Using computed infrared intensities for the reduction of vibrational configuration interaction bases

Vincent Le Bris, Marc Odunlami, Didier Bégué, Isabelle Baraille, Olivier Coulaud

PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d0cp00593b

Using computed infrared intensities for the reduction of vibrational configuration interaction bases†

Vincent Le Bris, *[a] Marc Odunlami, [a] Didier Bégué, [a] Isabelle Baraille[a] and Olivier Coulaud[b]

Received 3rd February 2020,
Accepted 11th March 2020

DOI: 10.1039/d0cp00593b

rsc.li/pccp

The Adaptive Vibrational Configuration Interaction (A-VCI) algorithm is an iterative process that computes the anharmonic spectrum of a molecule using nested bases to discretize the Hamiltonian operator. For large molecular systems, the size of the discretization space and the computation time quickly become prohibitive. It is therefore necessary to develop new methods to further limit the number of basis functions. Most of the time, the interpretation of an experimental infrared spectrum does not require the calculation of all eigenvalues but only those corresponding to vibrational states with significant intensity. In this paper, a technique that uses infrared intensities is introduced to select a subset of eigenvalues to be precisely calculated. Thus, we build smaller nested bases and reduce both the memory footprint and the computational time. We validate the advantages of this new approach on a well-studied 7-atom molecular system (C_2H_4O), and we apply it on a larger 10-atom molecule ($C_4H_4N_2$).

1 Introduction

The computation of anharmonic vibrational spectra is a hot topic because of its usefulness as a tool for the interpretation of infrared spectra in fields as diverse as biochemistry, chemical reactivity, interstellar chemistry, complex matrix chemistry and material chemistry. The complexity of vibrational data, particularly in spectral regions with high density of states (mid-IR) or with very low active signals (near IR), makes the use of predictive modeling an essential support for the interpretation of experimental spectra. The main challenge concerns the ability to compute the spectrum of complex Hamiltonians in affordable computational times and with worthwhile accuracy. Complexity arises when the size of the molecular system under consideration becomes large, but also when relevant terms, such as Coriolis corrections, are added to the Hamiltonian. In these cases, new methods and algorithms are required.

The Adaptive Vibrational Configuration Interaction (A-VCI) algorithm[1,2] has been developed to effectively reduce the number of vibrational states used in the Configuration Interaction (CI)[3–6] process. It builds iteratively nested bases to discretize the Hamiltonian operator within a large CI approximation space by considering an *a posteriori* error estimator for the convergence of the method and to select the most relevant directions to expand the discretization space. Currently, the robustness and reliability of this method have been tested on molecules of 4 atoms (formaldehyde), 6 atoms (acetonitrile) and 7 atoms (ethylene oxide) for calculated vibrational spectra up to 3000 cm^{-1}. In order to be able to process larger systems, it is mandatory to control the memory footprint and the CPU time of the algorithm. The complexity depends on the size of the generated bases, the required accuracy and also on the number of eigenvalues searched for. For larger molecules, the size of the CI approximation space increases exponentially as well as the density of vibrational states, especially in high energy regions. These two points lead to extremely large basis sizes (and therefore matrices) making the calculation of eigenvalues difficult.

Traditionally, pruning techniques[7–13] have been used extensively to reduce the number of elements in the CI approximation space. The most commonly used pruning conditions are usually written as follows

$$\sum_{i=1}^{3N-6} g_i(n_i) \leq b, \tag{1}$$

where g_i is an arbitrary function. The choice of the value of b is the result of a trade-off between the size of the approximation basis and the accuracy required for the eigenvalues with respect to the Hamiltonian continuous spectrum. However, for molecules with more than 7 or 8 atoms, this method does not reduce the approximation space sufficiently to obtain satisfactory accuracies.

A complementary idea to be able to process larger molecules is to take into account information on vibrational states that are active in

[a] Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IPREM, Institut des Sciences Analytiques et de Physicochimie pour l'Environnement et les Matériaux, UMR5254, Pau, France. E-mail: vincent.lebris@univ-pau.fr
[b] HiePACS project-team, Inria Bordeaux Sud-Ouest, 200, avenue de la Vieille Tour, F-33405 Talence Cedex, France

† Electronic supplementary information (ESI) available: The PES, Coriolis and dipole moment coefficients used in this work for C_2H_4O and $C_4H_4N_2$. Detailed calculation results are also provided for the molecular systems considered in this work. See DOI: 10.1039/d0cp00593b

infrared or Raman spectroscopy. The mode-tracking technique[14,15] applies this strategy for the diagonalization of the Hessian matrix within the harmonic approximation. The Intensity-Carrying Modes (ICM) theory[16] also uses this idea. The latter method generates a small number of pseudo-modes with non-zero dipole derivatives in order to reduce the dimension of the problem. ICM theory is used as an efficient starting point[17] for the Intensity Tracking technique,[18] an iterative algorithm that selects only the spectroscopically active modes.

This paper is intended to be a proof of concept to demonstrate the feasibility of applying this idea within an anharmonic framework using the A-VCI formalism. Since the computational cost of A-VCI mainly depends on the size of the basis obtained at each iteration, the choice of the enlargement strategy is a key point to optimize the performance of the algorithm. Selection techniques, based on the residual vector components, offer a good trade-off between the memory footprint and the execution time.[2] Nevertheless, the behavior of these component-wise strategies is strongly connected to the accuracy required on the final eigenpairs. Indeed, the basis functions selected for a given threshold will not necessarily be selected when a better accuracy is required (*i.e.* for a lower value of the threshold). In addition, the expansion rate of the A-VCI adaptive basis increases with the frequency range of the spectrum under study, *i.e.* the number of eigenpairs required. In this work, we explore the idea that the information on IR intensities can significantly reduce the number of eigenvalues that need to be accurately calculated. Calculation of IR intensities beyond electrical harmonicity was implemented in the A-VCI algorithm to efficiently reduce the size of the problem without any loss of accuracy on a subset of eigenpairs. In addition, knowledge of IR intensities provides additional information that is useful for experimental interpretation of non-fundamental bands in experimental spectra. The calculation of IR intensities is already considered in the literature[19–21] and is based on a non-linear dependence between the dipole moment and the normal coordinates.

The context of the variational approach used in this paper is presented in Section 2.1. Then, a brief presentation of the A-VCI method is provided in Section 2.2.1. In Sections 2.2.2 to 2.2.5, new strategies to achieve faster convergence are presented: the energy pruning of the search space, a new accuracy-independent basis augmentation technique and the intensity selection of vibrational states. Section 2.3 presents how the Coriolis terms and the intensities are efficiently computed in the A-VCI framework. Finally, to highlight the interest of these new developments, we present numerical results for ethylene oxide (C_2H_4O), leading to a reduction of the basis size of almost 50%, and then, on the more challenging pyrazine molecule ($C_4H_4N_2$), for which it would have been impossible to obtain such precise results without intensity selection.

2 Formalism, algorithm overview and new developments

2.1 Context

The vibrational frequencies of a molecule are obtained by calculating the eigenvalues of the Hamiltonian operator $\mathscr{H}$. Given a N-atom molecular system with $D = 3N - 6$ degrees of freedom, we consider the vibrational Hamiltonian

$$\mathscr{H}(\mathbf{q}) = \mathscr{H}_0(\mathbf{q}) + \mathscr{V}(\mathbf{q}) + \mathscr{C}(\mathbf{q}), \tag{2}$$

with $\mathscr{H}_0(\mathbf{q}) = \sum_{i=1}^{D} \frac{\omega_i}{2}(p_i^2 + q_i^2)$ the harmonic operator, $\mathscr{V}(\mathbf{q})$ the anharmonic Potential Energy Surface (PES), $\mathscr{C}(\mathbf{q})$ the second order Coriolis correction, $\boldsymbol{q} = (q_1,q_2,\ldots,q_D)$ the normal dimensionless coordinates and the conjugate momentum $p_i = -\mathrm{i}\frac{\partial}{\partial q_i}$. The PES is a Taylor expansion of order S, which is a sum of products of monomials. The Coriolis contribution is written as

$$\mathscr{C}(\mathbf{q}) = \sum_{\alpha=x,y,z} B_\alpha \sum_{i,j\neq i,k,l\neq k} \zeta_{ij}^{\alpha}\zeta_{kl}^{\alpha} q_i p_j q_k p_l \sqrt{\frac{\omega_j\omega_l}{\omega_i\omega_k}}, \tag{3}$$

with ω_i the harmonic frequency associated with the q_i coordinate, B_α the rotational constant associated with the $\alpha = x,y,z$ axis (in cm^{-1}), and ζ_{kl}^{α} the Coriolis constant coupling q_k and q_l through rotation along the α axis, with $\zeta_{kl}^{\alpha} = -\zeta_{lk}^{\alpha}$.

Let Π be the space spanned by the eigenfunctions of the harmonic operator, $\mathscr{H}_0$. These eigenvectors, $\phi_{\mathbf{n}}^0$, write as the product of D one-dimensional Hermite functions of degrees $\mathbf{n} = \{n_1,n_2,\ldots,n_D\}$. Let $\mathbf{d} = \{d_1,d_2,\ldots,d_D\}$ the maximal degree of these functions. We define the approximation space $\Pi_{\mathbf{d}}$ as a subspace of Π by

$$\Pi_{\mathbf{d}} = \left\{ \varphi_{\mathbf{n}}^0 \middle/ \mathbf{n} \in \prod_{i=1}^{D} [0, d_i] \right\}.$$

The calculation of the spectrum of the Hamiltonian operator leads to the computation of the eigenvalues of the matrix H, the discretization of the operator $\mathscr{H}$ in $\Pi_{\mathbf{d}}$. In the framework of the variational approach, the coefficients of the H matrix are written as

$$H_{(i,j)} = \langle \phi_{\mathbf{i}}^0(\boldsymbol{q}) | (\mathscr{H}_0(\boldsymbol{q}) + \mathscr{V}(\boldsymbol{q}) + \mathscr{C}(\boldsymbol{q}))\phi_{\mathbf{j}}^0(\boldsymbol{q}) \rangle.$$

The calculation of the $\langle \phi_{\mathbf{i}}^0(\mathbf{q}) | \mathscr{C}(\boldsymbol{q})\phi_{\mathbf{j}}^0(\mathbf{q}) \rangle$ is given by the formulae in ref. 22 whereas the formulae for the integrals involving the harmonic operator and the anharmonic PES are given in our previous work.[2]

Let $(E_k,\mathbf{X}_k)$ be the kth eigenpair of the $m \times m$ matrix H. As shown in ref. 19 the intensity I_k between the vibrational state k and the ground state $(E_0,\mathbf{X}_0)$ is

$$I_k = \frac{8\pi^3 N_{\mathrm{A}}}{3hc(4\pi\varepsilon_0)}(E_k - E_0) \sum_{\alpha=x,y,z} |R_{\alpha,k}|^2,$$

where N_{A} is the Avogadro's number, $R_{\alpha,k}$ the transition dipole moment between the states 0 and k in the α-direction, E_0 and E_k the first and the $(k+1)$th eigenvalues. Converting the constant $\frac{8\pi^3 N_{\mathrm{A}}}{3hc(4\pi\varepsilon_0)}$ to standard units leads to the following formula

$$I_k = 16.194105 \times (E_k - E_0) \sum_{\alpha=x,y,z} \left|R_{\alpha,k}\right|^2 \ \mathrm{km\ mol}^{-1}, \tag{4}$$

with E_0 and E_k in cm^{-1}, and $R_{\alpha,k}$ in a.u. The transition dipole moment is written as

$$R_{\alpha,k} = \langle \mathbf{X}_0 | \mu_\alpha(\mathbf{q}) \mathbf{X}_k \rangle = \langle \mathbf{X}_k | \mu_\alpha(\mathbf{q}) \mathbf{X}_0 \rangle, \quad (5)$$

where $\boldsymbol{\mu} = (\mu_x, \mu_y, \mu_z)$ is the dipole operator. Each dipole moment surface μ_α is a Taylor expansion with respect to the D-dimensional $\mathbf{q}$ variable and is written as

$$\mu_\alpha(\mathbf{q}) = \sum_{\|\mathbf{s}\|_1 = 1}^{\tilde{S}} C_{\alpha,\mathbf{s}} \mathbf{q}^{\mathbf{s}},$$

where $||\cdot||_1$ is the usual 1-norm and $\tilde{S}$ is its maximal degree, which verifies $\tilde{S} \leq S - 1$, since $\boldsymbol{\mu}$ is the first derivative of the energy with respect to the electric field.

2.2 The A-VCI algorithm

2.2.1 Quick algorithm overview. The A-VCI algorithm[1,2] is an iterative procedure used to calculate the first F eigenpairs of the Hamiltonian, discretized in a very large space while guaranteeing the accuracy of the calculation. At the end of the computation we obtain a representation of the F smallest eigenpairs in a subset B of $\Pi_{\mathbf{d}}$.

The key point of the approach is the decomposition of the image by the operator $\mathscr{H}$ of the subspace $B^{(j)}$ at iteration j, noted $\mathscr{H}(B^{(j)})$, in the direct sum of two orthogonal spaces $B^{(j)}$ and $B_R^{(j)}$. To improve the readability of the notations, the subscript (j) will be omitted in the following.

We denote by $\tilde{H}$ the discrete representation $\mathscr{H}$ of in $\mathscr{H}(B)$, and by H (resp. H_R) the discrete representation of $\mathscr{H}$ in B (resp. B_R):

$$\tilde{H} = \begin{pmatrix} H & H_R^T \\ H_R & \cdots \end{pmatrix}$$

The estimate of the difference between an eigenpair $(E,\mathbf{X})$ of the operator discretized in B and the corresponding eigenvalue $\tilde{E}$ in $\mathscr{H}(B)$ is written

$$|E - \tilde{E}| \leq ||H_R \mathbf{X}||_2,$$

where $||\cdot||_2$ is the usual Euclidian norm. If $||H_R\mathbf{X}||_2$ is small, then E is a good approximation of the eigenvalue $\tilde{E}$ in the larger space $\mathscr{H}(B)$.

We briefly recall the main steps of the A-VCI algorithm. First, we define an initial basis B_0 belonging to the $\Pi_{\mathbf{d}}$ space. This basis contains at least the first $F + 1$ functions of $\Pi_{\mathbf{d}}$ sorted by ascending energies corresponding to the diagonal of the operator (2).

Then, we construct the sparse structures (*i.e.* only the row and column indices of the non-zero elements) of the matrices $H = B_0^T \mathscr{H}(B_0)$ and $H_R = B_0^T \mathscr{H}(B_0)$. During this step, we also build the set of admissible basis elements $B_R = \mathscr{H}(B) \backslash B$, which is needed to compute the sparse structure of the matrix H_R. The iterative procedure begins by calculating the terms of the Hamiltonian matrix, then the first F eigenpairs are computed by an iterative eigensolver.

To check the convergence of the algorithm, we evaluate the scaled residual vectors for all eigenpairs $(E_\ell, \boldsymbol{X}_\ell)$

$$\mathbf{r}_\ell = H_R \mathbf{X}_\ell / E_\ell. \quad (6)$$

If the maximum value of the norms $||\boldsymbol{r}_\ell||_2$ for $\ell = 1,\ldots,F$ is lower than a threshold ε, the method has converged. This evaluation requires the calculation of the non-zero elements of H_R involved in each component of the scaled residual vectors. If the convergence is not achieved, we build the new active space B in the next iteration by adding the elements selected from B_R. Finally, we update the sparse structures of the two matrices with the newly added basis elements.

When the convergence is reached, the eigenpairs are not only the eigenpairs of the Hamiltonian discretized in B, but also a good approximation of those of the Hamiltonian discretized in $B \oplus B_R$.

2.2.2 Reduction of the approximation space. The first way to reduce the size of the active space B is to reduce the size of the product space $\Pi_{\mathbf{d}}$. First, we set the following energetic criterion to approximately reach the same energy $E_{\max}$ in each direction of $\Pi_{\mathbf{d}}$:

$$d_i = 1 + \left\lfloor \frac{E_{\max}}{\omega_i} \right\rfloor. \quad (7)$$

Then, we introduce a global energetic pruning condition on the functions of $\Pi_{\mathbf{d}}$ to reduce its size even more. The resulting approximation space $P_{E_{\max}}$ is defined by

$$P_{E_{\max}} = \left\{ \varphi_{\mathbf{n}}^0 \in \Pi_{\mathbf{d}} \text{ such that } \sum_{i=1}^{D} \omega_i n_i \leq E_{\max} \right\}. \quad (8)$$

This idea is widely used to efficiently reduce the size of vibrational bases (see ref. 7, 11–13).

2.2.3 Collective component-wise strategy (CCW_ε). The performance of the method depends on how the active space increases at each iteration of the algorithm. In the previous paper,[2] several strategies have been introduced to limit the number of terms to be added at each iteration while ensuring the decrease of the residual norm. We now briefly recall the $\mathrm{CCW}_\varepsilon(p)$ approach used in ref. 2.

At any given iteration of the A-VCI algorithm, let K be the set of the non-converged eigenpairs $(E_\ell, \mathbf{X}_\ell)$, *i.e.* $K = \{\ell \in \{1,\ldots,F\}$ with $||\mathbf{r}_\ell||_2 > \varepsilon\}$, where $\mathbf{r}_\ell$ is the scaled residual vectors defined in (6). We introduce the mean residual vector $\mathbf{R}$ such that each component is

$$R_i = \frac{1}{k} \sum_{\ell \in K} |(\mathbf{r}_\ell)_i| \quad \text{with} \quad i = 1, \ldots, m_R,$$

where m_R (resp. k) is the size of space B_R (resp. K). The main objective is to identify the significant components in $\mathbf{R}$ to determine the elements to be added to B. The way in which the elements are selected will result in a trade-off between the speed of convergence and the basis enlargement. We denote by

$$M = \left\{ i \in \{1, \ldots, m_R\} \text{ such that } |R_i| > \frac{\varepsilon}{\sqrt{m_R}} \right\}$$

the set of indices of admissible elements of B_R and we define the generalized mean with respect to M by

$$\mathrm{mean}(\varepsilon, p) = \sqrt[p]{\frac{1}{\mathrm{card}(M)} \sum_{i \in M} |R_i|^p}.$$

Finally, the subset A of elements to add is defined by

$$A = \{\mathbf{n}_i \text{ such that } \boldsymbol{n}_i^{\mathrm{T}}\boldsymbol{R} > \mathrm{mean}(\varepsilon,p)\}.$$

One difficulty is to choose p in order to obtain the best trade-off between the growth of the basis size and the convergence rate. We refer the reader to our previous work[2] for more details on this strategy.

2.2.4 Accuracy independent strategy (CCW). The main drawback of the previous strategy is that the final bases are not necessarily nested when the ε parameter decreases. However, for large systems, it is necessary to be able to decrease progressively ε in order to reach the best accuracy computationally achievable. Indeed, in the CCW_ε approach described in Section 2.2.3, the mean residual vector $\boldsymbol{R}$ is based on eigenpairs that have not yet converged and they can be different depending on the value of ε. In addition, ε is also involved in the computation of M, and thus in mean(ε,p). To overcome this problem, we start by modifying the definition of the mean residual vector by taking into account all the eigenpairs

$$R_i = \frac{1}{F}\sum_{\ell=1}^{F} |(\mathbf{r}_\ell)_i| \text{ with } i = 1,\ldots,m_R.$$

We consider the generalized mean which is now independent of ε

$$\mathrm{mean}\ (p) = \sqrt[p]{\frac{1}{m_R}\sum_{i=1}^{m_R} |R_i|^p}.$$

Finally, in this strategy called CCW(p), the elements to add to B are selected as follows

$$A = \{\boldsymbol{n}_i \text{ such that } \boldsymbol{n}_i^{T}\boldsymbol{R} > \mathrm{mean}(p)\}.$$

The ε parameter is now used only as a convergence criterion to stop the iterative procedure. This specificity allows the successive use of the A-VCI method as an ideal restart procedure to gradually achieve greater accuracy. As in our previous strategies,[1,2] the choice of p is crucial to obtain an optimal trade-off between the number of iterations to reach the convergence, and the number of basis elements added at each iteration.

2.2.5 Selection strategy based on intensities (SI). The CCW_ε and CCW approaches are used to select basis elements from B_R, thus restricting the number of vibrational states to be added at each iteration. However, these approaches take into account every eigenpair. The idea here is to more drastically restrict the number of selected states by considering only a subset of eigenvalues that are of interest from an experimental point of view. By selecting only the states with significant intensity, only the states useful to decrease the components of the residual vector corresponding to this subset will be added to B.

We introduce the ε_I threshold and define the subset of eigenpairs we wish to consider by

$$K_{\varepsilon I} = \{(E_\ell,\mathbf{X}_\ell),\ \text{such that } I_\ell > \varepsilon_I \quad \ell = 1,\ldots F\},$$

where I_ℓ is the intensity of eigenpair $(E_\ell,\mathbf{X}_\ell)$ defined in (4). We construct the scaled residue (6) only for the eigenpairs of $K_{\varepsilon I}$ to check whether they have converged, and then we apply the selection strategy to that subset only.

In eqn (5), the computation of I_ℓ for every eigenpair $(E_\ell,\mathbf{X}_\ell)$ depends on the first eigenvector $\mathbf{X}_0$. Since our process is iterative, it is necessary to have a good approximation of the first eigenvector $\mathbf{X}_0$ to expect a good approximation of I_k. Therefore, we must ensure the convergence of the first eigenpair before using the selection strategy based on intensities.

This new strategy, called Selection by Intensity (SI), has two main steps: in the first step, the algorithm uses the traditional approach (CCW_ε or CCW) on the first F eigenpairs to enrich the basis until the convergence of the first eigenvector $\mathbf{X}_0$ is reached. In the second step, the algorithm continues only for the states with an intensity greater than ε_I.

2.3 Algorithmic details

2.3.1 Sparse structure of the Coriolis matrix. Due to the polynomial form of the PES part of the operator and the properties of the Hermite functions, it is possible to efficiently calculate the sparse structure of the Hamiltonian matrix, which is the most time-consuming step.

However, in the Coriolis contribution, first and second derivatives terms prevent us from directly calculating the corresponding sparse structure in the same way as for the PES. To do this, it is possible to define a pseudo-surface in polynomial form which will contain the sparse structure corresponding to the Coriolis couplings. For a given basis function $\phi^0_{\mathbf{n}}$, the subset of basis functions that are connected to it by a single Coriolis term is

$$C_{i,j,k,l}(\mathbf{n}) = \{\phi^0_{\mathbf{m}}/q_i p_j q_k p_l \phi^0_{\mathbf{n}} | \phi^0_{\mathbf{m}} \neq 0\}.$$

The full subset of basis functions that are connected to $\phi^0_{\mathbf{n}}$ by the Coriolis couplings is

$$C_{\mathrm{Cor}}(\mathbf{n}) = \bigcup_{i,j\neq i,k,l\neq k} C_{i,j,k,l}(\mathbf{n}).$$

Let us consider two Hermite functions ψ^0_n and ψ^0_m related to a normal dimensionless coordinate q and its corresponding conjugate momentum p. Thanks to the properties of the Hermite functions, we have

$$\{\psi^0_m / q\psi^0_n | \psi^0_m \neq 0\} = \{\psi^0_m / \langle p\psi^0_n | \psi^0_m \neq 0\},$$

and

$$\begin{aligned}\left\{\psi^0_m / \left\langle q^2\psi^0_n \mid \psi^0_n\right\rangle \neq 0\right\} &= \left\{\psi^0_m / \left\langle qp\psi^0_n \mid \psi^0_m\right\rangle \neq 0\right\} \\ &= \left\{\psi^0_m / \left\langle pq\psi^0_n \mid \psi^0_m\right\rangle \neq 0\right\} \\ &= \left\{\psi^0_m / \left\langle p^2\psi^0_n \mid \psi^0_m\right\rangle \neq 0\right\}.\end{aligned}$$

This shows that $q\psi^0_n|\psi^0_m$ and $p\psi^0_n|\psi^0_m$ are non-zero for the same values of n and m. This is also true for integrals with q^2, qp, pq and p^2. Each Coriolis term $\zeta^{\alpha}_{ij}\zeta^{\alpha}_{kl}q_i p_j q_k p_l$ involves at most four variables with indices i,j,k,l, where $1 \leq i,j,k,l \leq D$. Since $j \neq i$ and $l \neq k$, the number of times each index appears is at most 2, so that these terms only involves products of type q, p, q^2, qp, pq and p^2. Consequently, we have the following proposition.

Proposition 1 Consider the normal dimensionless coordinate $\mathbf{q} = (q_h)_{h=1,\ldots,D}$, the multi-index (i,j,k,l), where $j \neq i$ and $l \neq k$, and $\tilde{s}_h(i,j,k,l)$ is the number of times the index h appears in the multi-index (i,j,k,l). Then,

$$C_{i,j,k,l}(\mathbf{n}) = \left\{ \varphi^0_{\mathbf{m}} \middle/ \left\langle \prod_{h=1}^{D} q_h^{\tilde{s}_h(i,j,k,l)} \varphi^0_{\mathbf{n}} \middle| \varphi^0_{\mathbf{m}} \right\rangle \neq 0 \right\}.$$

We call Coriolis Pseudo-Surface (CPS) the polynomial form consisting of the monomials $\prod_{h=1}^{D} q_h^{\tilde{s}_h(i,j,k,l)}$. The full subset of basis functions that are connected to $\phi^0_{\mathbf{n}}$ by this Coriolis Pseudo-Surface is

$$C_{\mathrm{CPS}}(\mathbf{n}) = \bigcup_{i,j\neq i,k,l\neq k} \left\{ \varphi^0_{\mathbf{m}} \middle/ \left\langle \prod_{h=1}^{D} q_h^{\tilde{s}_h(i,j,k,l)} \varphi^0_{\mathbf{n}} | \varphi^0_{\mathbf{m}} \right\rangle \neq 0 \right\}.$$

Thanks to Prop. 1, it is easy to see that the two subsets $C_{\mathrm{CPS}}(\mathbf{n})$ and $C_{\mathrm{Cor}}(\mathbf{n})$ are identical. As a result, we can now quickly determine the sparse structure resulting from Coriolis couplings in the same way as for the PES.

2.3.2 Fast evaluation of the intensity. Calculating the intensities (4) requires a fast evaluation of the transition dipole moment (5), which is written in matrix form $R_{\alpha,\ell} = \mathbf{X}_\ell^T M^\alpha \mathbf{X}_0$, where $M^\alpha_{(i,j)} = \langle \phi^0_{\mathbf{i}} | \mu_\alpha \phi^0_{\mathbf{j}} \rangle$. Since only the non-zero elements of M^α are involved in the matrix-product $M^\alpha \mathbf{X}_0$, we have to consider the sparse structure of each matrix M^x, M^y and M^z.

From a computational point of view, the idea is to build a single sparse structure in order to save memory. Let $\mathscr{S}$ be the set of multi-indices $\mathbf{s}$ such that there is at least one α (x, y, or z) with a non-zero coefficient $C_{\alpha,\mathbf{s}}$ in μ^α (*i.e.* $\mathscr{S} = \{\mathbf{s} / C_{x,\mathbf{s}} \neq 0$ or $C_{y,\mathbf{s}} \neq 0$ or $C_{z,\mathbf{s}} \neq 0\}$).

Alogorithm 1: Intensity evaluation algorithm
Data: $(E_\ell, \mathbf{X}_\ell)_{\ell=0}^{F-1}$ the eigenpairs
Result: I_ℓ, the intensity vector for the eigenvalues E_ℓ when $\ell > 0$
Build the set $\mathscr{S}$;
Construct on the fly the vectors $\mathbf{Y}_0^\alpha = M^\alpha \mathbf{X}_0$ for $\alpha = x,y,z$; for $\ell = 1$ to $F - 1$ do
  Evaluate: $R_\alpha = \mathbf{X}_\ell^T \mathbf{Y}_0$ for $\alpha = x,y,z$;
  Build: $I_\ell = C_I(E_\ell - E_0)(R_x{}^2 + R_y{}^2 + R_z{}^2)$;

Algorithm 1 explains how the intensities are computed. Once the set $\mathscr{S}$ is built, we evaluate on the fly three sparse matrix-vector products (line 2) to obtain $\mathbf{Y}_0^\alpha$ from $\mathbf{X}_0$. Finally, for each calculated eigenvalue, we construct the three dipole transition moments (line 3) and add them together to obtain the intensity *via* the formula (4).

3 Results and discussion

The aim of this section is to establish a proof-of-concept for the intensity screening strategy with two test molecules. We use the potential energy surfaces (PES) already available in the literature and considered as references by several authors (see ESI† for the force constants). It is not intended to obtain the most accurate results to reproduce the experimental data. The idea is to show that the new options of the A-VCI method allow excellent control of the error coming from the discretization of the Hamiltonian, for which the discrete spectrum is well known. Moreover, the screening strategy introduced in this paper is intended for a finer control on a subset of eigenvalues of interest.

The first computations aim to validate the concepts introduced in the previous section on ethylene oxide (C_2H_4O), using a PES already published in previous works.[2,23,24] This PES was approximated by a Taylor series in normal dimensionless coordinates at the CC/B3//cc-pVTZ (4th order) level of calculation. The dipole moment and Coriolis coefficients were calculated at a B3LYP//6-31+G(d,p)[25] level of calculation. Finally, the calculations on pyrazine ($C_4H_4N_2$) will demonstrate the feasibility and efficiency of our latest developments to calculate the spectrum of a 10-atom molecular system.

3.1 Numerical environment

The C++ code takes advantage of the OpenMP shared-memory paradigm for parallelization. We consider ARPACK[26] to solve the eigenvalue problem at each iteration. The number of Arnoldi vectors generated is set to $(2F + 1)$, and the convergence criterion to 10^{-16}. Due to the large dimensions (≥ 15) of the calculations presented in this section, we use the 64 bits interface of LAPACK/ARPACK for integers. Moreover, we do not store in memory the coefficients of the H_R matrix[1,2] so the scaled residues (6) are computed on the fly to limit the memory footprint.

Extensive testing was conducted on this computer: 64-core Intel Xeon Gold SKL-6130 node running at 2.1 GHz with 3TB of shared memory. The Intel compiler (2019.3.199) is used with the following options: -O3 -qopenmp.

3.2 Influence of the Coriolis contributions

Fig. 1 reports the energy deviations on computed eigenvalues for ethylene oxide. The influence of off-diagonal Coriolis terms is very small compared to the effect of the diagonal terms (all eigenvalues, frequencies, and assignments are reported in the ESI†). This is con-

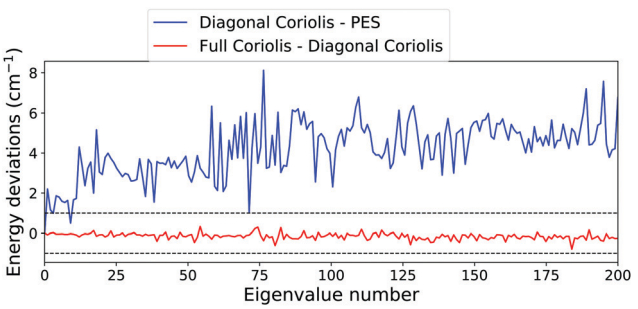

Fig. 1 Energy deviations (cm^{-1}) on computed eigenvalues for C_2H_4O. Differences between the operator with Coriolis diagonal terms and PES terms only are in blue (—). Differences between the operator with every Coriolis terms and diagonal Coriolis terms only are in red (—). Calculations are performed with $F = 200$, $p = 8$, $\varepsilon = 0.005$, and $E_{\max} = 25\,000$ cm^{-1}.

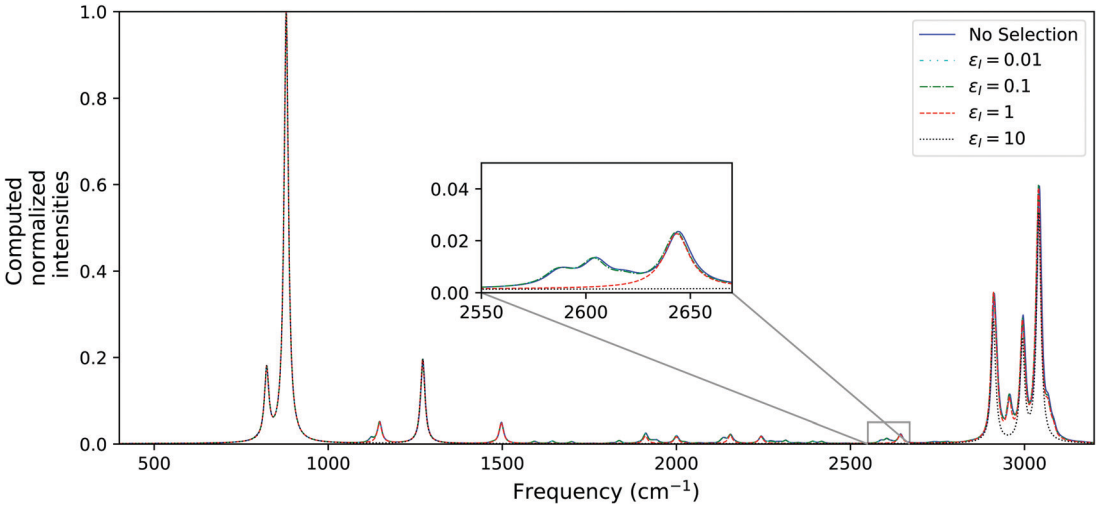

Fig. 2 Influence of the ε_I (km mol^{-1}) parameter using the SI strategy for C_2H_4O. The bandshapes have been created using a Lorentzian profile with a width at half height of 15 cm^{-1} and normalized.

sistent with the results reported in the literature[27,28] and these terms should only be added when high accuracy is needed (<1 cm^{-1}). Since the deviations caused by non-diagonal terms are lower than 1 cm^{-1}, only diagonal terms will be used in the following computations.

3.3 Screening by intensity

In this section, we provide further numerical experiments to assess the influence of the parameter ε_I in the new selection strategy. In Fig. 2, we plotted the numerical spectra for different values of ε_I with $F = 200$, $p = 8$, $E_{max} = 25\,000$ cm^{-1} and $\varepsilon = 0.005$ compared to a reference calculation without selection. These spectra are obtained from numerical values using a Lorentzian profile[29] with a width at half height of 15 cm^{-1}. These lineshapes are centered on the frequencies with an intensity above ε_I, and the peak heights correspond to a normalization of the molar absorption coefficient based on the predicted intensities (see ESI† for the numerical results).

For $\varepsilon_I = 1.0$ km mol^{-1}, and especially for $\varepsilon_I = 10.0$ km mol^{-1}, Fig. 2 shows that several frequencies were not taken into account by the selection algorithm. This absence represents a substantial change to the final profile. However, for $\varepsilon_I = 0.1$ km mol^{-1}, the selection has not significantly affected the profile compared to the one obtained without the selection. For example, in the 2600 cm^{-1} region, the $\varepsilon_I = 10.0$ km mol^{-1} computation misses all the bands. The $\varepsilon_I = 1.0$ km mol^{-1} computation succeeds in recovering one band, whereas the $\varepsilon_I = 0.1$ km mol^{-1} calculation recovers most of the significant information. This leads us to conclude that the $\varepsilon_I = 0.1$ km mol^{-1} threshold is a good trade-off between efficiency and accuracy for a qualitative study of the spectrum.

Let us first compare in Table 1 the efficiency of the SI strategy for an accuracy of 0.005 on the 200 first eigenpairs. As the number of frequencies is reduced by the selection algorithm, we obtain smaller basis sizes and execution times.

It should be mentioned that the number of frequencies with an intensity above a given threshold ε_I (*i.e.* selected by the algorithm) is sightly different from the number of frequencies with an intensity above the same threshold when no selection is made. The intensities

Table 1 Intensity screening for C_2H_4O. The parameters are $\varepsilon = 0.005$, $F = 200$, $p = 8$ and $E_{max} = 25\,000$ cm^{-1}

ε_I (km mol^{-1})	Selected eigenvalues	Final basis size	#(BR)	Total time (s)	Number of iterations
10.0	7	384 604	24 508 221	468	10
1.0	24	985 134	44 411 833	1589	11
0.1	71	1 797 093	60 844 318	3211	11
0.01	111	2 463 607	71 571 482	4918	11
0.0	200	3 473 266	91 128 462	7139	11

corresponding to these missing frequencies have the same order of magnitude as the threshold ε_I. When the selection begins (second step of the algorithm), the corresponding eigenvectors are not perfectly represented in the basis and lead to an inaccurate approximation of their intensities. Since the values of the resulting intensities are near ε_I, but slightly lower, the corresponding states are not selected by the algorithm. For these particular states, the added information thereafter does not entail the convergence of their eigenvectors. However, this side effect is not of major importance since it vanishes by reducing the ε_I parameter. For example, with $\varepsilon_I = 0.1$ km mol^{-1}, the final basis size and the total time are almost

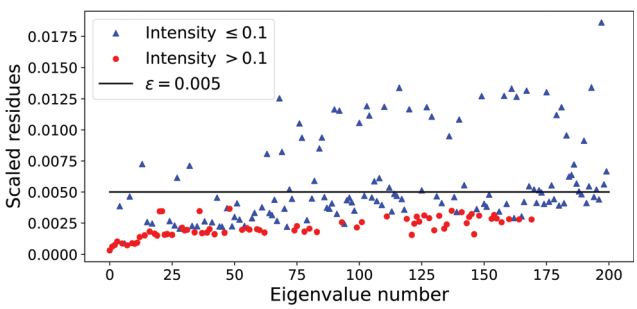

Fig. 3 Scaled residues using the SI strategy ($\varepsilon_I = 0.1$ km mol^{-1}). Calculations are done for C_2H_4O with the following parameters: $F = 200$, $p = 8$, $\varepsilon = 0.005$ and $E_{max} = 25\,000$ cm^{-1}.

Table 2 Summary of the different computations for $C_4H_4N_2$. The parameters are F = 2400, p = 8 and E_{max} = 22 000 cm^{-1}

ε	Selected eigenvalues	Final basis size	#(BR)	Total time (s)	Eigensolver time (s)
0.05	62	55 315	38 720 120	10 284	9992
0.04	60	108 733	68 315 887	22 821	22 159
0.03	57	190 490	106 409 050	56 283	54 873
0.02	57	725 390	301 805 985	157 593	151 914

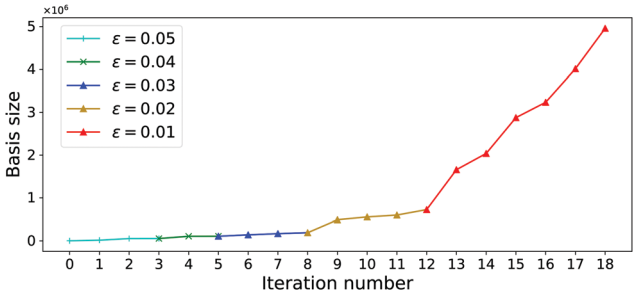

Fig. 4 Evolution of the basis size using the SI strategy (ε_I = 0.1 km mol^{-1}) for $C_4H_4N_2$ with the following parameters: F = 2400, p = 8 and E_{max} = 22 000 cm^{-1}.

halved without any significant qualitative impact on the spectrum profile.

Fig. 3 presents the scaled residues for every eigenvalue at the convergence of the algorithm. The horizontal line represents the selected value of ε = 0.005. The red dots correspond to the 71 eigenvalues with an intensity higher than ε_I = 0.1 km mol^{-1}, while the blue triangles are for those with an intensity lower than ε_I. The first eigenpair (E_0,$\mathbf{X}_0$) converges after 5 iterations and the full run takes 1 h 54 min to converge in 11 iterations with 32 cores. At the convergence, only 54 eigenvalues have a residue greater than ε = 0.005. This means that 75 eigenvalues with an intensity lesser than 0.1 km mol^{-1} have also converged even though they were not selected by the algorithm. The final basis size is 1 797 093. In summary, the new intensity-based selection strategy allows us to reliably address this problem using less than half the computational resources needed. The results of this calculation are compared with the experimental data in the ESI.†

3.4 Application to a 10-atom molecular system

To push the limits of the method, we consider a larger molecular system. The harmonic coefficients of the PES generated were calculated at the CCSD(T)//cc-pVTZ level. The anharmonic part of the potential (composed of 367 cubic terms and 666 quartic terms), the Coriolis coefficients, and the dipole moment second order terms were all computed at the B3LYP//6-31+G(d,p) level. We consider an intensity criterion ε_I set to 0.1 km mol^{-1} which appears as a good trade-off between computational time, accuracy on the eigenvalues, and relevant experimental data.

For this challenging computation, 2400 eigenvalues are needed to reach the 3000 cm^{-1} region. The strategy involves lowering ε gradually from 0.05 to 0.02, using the previously computed basis as a starting point for the next calculation. On a side note, the first step of the screening algorithm (*i.e.* the convergence of the first eigenpair) only occur during the first run (ε = 0.05). Table 2 shows details about each computation. It is important to note that the eigensolver always represents about 97% of the total CPU time. This is the major limitation for large sized systems due to the high density of vibrational states in the high energy region of the spectrum.

The last run (with ε = 0.01) fails to complete in a reasonable time (<30 days). The 6 iterations that went well took 26 days to complete (with 6 days for the last one to go from a matrix of 4 015 680 to 4 960 271 elements). The shared memory version of ARPACK is unable to compute 2400 eigenvalues for a matrix of this size in the allotted time. Only 7 eigenvalues have not converged among the 57 with an intensity above the threshold ε_I = 0.1 km mol^{-1}.

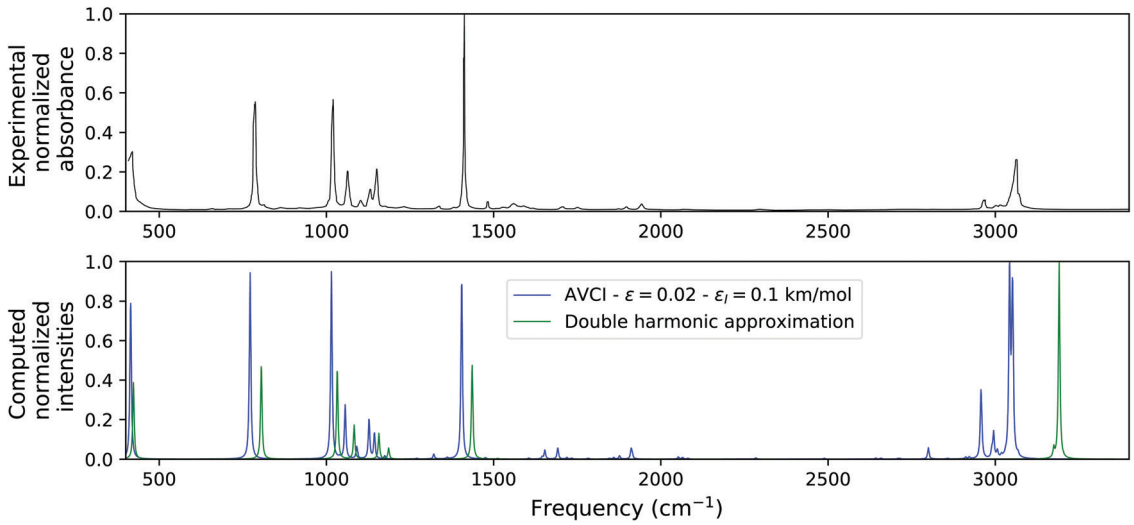

Fig. 5 Comparison between experimental data for $C_4H_4N_2$ and an A-VCI computation using SI (ε_I = 0.1 km mol^{-1}). The parameters are F = 2400, p = 8, E_{max} = 22 000 cm^{-1} and ε = 0.02. The bandshapes have been created using a Lorentzian profile with a width at half height of 15 cm^{-1} and normalized.

Table 3 A-VCI calculation for the first F = 2400 eigenvalues of $C_4H_4N_2$ with the SI method (ε_I = 0.1 km mol^{-1}). The parameters are p = 8, E_{max} = 22 000 cm^{-1} and ε = 0.02. Only the eigenpairs selected by the SI method are provided. Their position number in the full spectrum is given in parentheses, as well as the eigenvector coefficients used to make the attributions

Number	Frequency (cm^{-1})	Intensity (km mol^{-1})	Assignment
1(2)	414.64	26.45	ω_2(0.97), $\omega_2 + \omega_{10}$(0.16)
2(8)	771.76	31.30	ω_6(0.96), $\omega_6 + \omega_{10}$(0.17)
3(17)	1014.97	31.16	ω_{11}(0.96), $\omega_{10} + \omega_{11}$(0.19)
4(18)	1039.31	0.20	$\omega_1 + \omega_4$(0.97), $\omega_1 + \omega_4 + \omega_{10}$(0.16)
5(19)	1056.12	8.89	ω_{12}(0.9), $\omega_1 + \omega_5$(0.33)
6(21)	1090.84	2.02	$\omega_1 + \omega_5$(0.9), ω_{12}(0.33)
7(24)	1127.42	6.62	ω_{13}(0.9), $\omega_2 + \omega_5$(0.32)
8(25)	1143.61	4.24	ω_{14}(0.93), $\omega_{10} + \omega_{14}$(0.25)
9(27)	1174.44	0.51	$\omega_2 + \omega_5$(0.91), ω_{13}(0.32)
10(34)	1269.78	0.12	$\omega_1 + \omega_8$(0.92), $\omega_2 + \omega_7$(0.16)
11(37)	1321.10	0.82	$\omega_2 + \omega_7$(0.93), $\omega_2 + \omega_7 + \omega_{10}$(0.15)
12(44)	1360.87	0.22	$\omega_2 + \omega_8$(0.95), $\omega_2 + \omega_8 + \omega_{10}$(0.16)
13(49)	1404.53	30.08	ω_{17}(0.94), $\omega_{10} + \omega_{17}$(0.2)
14(60)	1475.52	0.21	ω_{18}(0.95), $\omega_{10} + \omega_{18}$(0.2)
15(86)	1605.21	0.16	$\omega_3 + \omega_{11}$(0.95), $\omega_3 + \omega_{10} + \omega_{11}$(0.19)
16(92)	1645.30	0.35	$\omega_3 + \omega_{12}$(0.73), $\omega_6 + \omega_7$(0.52)
17(94)	1653.14	1.49	$\omega_6 + \omega_7$(0.69), $\omega_3 + \omega_{12}$(0.49)
18(108)	1691.72	1.85	$\omega_6 + \omega_8$(0.86), $\omega_7 + \omega_9$(0.26)
19(113)	1719.04	0.27	$\omega_5 + \omega_9$(0.92), $\omega_6 + \omega_7$(0.22)
20(119)	1732.92	0.17	$\omega_3 + \omega_{14}$(0.92), $\omega_3 + \omega_{10} + \omega_{14}$(0.25)
21(137)	1782.01	0.14	$\omega_6 + \omega_{10}$(0.91), $\omega_6 + 2\omega_{10}$(0.26)
22(157)	1845.69	0.12	$\omega_4 + \omega_{14}$(0.93), $\omega_4 + \omega_{10} + \omega_{14}$(0.25)
23(166)	1858.86	0.26	$\omega_1 + \omega_{19}$(0.94), $\omega_1 + \omega_{10} + \omega_{19}$(0.2)
24(169)	1876.49	0.55	$\omega_7 + \omega_9$(0.89), $\omega_6 + \omega_8$(0.29)
25(184)	1911.19	1.56	$\omega_8 + \omega_9$(0.73), $3\omega_1 + \omega_8$(0.41)
26(186)	1914.25	0.72	$3\omega_1 + \omega_8$(0.66), $\omega_8 + \omega_9$(0.52)
27(259)	2052.19	0.36	$\omega_7 + \omega_{14}$(0.9), $\omega_7 + \omega_{10} + \omega_{14}$(0.24)
28(266)	2064.78	0.27	$\omega_{10} + \omega_{12}$(0.84), $\omega_1 + \omega_5 + \omega_{10}$(0.3)
29(277)	2080.72	0.13	$\omega_8 + \omega_{13}$(0.9), $\omega_2 + \omega_5 + \omega_8$(0.29)
30(453)	2283.81	0.19	$\omega_{12} + \omega_{15}$(0.83), $\omega_1 + \omega_5 + \omega_{15}$(0.32)
31(702)	2488.56	0.14	$\omega_9 + \omega_{19}$(0.89), $\omega_9 + \omega_{10} + \omega_{19}$(0.22)
32(991)	2642.48	0.16	$\omega_{13} + \omega_{19}$(0.87), $\omega_2 + \omega_5 + \omega_{19}$(0.26)
33(1023)	2657.88	0.15	$\omega_{14} + \omega_{19}$(0.77), $\omega_1 + 2\omega_2 + \omega_4 + \omega_5$(0.36)
34(1143)	2708.17	0.10	$\omega_1 + 2\omega_6 + \omega_8$(0.46), $\omega_1 + \omega_3 + \omega_6 + \omega_{11}$(0.36)
35(1162)	2714.71	0.11	$\omega_{14} + \omega_{20}$(0.76), $\omega_{10} + \omega_{14} + \omega_{20}$(0.24)
36(1381)	2799.75	1.90	$\omega_{16} + \omega_{18}$(0.86), ω_{23}(0.19)
37(1558)	2856.98	0.12	$\omega_2 + \omega_9 + \omega_{18}$(0.87), $\omega_1 + \omega_2 + \omega_3 + \omega_{18}$(0.19)
38(1742)	2911.31	0.29	$\omega_1 + \omega_5 + 2\omega_7$(0.47), $\omega_1 + \omega_6 + \omega_7 + \omega_9$(0.42)
39(1794)	2921.67	0.32	$\omega_{17} + \omega_{19}$(0.66), $2\omega_8 + \omega_{11}$(0.5)
40(1918)	2956.39	0.54	$\omega_1 + \omega_2 + 2\omega_3 + \omega_{11}$(0.85), $3\omega_1 + \omega_2 + \omega_3 + \omega_{11}$(0.28)
41(1922)	2956.98	9.36	$\omega_{17} + \omega_{20}$(0.64), $\omega_3 + \omega_{14} + \omega_{15}$(0.35)
42(1929)	2959.09	2.66	$\omega_3 + \omega_{14} + \omega_{15}$(0.77), $\omega_{17} + \omega_{20}$(0.29)
43(2045)	2989.40	1.02	$\omega_{18} + \omega_{19}$(0.82), $\omega_{17} + \omega_{20}$(0.31)
44(2046)	2990.13	0.36	$\omega_4 + \omega_{12} + \omega_{15}$(0.8), $\omega_2 + \omega_4 + \omega_6 + \omega_{12}$(0.3)
45(2062)	2994.76	4.24	ω_{22}(0.65), $\omega_{18} + \omega_{20}$(0.41)
46(2093)	3002.95	0.14	$\omega_1 + \omega_5 + 2\omega_9$(0.62), $2\omega_9 + \omega_{12}$(0.36)
47(2108)	3006.17	0.96	$\omega_3 + \omega_{10} + \omega_{17}$(0.8), $\omega_3 + 2\omega_{10} + \omega_{17}$(0.26)
48(2110)	3006.43	0.18	$\omega_1 + \omega_2 + \omega_3 + \omega_6 + \omega_8$(0.45), $\omega_3 + \omega_4 + \omega_6 + \omega_8$(0.25)
49(2164)	3018.50	0.44	$\omega_2 + \omega_6 + \omega_8 + \omega_9$(0.39), $\omega_6 + \omega_8 + \omega_{16}$(0.37)
50(2210)	3030.52	0.13	$2\omega_3 + \omega_4 + \omega_{13}$(0.68), $\omega_6 + \omega_8 + \omega_{16}$(0.32)
51(2211)	3030.81	0.15	$2\omega_3 + \omega_4 + \omega_{13}$(0.55), $\omega_6 + \omega_8 + \omega_{16}$(0.4)
52(2255)	3042.48	29.57	ω_{23}(0.5), $\omega_5 + \omega_6 + \omega_{19}$(0.29)
53(2265)	3044.12	1.78	$3\omega_1 + \omega_2 + \omega_6 + \omega_8$(0.35), $\omega_2 + \omega_5 + \omega_7 + \omega_8$(0.34)
54(2281)	3046.91	1.41	$\omega_5 + \omega_6 + \omega_{19}$(0.53), $\omega_1 + 2\omega_7 + \omega_8$(0.46)
55(2291)	3049.64	5.98	$\omega_5 + \omega_6 + \omega_{19}$(0.61), $\omega_2 + \omega_5 + \omega_7 + \omega_8$(0.29)
56(2296)	3051.71	24.43	ω_{23}(0.46), $\omega_2 + \omega_6 + \omega_8 + \omega_9$(0.36)
57(2352)	3067.78	0.52	$\omega_{18} + \omega_{20}$(0.65), $\omega_3 + \omega_{10} + \omega_{18}$(0.48)

Fig. 4 represents the evolution of the basis size with each iteration. The exponential growth rate of the basis needed to reach higher accuracy has a strong impact on the ability to solve the eigenproblem, especially for this many eigenvalues.

The results obtained for ε = 0.02 are compared to experimental data[30,31] on Fig. 5. Despite the fact that ε is relatively high and the 4th order force field used may not represent well the potential energy, the A-VCI results are in very good agreement with these data. However, a few differences in the intensity repartition can be observed. In the 1100–1300 cm^{-1} region, there is an inversion regarding the last 2 bands. There is also an overrepresentation of the bands in the 3000 cm^{-1} region, probably attributable to the large number of active frequencies accumulating in this Lorenzian representation.

There are also minor discrepancies on the position of a few bands caused by the precision limitation on this calculation ($\varepsilon = 0.02$). In particular, the $\omega_6 + \omega_7$ (1653.14 cm^{-1}), $\omega_6 + \omega_8$ (1691.72 cm^{-1}) bands and, to a lesser extent, the $\omega_7 + \omega_9$ (1876.49 cm^{-1}), $\omega_8 + \omega_9$ (1911.19 cm^{-1}) bands are slightly shifted toward the low-energy side of the spectrum. This figure also compares these results with a calculation in double harmonic approximation to highlight the effect of mechanical and electrical anharmonicity. Complete numerical results are presented in Table 3. The detailed results obtained for ε ranging from 0.03 to 0.05 are provided in the ESI,† as well as numerical results in double harmonic approximation.

4 Conclusion

One of the main challenges for the computation of anharmonic spectra resides in the choice of a suitable basis to discretize a given vibrational Hamiltonian. We have previously shown that the A-VCI method provides the beginning of a solution by sparingly selecting basis functions in a hierarchical manner, allowing some control over the accuracy of the computed eigenvalues.

In this paper, we managed to reduce the size of the nested bases of the A-VCI algorithm, which allowed us to calculate the IR spectrum of a 10-atom molecule. First, using an energy pruning method, we reduced the total number of available basis functions. Second, the basis-increasing technique, independent of the accuracy of the algorithm, allowed us to progressively refine the results to achieve better accuracy. Finally, the computation of IR intensities allowed us to select only the vibrational states that are active in infrared, thus reducing the size of the successive bases, as well as the memory footprint, while accelerating the computational time.

These developments have been validated on a system with 7 atoms (C_2H_4O), without any significant loss in the accuracy of the eigenvalues of interest, with a 55% computational time reduction in the case we retained (ε_I = 0.1 km mol^{-1}). As a conclusion, we have shown that the SI approach made it possible to calculate 2400 eigenvalues of a molecule with 10 atoms ($C_4H_4N_2$) with a reasonable accuracy. Such calculation would be unattainable without intensity selection. For these large systems, we have clearly shown that the main difficulty lies in the ability of the eigensolver to compute many eigenvalues in a reasonable time, especially when the matrix grows very large. This step corresponds almost entirely to the computational cost (97%). Moreover, the eigensolver is not efficient in such iterative process as it rebuilds from scratch the Krylov subspace at each iteration. Thus, the study of large systems involves working on a new category of eigensolvers. In our case, since the bases are nested from one iteration to the next, one possible improvement consists in using the previous Krylov subspace as a starting point to compute the current Krylov subspace.

In order to calculate the IR spectra of large molecules with good accuracy, we have to work in two complementary directions. First, we must improve the parallelism to decrease the pressure on the memory and speed up the eigenvalue solver. To do this, we have to move to a distributed version of the algorithm. This should allow us to calculate more eigenvalues on larger matrices. Second, there is also a need to improve the quality of the Hamiltonian operator by using higher order force fields that would better represent the potential energy surfaces in order to approximate the experimental data.

Finally, the selection strategy based on intensities has proven to be a success in obtaining accurate results on the frequencies and IR intensities of target vibrational states of a 10-atom molecular system.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocations 2017-A0030810099 and 2018-A0050810638 made by GENCI. Some experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr/). Numerical computations were also carried out using the MCIA platform (Mésocentre de Calcul Intensif Aquitain, see www.mcia.univ-bordeaux.fr). Finally we acknowledge the Direction du Numérique of the Université de Pau et des Pays de l'Adour for the computing facilities it provided us.

Notes and references

1 R. Garnier, M. Odunlami, V. Le Bris, D. Bégué, I. Baraille and O. Coulaud, *J. Chem. Phys.*, 2016, **144**, 204123.
2 M. Odunlami, V. Le Bris, D. Bégué, I. Baraille and O. Coulaud, *J. Chem. Phys.*, 2017, **146**, 214108.
3 J. Bowman, K. Christoffel and F. Tobin, *J. Phys. Chem.*, 1979, **83**, 905–920.
4 K. Christoffel and J. Bowman, *Chem. Phys. Lett.*, 1982, **85**, 220–224.
5 T. Thompson and D. Truhlar, *Chem. Phys. Lett.*, 1980, **75**, 87–90.
6 G. Rauhut, *J. Chem. Phys.*, 2004, **121**, 9313–9322.
7 J. Cooper and T. Carrington, *J. Chem. Phys.*, 2009, **130**, 214110.
8 G. Avila and T. Carrington Jr., *J. Chem. Phys.*, 2011, **134**, 054126.
9 G. Avila and T. Carrington, *J. Chem. Phys.*, 2012, **137**, 174108.
10 G. Avila and T. Carrington Jr., *J. Chem. Phys.*, 2013, **139**, 134114.
11 T. Halverson and B. Poirier, *Chem. Phys. Lett.*, 2015, **624**, 37–42.
12 G. Avila and T. Carrington, *Chem. Phys.*, 2016, 1–16.
13 J. Brown and T. Carrington, *J. Chem. Phys.*, 2016, **145**, 144104.
14 M. Reiher and J. Neugebauer, *J. Chem. Phys.*, 2003, **118**, 1634–1641.
15 J. Neugebauer, C. Herrmann, S. Schenk and M. Reiher, AKIRA - Mode-Tracking of Molecular Vibrations, 2009,

https://ethz.ch/content/dam/ethz/special-interest/chab/physical-chemistry/reiher-dam/documents/Software/akira.pdf.
16 H. Torii, *J. Comput. Chem.*, 2002, **23**, 997–1006.
17 S. Luber and M. Reiher, *Chem. Phys. Chem.*, 2009, **10**, 2049–2057.
18 K. Kiewisch, S. Luber, J. Neugebauer and M. Reiher, *CHIMIA*, 2009, **63**, 270–274.
19 R. Burcl, S. Carter and N. C. Handy, *Chem. Phys. Lett.*, 2003, **380**, 237–244.
20 D. Bégué, I. Baraille, P. A. Garrain, A. Dargelos and T. Tassaing, *J. Chem. Phys.*, 2010, **133**, 34102.
21 H. G. Kjaergaard, B. R. Henry, H. Wei, S. Lefebvre, T. Carrington, O. Sonnich Mortensen and M. L. Sage, *J. Chem. Phys.*, 1994, **100**, 6228–6239.
22 P. Carbonnière, A. Dargelos and C. Pouchan, *Theor. Chem. Acc.*, 2010, **125**, 543–554.
23 E. Lesko, M. Ardiansyah and K. R. Brorsen, *J. Chem. Phys.*, 2019, **151**, 164103.
24 J. Brown and T. Carrington, *J. Chem. Phys.*, 2016, **145**, 144104.
25 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Revision D.01*, Gaussian Inc., Wallingford CT, 2009.
26 R. Lehoucq, D. Sorensen and C. Yang, *ARPACK Users' Guide*, SIAM, 1998.
27 P. Carbonniere and V. Barone, *Chem. Phys. Lett.*, 2004, **392**, 365–371.
28 M. Neff, T. Hrenar, D. Oschetzki and G. Rauhut, *J. Chem. Phys.*, 2011, **134**, 064105.
29 J. Spanget-Larsen, Infrared Intensity and Lorentz Epsilon Curve from 'Gaussian' FREQ Output, 2015.
30 V. K. Shen, D. W. Siderius, W. P. Krekelberg and H. W. Hatch, NIST Standard Reference Simulation Website, 2017, https://webbook.nist.gov/cgi/cbook.cgi?ID=C290379Units=SIMask=80#IR-Spec.
31 A. L. Smith, *The Coblentz Society Desk Book of Infrared Spectra, The Coblentz Society Desk Book of Infrared Spectra*, The Coblentz Society, 2nd edn, 1982.