npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01400-9

Fast prediction of anharmonic vibrational spectra for complex organic molecules

Mattia Miotto [1,4] ✉ & Lorenzo Monacelli [2,3,4] ✉

Interpreting Raman and IR vibrational spectra in complex organic molecules lacking symmetries poses a formidable challenge. In this study, we propose an innovative approach for simulating vibrational spectra and attributing observed peaks to molecular motions, even when highly anharmonic, without the need for computationally expensive ab initio calculations. Our approach stems from the time-dependent stochastic self-consistent harmonic approximation to capture quantum nuclear fluctuations in atom dynamics while describing interatomic interaction through state-of-the-art reactive machine-learning force fields. Finally, we employ an isotropic charge model and a bond capacitor model trained on ab initio data to predict the intensity of IR and Raman signals.

Insight of chemical bonds and knowledge of molecular symmetries allow us to quickly analyze and predict the vibrational spectrum and the activity of Raman and IR modes in simple molecules. This allows the practical identification of well-known stretching and bending signatures of different chemical species. Symmetries constrain the vibrational patterns to align on high symmetry planes and axes, while determining the multiplicity of modes and their Raman or IR activity. However, the power of group theory diminishes in more complex organic molecules where few or no symmetries are present, like in complex sugars, lipids, amino acids, or proteins[1,2]. Interpreting the vibrational spectrum in these cases is much more challenging, as the number of Raman and IR active modes increases linearly with the number of atoms in the molecule. While intuition and experience are still fundamental assets to interpret vibrational spectra, theoretical predictions are a handy aid for experimenters. The accurate prediction of Raman and IR data often comes with major costs in the complexity of employing existing software and the computational resources required. High-level ab initio calculation needs powerful high-performance computers and a deep understanding of the theory behind these tools, while most empirical force fields lack the necessary precision. Moreover, atoms present in organic molecules are light and sensitive to sizable quantum fluctuations even at room temperature[3], which can trigger the anharmonic nature of molecular bonds, as it occurs in the OH stretching[3,4]. While coupling quantum nuclear dynamics with advanced quantum chemistry tools to describe the electronic ground state has proven highly effective in reproducing—with high accuracy—the vibrational behavior of small molecules, its computational cost prevents its daily applications in more challenging systems.

Significant steps forward have been made thanks to the availability of machine-learning force fields for the interatomic potential, which were able to bring near chemical accuracy at a computational cost that is just a fraction of any ab initio calculation[5–7]. Of particular interest is the ANI-1ccx model[8,9], trained on the GDB database of high-level coupled cluster CCSD(T) energies of organic molecules containing H, C, N, O atoms[10], and the more recent ANI-2x, which extends the force-field by including S, F and Cl atomic species[11], making up for ~90% of all drug-like molecules. The employment of such machine-learned force field allows the application of advanced tools to simulate the quantum nature of nuclei accounting for anharmonicity in both solids and molecules, like path-integral molecular dynamics (PIMD) (see, e.g., refs. 12–14). However, these calculations still remain time-consuming when run on ordinary computers. Moreover, PIMD is a static theory in which extracting the dynamical features of the vibrational spectrum remains an open challenge.

In this work, we introduce a method able to capture the quantum anharmonic dynamics of the nuclei while keeping a good description of the electronic degrees of freedom to predict and analyze the vibrational spectrum of molecules as big as small proteins, with a computational cost low enough to run on any common laptop for few hours. The method is distributed as an open-source web app named *FAbulA* : FAst viBroscopy of molecULes with Anharmonicity.

FAbulA combines the time-dependent Stochastic Self-Consistent Harmonic Approximation (tdSSCHA) to simulate the quantum and anharmonic motion of nuclei[15] with different force fields for the interatomic potential. In the following sections, we describe the underlying methodology of tdSSCHA and the implementation of *FAbulA* in detail. We also present several examples of applications of the method to illustrate its capabilities.

[1]Istituto Italiano di Tecnologia (IIT), Center for Life Nano & Neuro Science, Viale Regina Elena 291, Roma, 00161, Italy. [2]Sapienza University of Rome, Department of Physics, Piazzale Aldo Moro, 5, Rome, 00185, Italy. [3]Theory and Simulation of Materials (THEOS), and National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. [4]These authors contributed equally: Mattia Miotto, Lorenzo Monacelli. ✉e-mail: mattia.miotto@iit.it; lorenzo.monacelli@uniroma1.it

Results

Outline of the *FAbulA* method

The calculation of Raman and IR spectra of a molecule is a two-step procedure: one has to (i) evaluate the frequencies and polarization modes of the atomic vibrations and (ii) simulate the molecular response to light irradiation, either inelastic Raman scattering or infrared absorption. *FAbulA* can compute the molecular vibrational excitations with three degrees of accuracy: (i) harmonic, (ii) static anharmonic, and (iii) dynamic anharmonic (Fig. 1).

The Harmonic spectrum is temperature-independent, and the infinite lifetime of lattice vibrations gives rise to arbitrary narrow spectral lines. The harmonic excitations are evaluated through a finite-difference approach, where the atoms are slightly displaced from their average positions, and curvature of the energy landscape is obtained from the numerical derivative of the forces. In particular, all the calculations in this work are performed using either ANI-1ccx or ANI-2x[8]. These are very accurate neural network force fields trained on high-quality quantum chemistry calculations and can be selected from the web interface/command line package of *FAbulA* (details in the Methods).

In the static anharmonic approach, *FAbulA* dresses the harmonic spectrum with anharmonicities triggered by quantum and thermal fluctuations of nuclei. The anharmonic shift corrects both the frequencies and polarization modes of the harmonic calculation. The procedure is a mean-field technique, where the curvature of the energy landscape is evaluated on an average region of the phase space obtained as the equilibrium ensemble of an auxiliary self-consistent Harmonic system. In particular, we adopt the SSCHA[16]. The overall static anharmonic spectrum is evaluated from the resulting self-consistent harmonic potential. Like in the harmonic case, phonon peaks have infinite lifetimes, but their position and intensity depend on temperature. Further details on the implementation are discussed in the Methods.

Lastly, the complete dynamical anharmonic calculation simulates the quantum dynamics of nuclei in the anharmonic potential at any given temperature. This approach goes beyond the mean field, enabling the scattering and decay of vibrations, resulting in a complex spectrum displaying the natural lifetime broadening of peaks, overtones, combinational modes, and complex satellite features. The dynamical anharmonic calculation is obtained as the linear response of the time-dependent stochastic self-consistent harmonic approximation (tdSSCHA)[15,17]. The computational cost of each approach is incremental, and each stage is a required starting point for the next one: the harmonic dynamical matrices are needed to initialize the SSCHA, and the self-consistent harmonic potential is required to set the equilibrium condition for the time evolution in the tdSSCHA.

The harmonic, SSCHA, and tdSSCHA runs allow for the evaluation of frequencies, polarization modes, and spectral functions of the molecule. However, to estimate the activity and intensity of each mode, we must simulate the interaction between photons and vibrations. *FAbulA* provides two models to compute the dipole and polarizability induced by the atomic motion, which can be employed to simulate the Raman and IR spectra. In particular, the coupling between light and the atoms in the lattice is modeled with the effective charges (Z) and Raman tensors (Ξ), representing the variation of dipole (μ) and polarizability (α) in the molecule when an atom is displaced from its equilibrium position:

$$Z_{\alpha j\beta} = \frac{\partial \mu_\alpha}{\partial R_{j\beta}}, \qquad \Xi_{\alpha\beta i\gamma} = \frac{\partial \alpha_{\alpha\beta}}{\partial R_{i\gamma}}. \tag{1}$$

The intensity of Raman and IR peaks within the harmonic and SSCHA approximation is estimated from the squared projection of the effective charge and Raman tensor on the polarization of each vibration as

$$I_\mu^{(\mathrm{IR})} = \sum_{j\beta\alpha} \left(\frac{Z_{\alpha j\beta} e_\mu^{j\beta}}{\sqrt{m_j}} \right)^2, \tag{2}$$

where e_μ is the polarization vector of the μ-th vibrational mode, and m_j is the mass of atom j. We employ an analogous expression for the Raman, with a bit of extra care to cover the case of unpolarized light[18] by estimating the so-called Placzek rotation invariats[19]. The dynamical IR spectrum is evaluated from the dipole autocorrelation function, while the Raman spectrum is obtained from the polarizability. Therefore, there is no predefined one-on-one assignment of each peak with the corresponding lattice vibration, but it also accounts for overtones, natural broadening, and combinational modes. The spectrum is evaluated using a Lanczos algorithm to invert the interacting anharmonic Green's function[15]. More details are reported in the Methods section.

The computation of effective charges can be performed ab initio with linear response, as implemented in Quantum ESPRESSO[20] for small molecules. For bigger systems, we approximate the effective charges as an isotropic charge on each atom q_i,

$$Z_{\alpha i\beta} = q_i \delta_{\alpha\beta}, \tag{3}$$

where the value of q_i can be obtained from the input PDB structure, the topology files, or standard force-field parameters (point-charge model).

For the Raman intensities, we derived an analytical bond capacitor model, trained on a subset of ab initio estimations of small organic molecules, introduced and discussed in more detail in the Methods section. The bond capacitor model evaluates the Raman tensor for any organic molecule with a negligible computational cost compared to the calculation of vibrations.

Identification of vibrational modes from the spectrum

We tested the capability of *FAbulA* to evaluate the vibrational frequencies of organic molecules. Figure 2a displays the experimental Raman (left) and

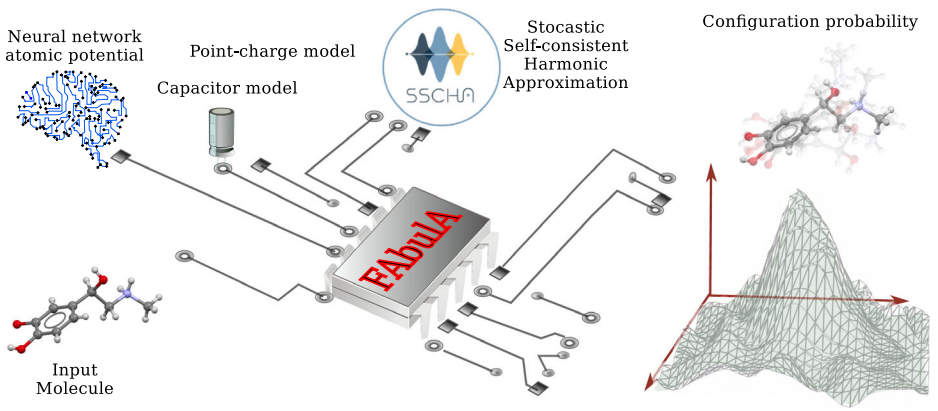

Fig. 1 | Schematic representation of the *FAbulA* method. Starting from an input molecule of known 3D structure, the method combines existing force fields with a simple capacitor model and point-charge model to build a starting dynamical matrix with an associated polarizability and polarization matrix, which are then minimized according to the SSCHA (Stochastic Self-Consistent Harmonic Approximation) theory. The SSCHA logo is reproduced with permission from the owner.

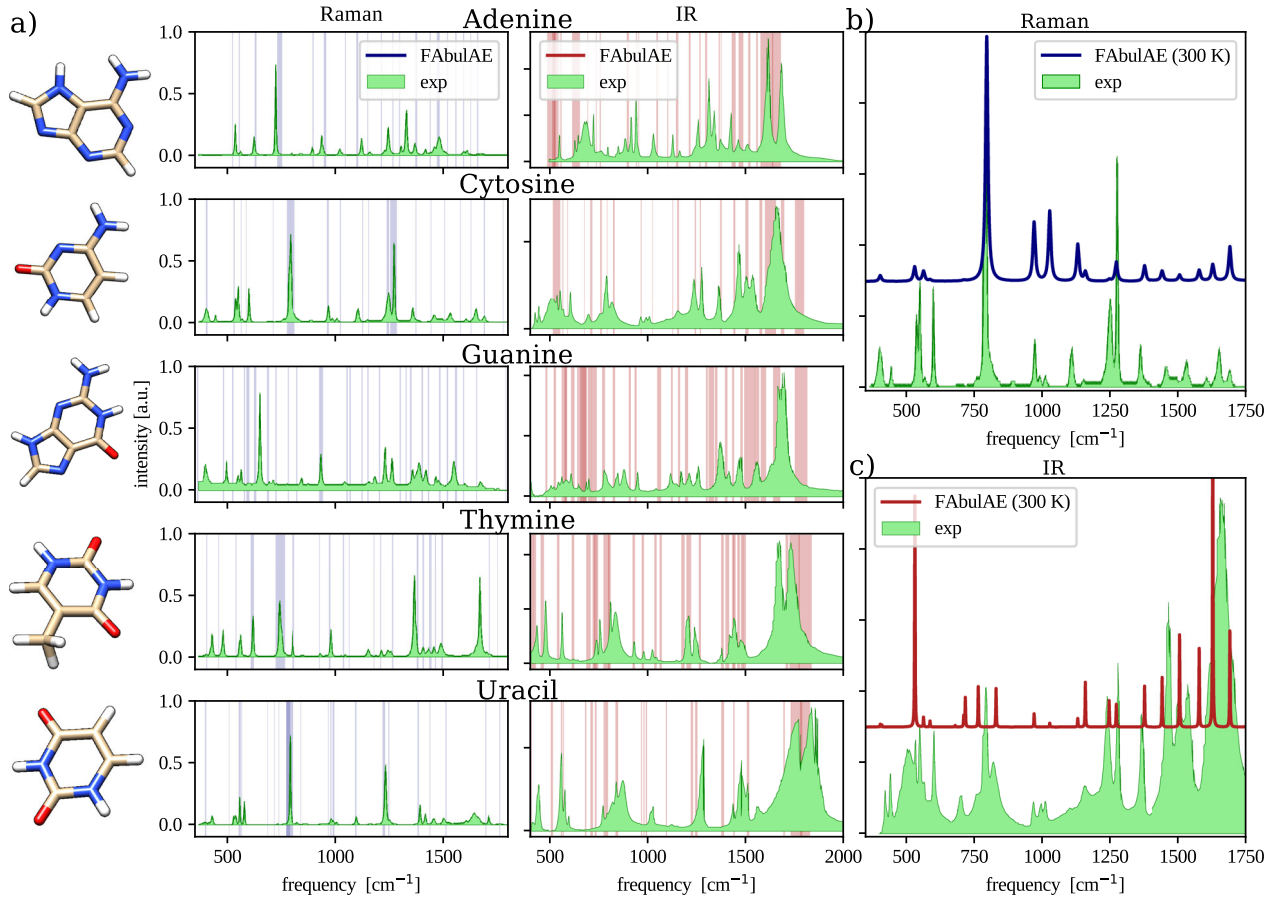

Fig. 2 | Comparison between experimental and predicted Raman and IR vibrational frequencies of the five main nucleobases. a From left to right, sticks representation of the considered nucleobasis, Raman, and infrared spectra. Green-shaded curves represent experimental Raman and infrared spectra of molecules in crystalline phase taken from the work of De Gelder[37] and Bec et al.[38], respectively. Blue and red vertical lines mark the positions of the active vibrational frequencies predicted by *FAbulA* with the SSCHA method. **b** Comparison between the experimental Raman spectrum of cytosine[37] and the prediction of *FAbulA* with the SSCHA method combined with the neural-network derived-polarization matrix to determine the relative intensities. **c** Same as in **b** but for the infrared spectrum of cytosine.

infrared (right) spectra of the five nucleic acids -i.e., adenine, cytosine, guanine, thymine, and uracil- as green curves in the 400–1700 cm^{-1} and 450–2000 cm^{-1} spectral ranges, respectively. Vertical blue and red lines mark the predicted active mode positions for Raman and infrared spectra, respectively. The thickness of the lines indicates the agreement between the position of the experiment and the predicted peaks. As one can see, *FAbulA* detects the peaks with higher spectral weight in both kinds of spectra. To determine active Raman/IR modes, we used our bond capacitor model to compute the polarizability dynamical autocorrelation and the effective charges computed ab initio (within DFPT as implemented in quantum ESPRESSO) for the dipole dynamical autocorrelation. Active modes are defined as those participating at least 0.01% of the total spectrum intensity. The unpolarized Raman spectrum of cytosine is reported in Fig. 2b, where we address the bond capacitor model efficacy. In particular, the overall relative intensity of the vibrational modes is well captured by the model, except for the 1250 cm^{-1} vibration, whose intensity is underestimated with respect to the experiment. Analogously, Fig. 2c reports the cytosine IR spectrum, where a similar good agreement is achieved. Overall, the *FAbulA* method estimates the relative intensities between the different modes at a small fraction of the computational cost required from a full ab initio simulation. In Supplementary Note 1, we report a careful test of the method versus ab initio simulation, separately assessing the impact on the vibrational spectra of the approximation of the potential energy landscape (the ANI neural-network force-field), the point-charge approximation for the IR intensity, and the bond capacitor model for the Raman tensor. Note that the obtained accuracy can be further increased considering the effect of the molecule's surrounding environment. As discussed in Supplementary Note 2, *FAbulA* has the potential to simulate molecules in solution by explicitly accounting for the solvent structure. Considering cytosine embedded in a water shell allows *FAbulA* harmonic prediction to describe the experimental infrared spectrum of cytosine with a good approximation. *FAbulA* also provides the possibility to inspect visually and quantitatively (bending and stretching components) the atomic motion associated with each vibrational mode, which can be helpful in comparing the predicted Raman/IR features with similar known modes of different molecules.

Peak relative intensity and width

While the SSCHA-based calculation provides a fast estimation of the vibrational mode positions considering anharmonic effects, it does not yield indications on the peak width. This information can be uncovered by employing the time-dependent SSCHA (see Methods for details). A comparison between the SSCHA and tdSSCHA approaches is provided in Fig. 3, where ethanol is considered as a case of study. Figure 3a, b shows the experimental Raman (panel a) and infrared (panel b) spectra of ethanol in the gas phase as green-shaded curves, together with the predictions of *FAbulA* using the SSCHA and tdSSCHA, respectively. As one can see, the positions of the prominent active peaks are significantly improved by employing the tdSSCHA. In addition, the time-dependent method also provides a quantitative estimation of the modes' lifetimes, i.e., the peak linewidths, as evident, for instance, in the bending (~420 cm^{-1}) and stretching (~920 cm^{-1}) modes of Fig. 3b and Table 1.

In Fig. 3c, d, we reported, instead, the comparison between experimental[21] and predicted absorbance infrared spectra for ethanol molecules in the gas phase. Also, the positions of predicted IR active modes

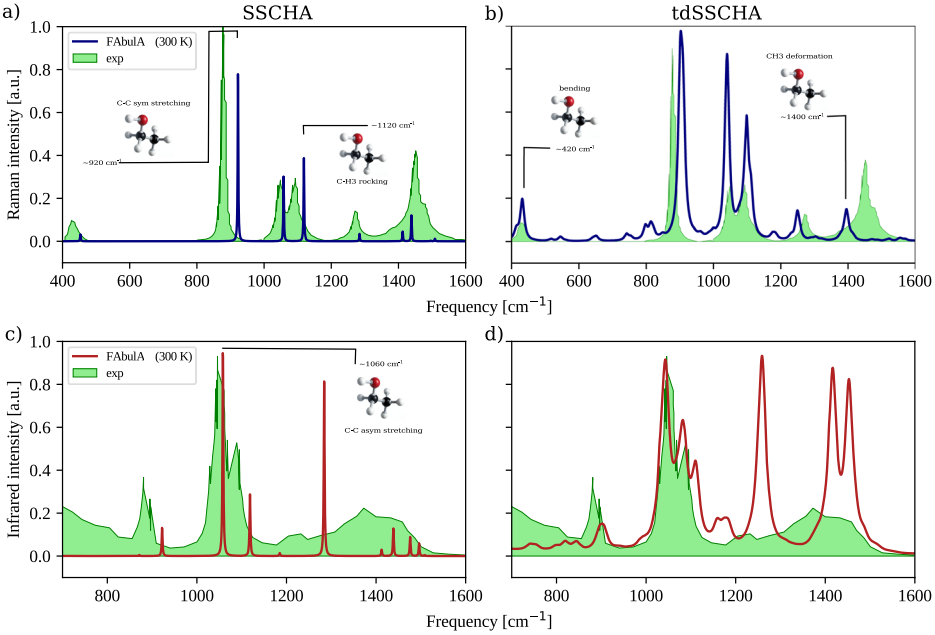

Fig. 3 | Comparison of *FAbulA* results using SSCHA or tdSSCHA method. a Comparison between the predicted (blue curve) and measured (green) Raman spectrum of the ethanol molecule. Experimental data for liquid ethanol are taken from ref. 39, while predictions are obtained via *FAbulA* with the SSCHA method. Frequencies of the four vibrational modes of higher intensity are reported. **b** Same as in **a** but using the tdSCHA method. **c** Same as in **a** but for the infrared spectrum of the ethanol molecule. Experimental data of liquid ethanol are taken from ref. 21. **d** Same as in **b** but for the infrared spectrum of the ethanol molecule.

Table 1 | Predicted vibrational modes of ethanol by *FAbulA* via the SSCHA, together with their percentage of stretching and bending components against the experimental Raman[36] active modes and relative assignments

Mode	Frequency [cm^{-1}]	Stretching	Bending	Exp. [cm^{-1}]	Assignment	Active mode
1	227.1	-	100%			
2	434	-	100%			
3	453.1	4%	96%			
4	636.7	-	100%			
5	871.5	-	100%			
6	922.3	38% [CC(21), CO(23)]	62%	883	C–C, C–O sym. stretching	R
7	1058.1	46% [CC (37), CO (28)]	54%	1051	C–C, C–O asym. stretching	R
8	1118.4	21% [CO (19), CH (2)]	79%	1096	CO stretch, C–H3 rock, in plane COH def.	R
9	1184.9	1%	99%			
10	1284.1	9% [CC (3), CO (5), OH (1)]	91%	1275.7	CH2 twist, in plane COH deformation	R
11	1291.7	2%	98%			
12	1412.4	8% [CC (5), CO (2), CH (1), OH (1)]	92%			
13	1438.5	23% [CC (16), CO (4), CH (2), OH (1)]	77%	1453.7	CH3 deformation	R
14	1476.0	1%	99%			
15	1496.1	7% [CC (4), CO (1), CH (2)]	93%			
16	1508.6	8% [CC (2), CO (3), CH (3)]	92%			
17	2904.1	86% [CC (3), CO (2), CH (81)]	14%	2880.6	CH2 symmetric stretch	R
18	2931.5	73% [CH]	27%	2928.7	CH3 stretch	R
19	2942.3	92%[CC (4), CH (88)]	8%			
20	3029.8	72% (CH)	28%			
21	3034.8	72% (CH)	28%			
22	3728.7	78% [CO (2), CH (1), OH (75)]	22%			

are in accordance with experimental data. In particular, both relative intensities and lifetimes of the C–C stretching modes in the region ~1060 cm^{-1} are well reproduced by the tdSSCHA.

Temperature dependence

Being able to account for anharmonicities, *FAbulA* allows us to study the temperature-dependence behavior of vibrational modes. To test the capability of *FAbulA*, we considered the case of study of myristic acid, a common saturated fatty acid composed of 44 atoms for which the Raman active modes show a significative temperature dependence (Fig. 4). In particular, Miranda et al.[22] collected unpolarized Raman spectra of myristic acid powders at different temperatures in the range of 20–300 K. At room temperature, myristic acid presents Raman active modes in the spectral range of 880–920 cm^{-1}: one centered at ~897 cm^{-1} associated to the C_{13}−C_{14} stretching and the other localized at ~910 cm^{-1} and assigned to C_1–C_2–C_3 stretching and C_1–H_2 wagging. Interestingly, increasing the

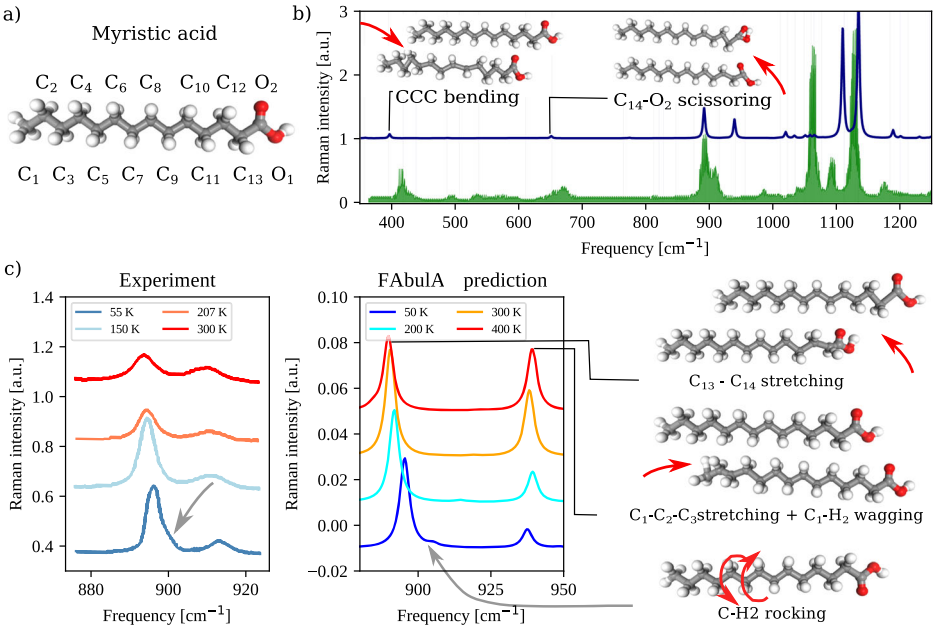

Fig. 4 | Temperature dependence of *FAbulA* results using the SSCHA method for myristic acid. **a** Balls and sticks representation of the myristic acid molecule with the associated atom nomenclature. **b** Comparison between predicted (blue curve) and measured (green) Raman spectrum of the myristic acid molecule. Experimental data of crystalline myristic acid are taken from ref. 37, while predictions are obtained via *FAbulA* with the SSCHA method. The frequencies of the two low-frequency vibrational modes of higher intensity are reported together with the corresponding atomic motion. **c** Comparison between experimental (left) and predicted (right) behaviors of the vibrational modes in the spectral region 850–950 cm^{-1} as a function of temperature. The experimental spectrum of crystalline myristic acid data is taken from Miranda et al.[22]. Atomic motions of the three observed peaks are shown together with their assigned modes.

temperature produces three significant changes: (i) a red-shift of the 897 cm^{-1} and the 910 cm^{-1} modes, (ii) a decrease in the relative differences between the intensities of the two modes, and (iii) the disappearance of a third peak in the Raman spectrum a temperature lower than ~60 K. Figure 4c compares experimental and predicted Raman spectra for myristic acid at different temperatures. *FAbulA* reproduces the correct active modes and their relative intensities compared with experiments. Moreover, we successfully explain both the red-shift of the 897 cm^{-1} mode upon increasing the simulation temperature as an anharmonic interaction and the presence of the third mode–predicted to be a $C{-}H_2$ rocking, as noted by ref. 23, disappearing upon heating. On the opposite, we do not reproduce the redshift of the 910 cm^{-1} mode, whose slight residual temperature dependence displayed in Fig. 4 is an effect of the stochastic sampling of the energy landscape.

Discussion

Spectroscopy investigation of the optical properties of organic molecules is one of the easiest and most common ways to characterize a sample. In particular, Raman and IR signals help understand sample composition and provide valuable insights into the structural and chemical properties of the molecules. While experimental apparati for standard Raman and IR measurements are commonly available, tools for the rapid characterization of the measured vibrational profiles are still largely missing, and the interpretation of the observed spectra is often handwaving when not based on time-consuming numerical ab initio simulations. Indeed, over the past decades, various methods have been developed to address this problem, ranging from semi-empirical approaches to highly accurate quantum mechanical computations[22]. Among the latter, the self-consistent harmonic approximation has emerged as a powerful and efficient tool for the computation of anharmonic thermodynamic properties and spectra of small to medium-sized crystalline systems[16]. Indeed, the method has been shown to provide highly accurate prediction for condensed matter systems, and it is considered the state of the art in the field[24–26].

In this work, we present *FAbulA*, a computational protocol for the fast computation of the vibrational modes of organic molecules, taking into account anharmonic effects. *FAbulA* is based on the union of the stochastic self-consistent harmonic approximation formalism, able to model the anharmonic fluctuations of atomic vibration[4,15,16] and neural network-based force-fields for the rapid evaluation of the atomic interaction of molecules.

The main features of static *FAbulA* calculation include rapid evaluation of the vibrational modes of the given organic molecule, evaluation of Raman and infrared active modes, and their relative intensities, as we showed in Fig. 2 using the five nucleic bases as examples. Notably, the static approach allows for assigning the atomic motions corresponding to each vibrational mode. We benchmarked the predicted modes of ethanol with the respective assigned nuclear motions, finding that *FAbulA* correctly classifies all ethanol principal modes (see Table 1). Additional benchmarks of the approximation for the Raman/IR intensities, the effect of the solvent, and the atomic energy landscape are analyzed in the Supplementary Information (see Supplementary notes 1–4). In addition, considering the anharmonic effects, our method permits us to consider the temperature-dependence behavior of the found vibrational modes. We tested its potential in the case of myristic acid, where an experimental investigation of the Raman spectrum of myristic acid at different temperatures highlighted a well-defined trend of its active modes in the $880-920\,cm^{-1}$ region, perfectly explained by the *FAbulA* calculations as shown in Fig. 4.

As the static approach can not provide information on mode bandwidths and overtones, we implemented a dynamical version of *FAbulA*, based on the time-dependent self-consistent harmonic approximation. The dynamic results systematically improve the agreement with the experimental data while simultaneously predicting the natural linewidth emerging from anharmonic scattering (see Fig. 3).

Based on combining a solid theoretical framework with neural-network algorithms to evaluate total energies and polarizability tensors efficiently, *FAbulA* can be applied to a wide range of organic molecules made of hundreds of atoms. Alternatively, the method can also account for the explicit effects of solvents or aggregate states thanks to the fast evaluation of the vibrational spectrum (more details in Supplementary Note 2).

Limitations of the current method comprehend the description of molecular rotations[27], as the corresponding atomic motion strongly deviates from the linear vibration of a harmonic oscillator. To mitigate this issue for the free molecular rotation, we lock both the average position and the global rotation of the molecule and study only the vibration of the internal degrees of freedom. This is correct in the approximation that the global rotation of the molecule can be decoupled from the other phonons, which is true if we neglect the centrifugal force due to the rotation of the other internal vibration. Indeed, the larger the molecule, the better this approximation becomes. A different case occurs for molecules that present substructures that can freely rotate within themselves, like paracetamol, where the final

methyl group (C–H3) presents rotational degrees of freedom that the SCHA method fails to describe[27].

Overall, our method combines state-of-the-art computational techniques with a user-friendly interface that permits even non-expert users to calculate the vibrational properties of organic molecules and obtain quick estimations of mode motions and temperature-dependent behaviors.

We envisage that *FAbulA* will shorten the distance between experimenters and computational/theoretical methods, allowing for a prompt comprehension of the vibrational properties of organic molecules.

Material and methods

The *FAbulA* method

The harmonic spectrum is evaluated using a finite-difference approach: the molecule is relaxed in the closest local minimum of the Born-Oppenheimer energy landscape (the total energy of the molecules where the ions are at position $\boldsymbol{R}$, assuming the electrons are always in the ground state). Then, a complete set of symmetry inequivalent atomic displacements is selected, for which we evaluate the forces $\boldsymbol{f}$ acting on the ions. We generate the remaining displacement, $\boldsymbol{u}$, and forces by applying the symmetry operations until we have a complete Cartesian set, for which we evaluate the harmonic force-constant matrix $\overset{(h)}{\Phi}$ inverting the set of equations

$$f_\alpha^i = -\sum_{\substack{j=1,\cdots,N \\ \beta=1,2,3}} \overset{(h)}{\Phi}{}_{\alpha\beta}^{ij} u_\beta^j; \tag{4}$$

Latin indices identify the atom, while Greek letters are the Cartesian component. The frequencies and polarization modes are obtained by diagonalizing the force-constant matrix divided by the masses:

$$\frac{\overset{(h)}{\Phi}{}_{\alpha\beta}^{ab}}{\sqrt{m_a m_b}} = \sum_{\mu=1}^{3N} \omega_\mu^2 e_{\mu\alpha}^a e_{\mu\beta}^b, \tag{5}$$

where ω_μ is the frequency of the μ-th mode, e_μ is the polarization vector defining the pattern of the vibrations, and the cross-section of the μ-th mode with the IR or Raman probe.

The harmonic force constant matrix $\overset{(h)}{\Phi}$ is employed as the starting point for the static anharmonic calculation, performed with the SSCHA[4,16,28,29]. The SSCHA method finds the force constant matrix that minimizes the Gibbs free energy, solution of the self-consistent equations

$$\left\langle \frac{d^2 V}{dR_\alpha^i dR_\beta^j} \right\rangle_{\tilde{\rho}\left[\mathscr{R},\overset{(s)}{\Phi}\right]} = \overset{(s)}{\Phi}{}_{\alpha\beta}^{ij}, \tag{6}$$

$$\left\langle \frac{dV}{dR_\alpha^i} \right\rangle_{\tilde{\rho}\left[\mathscr{R},\overset{(s)}{\Phi}\right]} = 0, \tag{7}$$

where $V(R)$ is the Born-Oppenheimer energy landscape, and the averages are performed with a Monte-Carlo algorithm, where the atomic configurations are extracted with an ionic position distributed according to $\tilde{\rho}[\mathscr{R}, \overset{(s)}{\Phi}]$:

$$\tilde{\rho}\left[\overset{(s)}{\Phi},\mathscr{R}\right](R) = \sqrt{\frac{\det \Upsilon}{(2\pi)^{3N}}} \exp\left[-\frac{1}{2}\sum_{\substack{ab=1,N \\ \alpha\beta=1,2,3}} \left(R_\alpha^a - \mathscr{R}_\alpha^a\right)\Upsilon_{\alpha\beta}^{ab}\left(R_\beta^b - \mathscr{R}_\beta^b\right)\right] \tag{8}$$

with

$$\Upsilon_{\alpha\beta}^{ab} = \sqrt{m_a m_b}\sum_{\mu=1}^{3N} \frac{2\overset{(s)}{\omega}{}_\mu}{\hbar(2n_\mu+1)} \overset{(s)}{e}{}_{\mu\alpha}^{a}\overset{(s)}{e}{}_{\mu\beta}^{b} \tag{9}$$

where $\overset{(s)}{\omega}$ and $\overset{(s)}{e}$ are frequencies and polarization vectors of $\overset{(s)}{\Phi}$, while n_μ is the Bose-Einstein occupation factor

$$n_\mu = \frac{1}{e^{\beta\hbar\overset{(s)}{\omega}_\mu} - 1}, \qquad \beta = \frac{1}{k_b T}. \tag{10}$$

The probability distribution for the atoms in (8) is Gaussian around the centroid position $\mathcal{R}$ with a width determined by the covariance matrix Υ^{-1} defined in (9). Notably, the covariance matrix takes into account the quantum zero-point motion of atoms, making our method able to correctly describe quantum-anharmonicity, particularly relevant when light atoms are present.

Thanks to the Gaussian form of (8), we can extract an ensemble of randomly distributed atomic configurations according to $\tilde{\rho}[\overset{(s)}{\Phi}, \mathcal{R}]$ almost instantaneously, without the need of any Metropolis or MD. These configurations evaluate the averages in (6) and (7). To speed up the self-consistency, we employ the importance sampling[16,30,31] to avoid recomputing the ensemble after each iteration and integrate by parts (6) to have only the first derivatives of the Born-Oppenheimer energy landscape[16,29]; so that we only need to compute the forces of the atomic configurations. More details on the implementation are reported in ref. 16.

Once (6) and (7) are satisfied, we use the $\overset{(s)}{\omega}$ and $\overset{(s)}{e}$ as vibrational modes to compute the Raman and IR. This is the static anharmonic approximation, as it considers changes in the molecules due to both temperature and quantum effects, as well as a self-consistent change of the vibrational frequencies in the distribution.

The dynamical anharmonic spectrum is obtained by inserting an explicit time dependence in the probability distribution $\tilde{\rho}[\overset{(s)}{\Phi}, \mathcal{R}]$ and imposing the stationary Action principle[15]. This procedure takes the name of time-dependent stochastic self-consistent harmonic approximation (tdSSCHA) and can be employed to evaluate the response function of the system under an external perturbation (Raman or IR). While this procedure concedes with computing the Raman or IR signal directly from the frequencies and polarization vectors in a perfectly harmonic system, this is not true in the presence of anharmonicity. In particular, scattering and interactions between different vibrational modes could result in the appearance of overtones or combinational modes, as well as change the linewidth by introducing a finite lifetime. The details of the tdSSCHA implementation are reported elsewhere[15,17].

IR and Raman intensity. Raman and IR intensities depend on the response function of dipole μ and polarizability α of the molecule. In particular, the IR intensity is related to the dipole-dipole response function

$$I_x^{(\mathrm{IR})}(\omega) \propto -\Im\int_{-\infty}^{\infty} e^{-i\omega t}\langle \mu_x(t)\mu_x(0)\rangle, \tag{11}$$

while the Raman is related to the autocorrelation of the polarizability tensor

$$I_{xy}^{(\mathrm{Raman})}(\omega) \propto -[n(\omega)+1]\Im\int_{-\infty}^{\infty} e^{-i\omega t}\left\langle \alpha_{xy}(t)\alpha_{xy}(0)\right\rangle. \tag{12}$$

The time dependence of μ and α is due to the movement of atoms in the lattice. By expanding the expression of the dipole and polarization around equilibrium, we get the one-phonon contribution to the vibrational

spectra[17]:

$$\mu_x(t) = \mu_x(0) + \sum_{i\alpha} \frac{\partial \mu_x}{\partial R_{i\alpha}} (R_{i\alpha}(t) - \mathscr{R}_{i\alpha}) \qquad Z_{xi\alpha} = \frac{\partial \mu_x}{\partial R_{i\alpha}} \tag{13}$$

$$\alpha_{xy}(t) = \alpha_{xy}(0) + \sum_{i\beta} \frac{\partial \alpha_{xy}}{\partial R_{i\beta}} (R_{i\beta}(t) - \mathscr{R}_{i\beta}) \qquad \Xi_{xyi\beta} = \frac{\partial \alpha_{xy}}{\partial R_{i\beta}} \tag{14}$$

where $Z_{xi\alpha}$ is the born effective charge, while $\Xi_{xyi\beta}$ is the Raman Tensor. Then, the IR and Raman intensity are obtained as

$$I_x^{(\mathrm{IR})}(\omega) \propto -\sum_{i\alpha j\beta} \frac{Z_{xi\alpha} Z_{xj\beta}}{\sqrt{m_i m_j}} \Im \underbrace{\int_{-\infty}^{\infty} e^{-i\omega t} \left\langle \sqrt{m_i}(R_{i\alpha}(t) - \mathscr{R}_{i\alpha}) \sqrt{m_j}(R_{j\beta}(0) - \mathscr{R}_{j\beta}) \right\rangle}_{G_{i\alpha j\beta}(\omega)} \tag{15}$$

$$I_x^{(\mathrm{Raman})}(\omega) \propto -[n(\omega)+1] \sum_{i\alpha j\beta} \frac{\Xi_{xyi\alpha} \Xi_{xyj\beta}}{\sqrt{m_i m_j}} \Im \underbrace{\int_{-\infty}^{\infty} e^{-i\omega t} \left\langle \sqrt{m_i}(R_{i\alpha}(t) - \mathscr{R}_{i\alpha}) \sqrt{m_j}(R_{j\beta}(0) - \mathscr{R}_{j\beta}) \right\rangle}_{G_{i\alpha j\beta}(\omega)} \tag{16}$$

where $G_{i\alpha j\beta}(\omega)$ is the dynamical vibrational Green's function. The self-consistent harmonic Green's function has the form

$$\overset{(s)}{G}(\omega) = \left[\omega^2 - \overset{(s)}{D}\right]^{-1} \qquad \overset{(s)}{D}_{i\alpha j\beta} = \frac{\overset{(s)}{\Phi}_{i\alpha j\beta}}{\sqrt{m_i m_j}}$$

where the eigenvalues of the dynamical matrix $\overset{(s)}{D}$ determine the position of the vibrational peaks. The dynamical tdSSCHA spectrum is evaluated by computing the interacting Green's function accounting for scattering between different atoms[15].

The IR and Raman activity is determined by the projection of effective charges and Raman tensor on the polarization vector of the vibration. Usually, these quantities are evaluated ab initio using density-functional theory or Hartree-Fock, either employing linear-response[32] or finite differences[33,34]. However, these calculations require considerable computational power for anything with more than 10 atoms. Instead, we approximate the tensorial dependences of $Z_{xi\alpha}$ as an isotropic charge q_i located on the atom, as usually assumed in classical force fields (as reported, for instance, in PDB or parametrization files):

$$Z_{xi\alpha} = \delta_{x\alpha} q_i \tag{17}$$

where δ_{ij} is the Kronecker delta.

The Raman tensor is far more challenging to evaluate as it is a property related to the bonding network of the molecule. For this reason, we trained a Neural-Network potential on a bond capacitor model of organic molecules. A tensorial capacitor $C^{(ij)}_{ijhk}$ is located between each couple of bounded atoms i and j in the molecule (as indicated in the topology file). The total polarizability is evaluated as a sum over all bonds of a local quantity that depends on the atomic bonds

$$\alpha_{xy} = \sum_{(i,j)\in \text{bonds}} \sum_{\alpha\beta} C^{(ij)}_{xy\alpha\beta} (R_{i\alpha} - R_{j\alpha})(R_{i\beta} - R_{j\beta}). \tag{18}$$

The Raman tensor of each atom is obtained from the derivative of the total polarizability

$$\Xi_{xyi\alpha} = 2 \sum_{\beta} \sum_{(k,i)\in \text{bonds}} C^{(ki)}_{xy\alpha\beta} (R_{i\beta} - R_{k\beta}) \tag{19}$$

Even if the C has 36 independent components for each bond, we assume $C^{(ij)}$ to depend only on the atomic species of atoms i and j and their distance in the bond to distinguish between single, double, and triple bonds. Under this assumption, we neglect the role played by the environment of each bond, and the capacitor tensor gains rotational symmetry around the bond, retaining only 6 independent quantities.

The independent values of $C^{(ij)}$ are obtained by training a neural-network model on a small dataset of molecules to correctly reproduce the Raman tensor evaluated within DFT with LDA approximation for exchange-correlation functional.

Used dataset and chosen parameters

The 3D structures in MOL2 format of nucleic basis, ethanol, and myristic acid were taken from the ZINC database[35]. Molecules' Raman and infrared spectra were calculated using the *FAbulA* webserver with 100,000 configurations, a maximum number of populations of 20 and the ANI-1ccx force field for the SSCHA minimization; the number of steps for the tdSCHA algorithm was set to 10,000. Finally, the following molecule-specific parameters were used: Nucleic basis: temperature was set to 300 K; Ethanol: temperature was set to 300 K; Myristic acid: temperature was set to 50, 150, 200, and 300 K.

Neural network

To get the parameters of the bond capacitor model, we implemented a multi-layer perceptron (MLP) regression algorithm via the sklearn.neural-network.MLPRegressor module within the scikit-learn library.

Webserver

The method is distributed as a free webserver, designed for seamless access and operation at: https://fabula.bio-groups.com, where users can find a complete guide to how to use the *FAbulA* tool.

Data availability

All relevant data are reported in figures and tables within the main text and the Supporting Information. A digitalized version can be made available upon reasonable request to the authors.

Code availability

We distribute the *FAbulA* as a web app encapsulated on a Docker container. The docker is available at https://hub.docker.com/r/mesonepigreco/fabula The basic codes for the SSCHA and tdSSCHA methods are available as open-source packages at www.sscha.eu.

Received: 23 January 2024; Accepted: 20 August 2024;
Published online: 10 October 2024

References

1. Rachwalski, M. Special issue: asymmetry and symmetry in organic chemistry. *Symmetry* **15**, 1363 (2023).
2. Carpenella, V. et al. High-pressure behavior of δ-phase of formamidinium lead iodide by optical spectroscopies. *J. Phys. Chem. C.* **127**, 2440–2447 (2023).
3. Cherubini, M., Monacelli, L. & Mauri, F. The microscopic origin of the anomalous isotopic properties of ice relies on the strong quantum anharmonic regime of atomic vibration. *J. Chem. Phys.* **155**, 184502 (2021).
4. Monacelli, L., Errea, I., Calandra, M. & Mauri, F. Pressure and stress tensor of complex anharmonic crystals within the stochastic self-consistent harmonic approximation. *Phys. Rev. B* **98**, 024106 (2018).
5. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
6. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).

7. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
8. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
9. Smith, J. S. et al. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **7**, 134 (2020).
10. Fink, T. & Reymond, J.-L. Virtual exploration of the chemical universe up to 11 atoms of c, n, o, f: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **47**, 342–353 (2007).
11. Devereux, C. et al. Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.* **16**, 4192–4202 (2020).
12. Li, C. & Voth, G. A. Using machine learning to greatly accelerate path integral ab initio molecular dynamics. *J. Chem. Theory Comput.* **18**, 599–604 (2022).
13. Bocus, M. et al. Nuclear quantum effects on zeolite proton hopping kinetics explored with machine learning potentials and path integral molecular dynamics. *Nat. Commun.* **14**, 1008 (2023).
14. Chmiela, S. et al. Accurate global machine learning force fields for molecules with hundreds of atoms. *Sci. Adv.* **9**, eadf0873 (2023).
15. Monacelli, L. & Mauri, F. Time-dependent self-consistent harmonic approximation: anharmonic nuclear quantum dynamics and time correlation functions. *Phys. Rev. B* **103**, 104305 (2021).
16. Monacelli, L. et al. The stochastic self-consistent harmonic approximation: calculating vibrational properties of materials with full quantum and anharmonic effects. *J. Phys. Condens. Matter* **33**, 363001 (2021).
17. Siciliano, A., Monacelli, L., Caldarelli, G. & Mauri, F. Wigner gaussian dynamics: Simulating the anharmonic and quantum ionic motion. *Phys. Rev. B* **107**, 174307 (2023).
18. Long, D. A. The Raman Effect. John Wiley & Sons Ltd, 1–340 (2002). https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470845767.fmatter_indsub.
19. Popov, M. N. et al. Raman spectra of fine-grained materials from first principles. *npj Comput. Mater.* **6**, 121 (2020).
20. Giannozzi, P. et al. Advanced capabilities for materials modelling with quantum ESPRESSO. *J. Phys. Condens. Matter* **29**, 465901 (2017).
21. Mudalip, S. K. A., Bakar, M. R. A., Adam, F. & Jamal, P. Structures and hydrogen bonding recognition of mefenamic acid form i crystals in mefenamic acid ethanol solution. *Int. J. Chem. Eng. Appl.* **4**, 124–128 (2013).
22. Miranda, J. et al. Phase transformation in the c form of myristic-acid crystals and DFT calculations. *Spectrochim. Acta A: Mol. Biomol. Spectrosc.* **208**, 97–108 (2019).
23. Saggu, M., Liu, J. & Patel, A. Identification of subvisible particles in biopharmaceutical formulations using raman spectroscopy provides insight into polysorbate 20 degradation pathway. *Pharm. Res.* **32**, 2877–2888 (2015).
24. Errea, I. et al. Quantum crystal structure in the 250-kelvin superconducting lanthanum hydride. *Nature* **578**, 66–69 (2020).
25. Monacelli, L., Errea, I., Calandra, M. & Mauri, F. Black metal hydrogen above 360GPa driven by proton quantum fluctuations. *Nat. Phys.* **17**, 63–67 (2020).
26. Monacelli, L., Casula, M., Nakano, K., Sorella, S. & Mauri, F. Quantum phase diagram of high-pressure hydrogen. *Nat. Phys.* **19**, 845–850 (2023).
27. Kapil, V., Engel, E., Rossi, M. & Ceriotti, M. Assessment of approximate methods for anharmonic free energies. *J. Chem. Theory Comput.* **15**, 5845–5857 (2019).
28. Errea, I., Calandra, M. & Mauri, F. Anharmonic free energies and phonon dispersions from the stochastic self-consistent harmonic approximation: application to platinum and palladium hydrides. *Phys. Rev. B* **89**, 064302 (2014).
29. Bianco, R., Errea, I., Paulatto, L., Calandra, M. & Mauri, F. Second-order structural phase transitions, free energy curvature, and temperature-dependent anharmonic phonons in the self-consistent harmonic approximation: Theory and stochastic implementation. *Phys. Rev. B* **96**, 014111 (2017).
30. Miotto, M. & Monacelli, L. Entropy evaluation sheds light on ecosystem complexity. *Phys. Rev. E* **98**, 042402 (2018).
31. Miotto, M. & Monacelli, L. TOLOMEO, a novel machine learning algorithm to measure information and order in correlated networks and predict their state. *Entropy* **23**, 1138 (2021).
32. Lazzeri, M. & Mauri, F. First-principles calculation of vibrational raman spectra in large systems: signature of small rings in crystalline sio_2. *Phys. Rev. Lett.* **90**, 036401 (2003).
33. Bastonero, L. & Marzari, N. Automated all-functionals infrared and Raman spectra. *npj Comput. Mater.* **10**, 55 (2024).
34. Togo, A., Chaput, L., Tadano, T. & Tanaka, I. Implementation strategies in phonopy and phono3py. *J. Phys. Condens. Matter* **35**, 353001 (2023).
35. Irwin, J. J. et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.* **60**, 6065–6073 (2020).
36. van Uden, N. W. A. et al. Solvation pressure as real pressure: I. ethanol and starch under negative pressure. *J. Phys. Condens. Matter* **15**, 1577–1584 (2003).
37. De Gelder, J. Raman spectroscopy as a tool for studying bacterial cell compounds. *Ghent University* (2008).
38. Beć, K. B. et al. IR spectra of crystalline nucleobases: combination of periodic harmonic calculations with anharmonic corrections based on finite models. *J. Phys. Chem. B* **123**, 10001–10013 (2019).
39. Burikov, S., Dolenko, T., Patsaeva, S., Starokurov, Y. & Yuzhakov, V. Raman and IR spectroscopy research on hydrogen bonding in water-ethanol systems. *Mol. Phys.* **108**, 2427–2436 (2010).

Acknowledgements

L.M. acknowledges the European Union under the program H2020 for the MSCA individual fellowship, grant number 101018714.

Author contributions

M.M. and L.M. conceived the research, developed the tool, and wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01400-9.

Correspondence and requests for materials should be addressed to Mattia Miotto or Lorenzo Monacelli.