization, both with and without \textit{a priori} reduction in mode coupling, and using loose, tight and very tight VPT2 screening thresholds of \(10^{-5}\) \(E_h\), \(10^{-6}\) \(E_h\) and \(10^{-7}\) \(E_h\), respectively.

\[
\begin{array}{|c|c|c|c|}
\hline
N_{inc} & 10^{-5} E_h & 10^{-6} E_h & 10^{-7} E_h \\
\hline
4MC & O(N_{mode}^2) & O(N_{mode}^2) & O(N_{mode}^3) \\
6MC & O(N_{mode}^2) & O(N_{mode}^2) & O(N_{mode}^3) \\
\hline
N_{eigs} & 10^{-5} E_h & 10^{-6} E_h & 10^{-7} E_h \\
\hline
4MC & O(N_{mode}^3) & O(N_{mode}^3) & O(N_{mode}^3) \\
6MC & O(N_{mode}^3) & O(N_{mode}^3) & O(N_{mode}^3) \\
\hline
\text{Run time} & 10^{-5} E_h & 10^{-6} E_h & 10^{-7} E_h \\
\hline
4MC & O(N_{mode}^6) & O(N_{mode}^6) & O(N_{mode}^7) \\
6MC & O(N_{mode}^8) & O(N_{mode}^8) & O(N_{mode}^9) \\
\hline
\text{Memory} & 10^{-5} E_h & 10^{-6} E_h & 10^{-7} E_h \\
\hline
4MC & O(N_{mode}^7) & O(N_{mode}^7) & O(N_{mode}^7) \\
6MC & O(N_{mode}^7) & O(N_{mode}^7) & O(N_{mode}^7) \\
\hline
\end{array}
\]

Although the errors in fundamental frequencies at \(V_{scr}(6,4MC)\)+VPT2 with a \(10^{-5}\) screening threshold (column 2, Table VI) push the upper limits of what may be considered acceptable, the major advantage of this approach is that both its run time and memory requirements scale as \(O(N_{mode}^6)\). The exact scaling behaviour is determined by the chemical nature of the molecule, and the extent of coupling between vibrational modes.

However, if more accurate results are desired, it is necessary to either increase the extent of mode-coupling or decrease the VPT2 screening threshold, or both. If runtime, but not memory, is the limiting factor, \(V_{scr}(6,4MC)\)+VPT2 with a tighter screening threshold of \(10^{-6}\) \(E_h\) is recommended. Conversely, if memory, but not runtime, is the limiting factor, \(V_{scr}(6)\)+VPT2

FIG. 3: Errors in VCI_{scr}(6,4MC)+VPT2 with different final screening thresholds: a) \(1 \times 10^{-5}\) \(E_h\); b) \(1 \times 10^{-6}\) \(E_h\). Data points are marked according to molecular size: 4 atoms (\(\times\)), 5 atoms (\(\Delta\)), 6 atoms (\(\bigcirc\)), 7 atoms (+), 8 atoms (\(\bigcirc\)).
with a looser screening threshold of $10^{-5}$ $E_h$ is more appropriate.

Finally, if computational resources are abundant, we recommend performing the VCI$_{scr}(6)$+VPT2 procedure iteratively over all screening thresholds until convergence to the desired level of accuracy is reached. Typically, this process will converge with errors in fundamentals of $<5$ cm$^{-1}$ by $10^{-6}$ $E_h$, but a tighter screening threshold is required for molecules with complicated resonance structures that exhibit mixing between overtones and fundamentals.

V. CONCLUSIONS

Vibrational configuration interaction is a simple and robust variational nuclear structure method that is easy to implement and parallelize. However, the need to include highly excited states in the wavefunction expansion even to describe fundamental frequencies results in poor scaling properties, and generally precludes its application to molecules containing more than five or six atoms, depending on molecular symmetry.

Therefore, we have tested a range of strategies to reduce the computational cost associated with VCI calculations for a collection of molecules containing up to 10 atoms. Truncating the sextic force field expansion at 4 mode representation gives substantial speed-up with minimal accuracy loss, whether screening procedures are applied to the VCI wavefunction or not. A priori reduction of mode coupling in the wavefunction also significantly reduces runtime, but introduces errors that grow as a function of molecule size. Iterative VPT2 screening of the VCI matrix decreases both runtime and memory requirements. Applying a subsequent VPT2 correction always improves the calculated fundamental frequencies at minimal additional computational cost.

Overall, we advocate using the VPT2-screened VCI procedure with post-hoc VPT2 correction that we denote VCI$_{scr}(6.4MC)$+VPT2, using a screening threshold of $10^{-5}$ $E_h$. This is the most efficient method that gives reasonable results, subject to the caveat that some highly anharmonic, highly coupled modes may incur substantial errors of up to 30 cm$^{-1}$, relative to benchmark VCI(6) results. For more accurate results with errors independent of molecular size, the VCI$_{scr}(6)$+VPT2 with a screening threshold of $10^{-6}$ $E_h$ is recommended. This procedure recovers fundamental frequencies with maximum errors $<5$ cm$^{-1}$ and average errors $<2$ cm$^{-1}$ for most modes of most molecules. A tighter screening threshold of $10^{-7}$ ensures that all fundamentals converge to within 5 cm$^{-1}$ of the unscreened VCI(6) results, even for problematic highly-coupled, highly-anharmonic cases.

Our open-source VPT2-screened VCI code has been released under the MIT licence, and may be freely downloaded from http://sourceforge.net/projects/pyvci-vpt2. We anticipate that the ability to accurately model vibrational spectra for a wider range of organic molecules than possible previously will stimulate further studies in this field and help solve applied problems in vibrational spectroscopy.

REFERENCES


52See supplemental material at [URL] for names and chemical structures of molecules in our test set, and computational scaling data.” .
