Molecular Screening for Terahertz Detection with Machine-Learning-Based Methods

Zsuzsanna Koczor-Benda*
Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom and Department of Chemistry, King's College London, London SE1 1DB, United Kingdom

Alexandra L. Boehmke, Angelos Xomalis, Rakesh Arul, Charlie Readman, and Jeremy J. Baumberg†
NanoPhotonics Centre, Cavendish Laboratory, Department of Physics, JJ Thompson Avenue, University of Cambridge, Cambridge CB3 0HE, United Kingdom

Edina Rosta‡
Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom and Department of Chemistry, King's College London, London SE1 1DB, United Kingdom

(Received 9 April 2021; revised 10 August 2021; accepted 7 September 2021; published 18 November 2021)

The molecular requirements are explored for achieving efficient signal up-conversion in a recently developed technique for terahertz (THz) detection based on molecular optomechanics. We discuss which molecular and spectroscopic properties are most important for predicting efficient THz detection and outline a computational approach based on quantum-chemistry and machine-learning methods for calculating these properties. We validate this approach by bulk and surface-enhanced Raman scattering and infrared absorption measurements. We develop a virtual screening methodology performed on databases of millions of commercially available compounds. Quantum-chemistry calculations for about 3000 compounds are complemented by machine-learning methods to predict applicability of 93 000 organic molecules for detection. Training is performed on vibrational spectroscopic properties based on absorption and Raman scattering intensities. Our top molecules have conversion intensity two orders of magnitude higher than an average molecule from the database. We also discuss how other properties like molecular shape and self-assembling properties influence the detection efficiency. We identify molecular moieties whose presence in the molecules indicates high activity for THz detection and show an example where a simple modification of a frequently used self-assembling compound can enhance activity 85-fold. The capabilities of our screening method are demonstrated on narrow-band and broadband detection examples, and its possible applications in surface-enhanced spectroscopy are also discussed.

DOI: 10.1103/PhysRevX.11.041035

Subject Areas: Atomic and Molecular Physics, Computational Physics

I. INTRODUCTION

Terahertz (THz) radiation has a high potential for applications in medical diagnostics, security screening, communication, astronomy, and many other fields [1–3]. The 0.1–30-THz range is often referred to as the THz gap, because the development of powerful yet affordable sources and efficient wideband detectors has been challenging for traditional electronics.

*z.koczor-benda@ucl.ac.uk
†jjb12@cam.ac.uk
‡e.rosta@ucl.ac.uk

The enhancement of Raman scattering signals in molecular nanocavities can potentially be harnessed in a recently proposed device for converting THz [or mid- and far-infrared (MIR and FIR, respectively)] radiation to visible or near-infrared (Vis and NIR, respectively) light [4,5], thus enabling optical detectors to be used for THz detection. To enhance the light-matter interaction, the molecules are placed in a set of two antennas operating on different scales [5,6]. A THz antenna focuses radiation at the design frequency over the molecular sample volume to enhance THz absorption via the surface-enhanced infrared absorption (SEIRA) [7] mechanism. A complementary optical antenna confines Vis or NIR light to $< 100\ \mathrm{nm}^3$ volumes, inducing surface-enhanced Raman scattering (SERS) [8] of molecules within the plasmonic nanocavity. Absorption of THz radiation by the molecules within the nanocavity causes vibrational excitation of a specific normal mode, which is then probed with a Vis or NIR laser (see Fig. 1). The increase in Raman

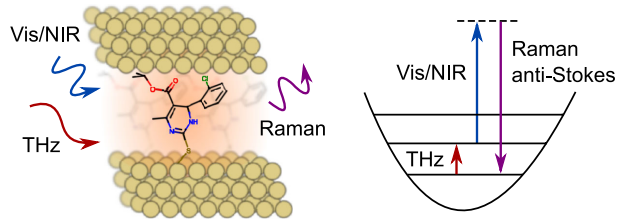

FIG. 1. The THz detection process. Absorption of THz radiation by a molecular vibrational mode in a molecular nanocavity increases population of excited vibrational levels, which is detected through the increased intensity of the SERS anti-Stokes signal of the vibrational mode in question.

anti-Stokes intensity of this mode signals the presence of THz radiation at the frequency of the normal mode. This is similar to the mechanism that is utilized in resonant sum-frequency generation (SFG) spectroscopy [9].

The detection technique requires strong simultaneous absorption and Raman activity for the vibrational mode [5]. For molecules with an inversion center, absorption and Raman activity of normal modes are mutually exclusive. However, there can be a significant simultaneous activity for nonsymmetric molecules, which means that careful selection of molecules is essential for the development of a highly efficient detector.

The optimal molecular design can be facilitated computationally by predicting accurate vibrational frequencies and intensities with cost-efficient methods. Additionally, secondary selection criteria incorporate requirements for experimental sample preparation and stability. A well-established method for depositing molecules in nanocavities is molecular self-assembly. On gold surfaces, which are often used in plasmonic devices, self-assembled monolayers (SAMs) of thiol-containing molecules provide high stability and reproducibility [10]. The capability of SAM formation and the cost of synthetic preparation also need to be considered for molecular selection. Here, we present a computational framework to optimally select molecules from databases with millions of compounds that are available for use in experimental applications. To enable the efficient prediction of molecular properties at this scale, we train machine-learning models on spectroscopic data from accurate quantum-chemistry (QC) calculations.

Machine learning (ML) for QC has seen rapid development in recent years, and now ML has become an invaluable tool in theoretical chemistry [11–13]. ML methods facilitate the discovery and design of new functional molecules and materials by enabling computational screening of millions of compounds [14] or generating new molecules [12,15] tailored for a specific application. ML methods have been successfully developed for the prediction of electronic properties of organic molecules [16–18] and transition-metal complexes [19] and spectroscopic properties such as electronic excitation spectra [20,21]. Highly accurate ML models have also been developed for accessing vibrational spectroscopic properties in molecular-dynamics calculations [22,23], including methods specifically for SERS applications [24]. These are, however, used for accelerating calculations for single compounds and, thus, not immediately applicable for molecular screening or chemical discovery. Predicting accurate vibrational frequencies, absorption, and Raman intensities for a large number of compounds remains a challenge computationally.

Here, we specifically develop an ML-based method that optimizes spectroscopic properties of molecules in plasmonic nanocavities for THz (MIR and FIR) detection applications. Our work focuses on a range of low-frequency vibrational modes between 1 and 30 THz and enables the selection of molecules with optimal THz conversion properties. Accordingly, we aim to address the following design aspects: (i) sensitivity—the ability of a molecule to detect THz radiation is assessed by ML models trained on results from QC calculations; (ii) integrability—molecular structure and potential self-assembling properties on a gold surface are investigated; (iii) availability—databases of commercially available compounds are explored.

We outline the underlying theory for the spectroscopic requirements of THz detection and then describe the computational framework for ML-based screening. We validate the QC approach by experimental THz absorption and Raman scattering measurements. Then, our ML results for database screening are analyzed and discussed, addressing the three design aspects and demonstrating specific narrow-band and broadband detection examples. We find that the highly efficient ML screening compares favorably to the costly QC calculations in terms of accuracy and that database screening can improve the desired spectroscopic properties by 2 orders of magnitude compared to randomly selected molecules. Finally, we discuss other possible applications of our database and screening method.

II. THEORY

The detector utilizes the increase in anti-Stokes intensity of a specific vibrational mode in the presence of THz radiation (Fig. 1) [5]. The anti-Stokes intensity for normal mode m is given by

$$I_m^{\mathrm{aS}} = nN\sigma_m^{\mathrm{aS}} I_L^{\mathrm{aS}}, \tag{1}$$

where n is the population of the vibration (Bose factor for thermal equilibrium), N is the number of molecules in the probed volume, σ_m^{aS} is the anti-Stokes cross section, and I_L^{aS} is the power density of the Vis/NIR laser used for inducing the Raman anti-Stokes effect. The presence of THz radiation at the frequency of the normal mode increases n through vibrational pumping, while various relaxation processes counteract this pumping. In the steady-state approximation, the change in population can be given by

$$\Delta n = \frac{\sigma_m^A I_L^A \tau_m}{h\nu^A}, \tag{2}$$

where σ_m^A is the absorption cross section, I_L^A and $h\nu^A$ are the power density and energy of the THz radiation and τ_m is the vibrational lifetime. This approximation is valid only if there is a moderate coupling between molecular vibration and THz light—in the strong coupling regime interesting new phenomena can occur [25,26], which could be utilized in other types of applications. In Eq. (2), indirect population gain through anharmonic effects (vibrational energy transfer between different modes) is considered to be negligible compared to direct pumping.

The increase in anti-Stokes intensity due to THz pumping can then be written as

$$\Delta I_m^{\mathrm{aS}} = \frac{I_L^A I_L^{\mathrm{aS}}}{h\nu^A} N \tau_m \langle \sigma_m^A \sigma_m^{\mathrm{aS}} \rangle. \tag{3}$$

In Eq. (3), the quantities dependent on the molecular properties are N, τ_m, σ_m^A, and σ_m^{aS}. In the current device, if we consider a well-structured SAM, N is inversely proportional to the surface area per molecule (S), which can be determined from experiments or, as in this work, estimated from geometrical parameters. The vibrational lifetime τ_m is determined by a range of relaxation processes, which are described in Ref. [5] for the current device. For simple molecules (e.g., CO and NO) physisorbed on metal substrates, τ_m varies from 1–2 ps to more than 10 ps [27]; however, it is nontrivial to predict how different vibrational modes behave, especially for more complex chemisorbed molecules as in our case. τ_m can also vary with the embedding of the molecules inside plasmonic cavities through the Purcell effect, and the modeling of this requires the calculation of electromagnetic contributions specific to the cavity and the exact positions of the molecules therein. Vibrational modes with large τ_m could possibly be targeted to enhance the sensitivity, but this is out of the scope of the current study.

In the current device, due to plasmonic effects, the THz and Raman in-out fields are all predominantly in the direction normal to the gold surfaces. Ideally, the cross sections in Eq. (3) would be averaged for the specific distribution of molecular orientations present in experiments [28]. However, detailed experimental studies on SAM structures are available only for a limited set of compounds, mostly alkyl-thiols and some simple aryl-thiols [29]. Even for these molecules, the optimal tilting of the molecular axis can vary from $5 \pm 5°$ to $49 \pm 5°$ [29]. For chemically more complex molecules, even less is known about the structure and the effects of depositing nanoparticles on top of the SAM, surface reconstruction, or stability [30]. Computational modeling of SAMs is also challenging [31], and it usually requires prior knowledge of packing parameters. Without access to wide-ranging experimental data or a fast and reliable method to predict orientation distributions for molecules in our database, we resort to a random distribution of molecular orientations, denoted by angle brackets in Eq. (3). Incubation time, temperature, solvent, etc., can all influence the distribution of orientations, and these will have to be optimized experimentally for the best molecules, to achieve a stable monolayer with the highest possible conversion efficiency.

In this paper, we focus on the molecular contributions to the product of Raman and absorption cross sections, which can be separated into the field enhancement E and the normal mode-dependent conversion intensity I_m^c:

$$\langle \sigma_m^A \sigma_m^{\mathrm{aS}} \rangle \sim E(g, n_M) I_m^c. \tag{4}$$

While E is dependent on several material and geometric factors of the THz and optical antennas, here we show only dependencies that are affected by the choice of molecule: the refractive index of the medium n_M and thickness g of the material in the nanogap. To achieve optimal enhancement of THz absorption, the THz antenna needs to be tuned to the molecular vibration of interest [7]. Regarding the enhancement of the Raman signal, we use the approximation [32] that the SERS maximum field enhancement is proportional to g^{-2}. Both absorption and Raman enhancements are affected by n_M, which is, however, expected to vary little between organic molecules of similar size and SAM properties.

I_m^c is calculated as

$$I_m^c = C \frac{(\bar{\nu}^{\mathrm{aS}} + \bar{\nu}_m)^4}{\bar{\nu}_m} \langle |\underline{e}\underline{\mu}'_m|^2 |\underline{e}\underline{\underline{\alpha}}'_m \underline{e}|^2 \rangle, \tag{5}$$

where C is a constant scaling factor (given in Sec. S1 in Supplemental Material [33]), $\bar{\nu}^{\mathrm{aS}}$ and $\bar{\nu}_m$ are the wave numbers of the Vis laser and the normal mode, respectively, $\underline{\mu}'_m$ is the dipole derivative vector, and $\underline{\underline{\alpha}}'_m$ is the polarizability derivative tensor. The aligned THz and Raman in-out field polarization vectors are all denoted by $\underline{e}$. Note that I_m^c is connected to the increase in anti-Stokes signal due to THz pumping, and as such it does not consider a thermal population but rather a population increase proportional to THz absorption intensity. The analytical formula for calculating the molecular orientation average in angle brackets is given in the Supplemental Material, Sec. S2 [33]. We note that Ref. [5] uses a similar quantity $\langle \eta_{\mathrm{pol}} \rangle = \langle |\underline{e}\underline{\mu}'_m|^2 |\underline{e}\underline{\underline{\alpha}}'_m \underline{e}|^2 \rangle / ||\underline{\mu}'_m||^2 ||\underline{\underline{\alpha}}'_m||^2$ referred to as the (orientation-averaged) local overlap of IR and Raman fields. Our definition does not separate the local overlap from the sensitivity of a normal mode to THz and Vis or NIR fields, resulting in a single quantity suitable for ranking normal modes for THz detection. We also benefit from the analytical calculation of the average, in contrast to numerical calculation of $\langle \eta_{\mathrm{pol}} \rangle$.

The conversion process is normal mode specific; thus, conversion intensities are additive. To rank a specific molecule for THz conversion, we consider all of its normal modes in the relevant frequency range in a target property defined as

$$P = \log\left(\sum_{m\in M} I_m^c\right), \tag{6}$$

where M is the set of vibrational normal modes of the molecule in the 30–1000 cm^{-1} (1–30 THz) range. P is then standardized with respect to the randomly selected molecules, so that it is in units of standard deviation of the random set (σ). We note that P appears to follow a normal distribution for the randomly selected set of molecules (Fig. 3). We also define similar absorption (A) and Raman Stokes scattering (R) target properties (see Supplemental Material, Sec. S3 [33]).

Apart from selecting molecules based on high P values, other molecular properties also need to be considered to achieve a significant anti-Stokes intensity increase. To maximize E and N, it is optimal to ensure that a single closely packed molecular layer is formed and to limit g to a couple of nanometers. This means that molecular geometry, SAM structure, and stability are also important factors when selecting molecules for the detector. As we show later, complex molecules are significantly more promising for the current application than commonly used SAMs, due to the high THz conversion rates they can achieve, but detailed experimental and computational investigations of their self-assembling properties are needed.

Other than the above-mentioned points, the availability of a compound also has to be considered. Compounds that are readily available commercially or are easy to synthesize are preferred to facilitate production.

III. COMPUTATIONAL DESIGN FRAMEWORK

We start by building an initial database from commercially available compound databases by screening for molecules containing a single thiol group, which facilitates adsorption on gold surfaces. The eMolecules database [34] contains 18 million in-stock and back-ordered compounds currently, of which 150 000 contain one thiol group. The MolPort database [35] contains 7 million in-stock compounds currently, of which 32 000 are monothiols. After removing duplicates and applying further restrictions on the size (<3 nm) and flexibility (<4 rotatable bonds), our database contains 93 000 unique compounds in total. For experimental trials, we also consider well-established self-assembling molecules from Sigma Aldrich [36] (about 20 compounds after prescreening). A notable difference from previous ML studies is the type of molecules that are included in the databases probed here. For our goals, it is essential to consider molecules that are known to be synthesizable and have an affinity to gold surfaces used in detector prototypes. This is very different from databases like QM9 [37] usually used as a benchmark in ML studies, as we have a much smaller number of compounds but with larger size (16 ± 4 nonhydrogen atoms compared to maximally nine in QM9) and consisting of a wider range of elements (S, Cl, Br, I, and P in addition to C, H, O, N, and F in QM9).

We address self-assembling properties of molecules in the database by measuring their similarity to known self-assembly materials. The similarity score s gives the maximum molecular similarity between the candidate molecule and a library of 110 known SAM materials (see Supplemental Material, Sec. S8 [33], for more information). We investigate the spectroscopic changes due to molecular orientation in the nanocavity for a set of test molecules and leave a detailed investigation of orientation effects for future studies. For estimating g and S, we use 3D molecular geometry and the approximation that all tested molecules have a similar orientation (perpendicular to the surface; see Supplemental Material, Sec. S6.2 [33]).

QC calculations are performed on 1300 randomly selected molecules from the database of 93 000. This QC database is then extended in several rounds to 3000 molecules by including those with the highest predicted activity based on our early ML models and other selection criteria. Density functional theory (DFT) calculations are performed at the B3LYP/def2-SVP level for molecules that have the thiol hydrogen atom exchanged to a single gold atom which is a sufficient first approximation of binding to the gold surface, as discussed in Sec. V. Further details of the computational methods are given in Supplemental Material, Sec. S4 [33].

ML training is performed for target properties P, A, and R separately, with elastic net (EN) and kernel ridge regression (KRR) with a Laplacian kernel. Feature selection with Lasso is applied to Morgan fingerprints [38] of various radii and lengths, and hyperparameter optimization of feature generation and model parameters is performed to yield best prediction accuracy (see also Supplemental Material, Sec. S4 [33]). Using trained ML methods, predictions are made for the unseen molecules from the database, and the best candidates are chosen for additional QC investigation based on their predicted P value. The ML training and prediction processes, as well as the augmentation of the QC database by the best candidates, are repeated a few times during the development of the screening method.

IV. EXPERIMENTAL VALIDATION

To validate the computational methodology, measurements are performed on a set of test molecules. As the THz detector designs are still being developed, up-converted detection cannot yet be evaluated; therefore, the absorption and Raman scattering experiments are performed separately. SERS measurements are performed for SAMs formed inside a plasmonic nanocavity. This nanocavity consists of a gold nanoparticle coupled to its mirror-image charges in a gold

mirror, forming a virtual dimer. A dielectric spacer layer made of a SAM of molecules adsorbed to the gold mirror yields a consistent gap size. Each NP in this nanoparticle-on-mirror (NPOM) geometry is interrogated individually through a microscope. SAM molecules located in the gap of the NPOM interact with the plasmonically enhanced optical field and chemically with the gold, producing SERS [32,39]. SERS measurements on NPOM structures provide a good approximation to the THz device architecture, but THz-frequency molecular vibrations are environmentally sensitive and there are many unknown factors of the state of molecules in SAMs, so as a fast first validation stage, we also measure Raman and absorption spectra in powder and solution phases.

Experimental methods are detailed in Supplemental Material, Sec. S5 [33]. To account for variations in overall intensity, multiple measurements per compound are performed. The spectral angle is used as a distance measure between measured and calculated spectra. In order to analyze measurements and enhance similarity of measured and calculated spectra, a range of postprocessing techniques are applied (see Supplemental Material, Sec. S6 [33], for details). Automatic frequency scaling is performed based on spectral distance. A clustering approach is applied to screen out contaminated or otherwise outlier experimental spectra. For modeling powder and solution measurements, the best-matching chemical form of each molecule is chosen from the various tautomers, ionic forms, etc., considered in the calculations. To correct for the number of molecules giving the measured signal, we use concentration (c) for solution measurements, molecular cell volume (V) for powder measurements, and occupied surface area (S) for NPOM measurements. The measured spectra are background corrected and integrated over the recorded spectral range (150–1200 cm^{-1} for powder, 110–1220 cm^{-1} for solution, 110–1200 cm^{-1} for NPOM Raman, and 580–1700 cm^{-1} for powder absorption). To check the correlation between measurements and calculations, we compare integrated Raman intensities

$$\tilde{R} = \int_{\bar{\nu}_1}^{\bar{\nu}_2} I^R d\bar{\nu}, \tag{7}$$

where I^R is the measured or calculated Raman Stokes intensity and $\bar{\nu}_1$ and $\bar{\nu}_2$ are, respectively, the minimum and maximum recorded wave numbers in experiments.

V. EXPERIMENTAL RESULTS

Comparison of measured spectra with calculations for the test molecules shows that characteristics of the Raman spectra are very well described by the current computational methods (see Supplemental Material, Sec. S6 [33], for individual spectra and spectral distances). Analyzing the spectral distance for narrower frequency ranges of the solution spectra reveals that the match between experiment and calculations is similarly accurate across most of the frequency range, with larger deviations only in the lowest frequency range (110–269 cm^{-1}). This is not surprising, as low-frequency modes are easily perturbed by the environment, and this effect is neglected in the current calculations. Calculations with a larger basis set (aug-cc-pVTZ) and polarizable continuum solvent model provide smaller spectral distances from solution-phase experiments (0.256 on average, compared to 0.298 with the def2-SVP basis set). The aug-cc-pVTZ basis set partially corrects the overestimation of low-wave-number mode intensities by the def2-SVP basis set.

A comparison of integrated Raman Stokes intensities between experiment (solution and NPOM) and calculation are shown in Figs. 2(a) and 2(b), while powder data are shown in Supplemental Material, Sec. S6.5 [33]. Similar plots for subranges of the solution spectra are given in Supplemental Material, Sec. VI. 3 [33]. Integrated intensities of solution measurements have low variance, with highest variations in the lowest frequency range (110–269 cm^{-1}). R^2 scores are above 0.80 for all subranges except 110–269 cm^{-1} and 427–586 cm^{-1}, which have 0.38 and 0.23, respectively. In the full frequency range, there is also a strong correlation between experiment and calculation [$R^2 = 0.81$, Fig. 2(a)]. Interestingly, the aug-cc-pVTZ basis set does not improve the correlation between integrated intensities ($R^2 = 0.79$ for the full frequency range; see Supplemental Material, Sec. VI. 3 [33]), which indicates that def2-SVP is able to reproduce the differences in integrated intensities between molecules sufficiently. Calculations with explicit solvent might be required to account for some of the remaining discrepancies in peak positions and intensities. This is expected to be most important for molecules that can form H bonds with the solvent but is out of the scope of the current study.

The current computational model provides a remarkably accurate match with NPOM spectra [Fig. 2(d)], with a mean spectral distance of 0.170 (0.157 with the aug-cc-pVTZ basis), which indicates that a single gold atom recovers the most immediate chemical effects of the molecular monolayer formed on the gold surface, consistent with previous studies [40]. The spectral distance is somewhat larger for molecule 8 than for the other molecules. This is due partly to the lower signal-to-noise ratio of the experimental spectrum and high variations of spectral features between measurements, which suggest that its SAM is not as uniform as the others.

Computational studies of thiophenol investigate the effect of including gold clusters [41] and slabs with periodic boundary conditions for different binding sites [42] on the vibrational spectrum. Our result with a single gold atom gives a spectrum very similar to the best matches with SERS experiments in these studies. Coordination geometries have the largest effect on the Raman intensities of 200–400 cm^{-1} modes, that have significant contributions from the sulfur atom [41]. Some of the main

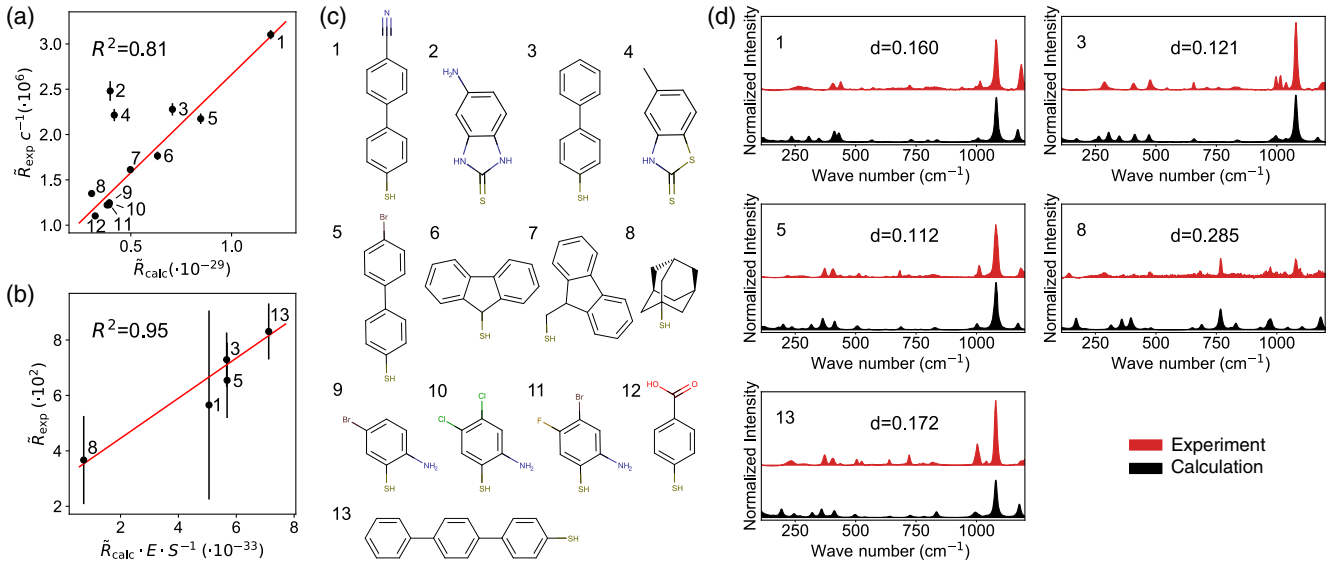

FIG. 2. Comparison of calculated and experimental Raman Stokes properties of a set of test molecules. (a) Corrected solution-phase ($\tilde{R}_{exp}/c$) vs calculated ($\tilde{R}_{calc}$) integrated intensities. (b) NPOM measured ($\tilde{R}_{exp}$) vs corrected calculated ($\tilde{R}_{calc}ES^{-1}$) integrated intensities. (c) Measured molecules. (d) Mean experimental and calculated spectra for NPOM systems along with spectral distances (d) between experiment and calculation.

discrepancies we observe between calculations and experiments are in this wave-number range, and we expect that simulations with gold clusters or slabs could resolve these. On the other hand, these investigations would further complicate orientation studies and database screening; thus, it is best to perform these investigations only on the highest-ranked molecules. Any remaining differences between the measured and calculated spectra are expected to be due to intermolecular interactions in the SAM, especially at low wave numbers. This can be investigated with molecular-dynamics simulations for specific molecules with known SAM structures, but, due to the lack of information about SAMs and high costs of simulations, such approaches are currently not available for the large-scale study of our database.

With corrections for E and S, the current computational model employing orientation-averaged quantities can predict integrated intensities very well [$R^2 = 0.95$ for both basis sets; see Fig. 2(b) for def2-SVP results]. Based on previous studies, molecules 3 and 13 are estimated to have a tilting angle of about 22° on gold [43], and molecules 1 and 5 are expected to have a similar tilt. To investigate the effects of molecular orientation, we determine Raman intensities for 0°- and 22°-tilted molecules as well (see Supplemental Material, Sec. VI. 4 [33]). We find that all four aromatic molecules are influenced similarly by tilting angle, with a larger decrease in $\tilde{R}_{calc}$ but practically unaffected spectral features. This behavior is explained by the high anisotropy of their polarizabilities (281–378 $bohr^3$), with highest polarizability along the long molecular axis (0° tilt). On the other hand, $\tilde{R}_{calc}$ of molecule 8 is much less sensitive to changes in tilting, while its spectral features vary more with tilting angle; this is in line with its low anisotropy (70 $bohr^3$). The similar, rodlike structure of molecules 1, 3, 5, and 13 causes a pronounced decrease of gap size with tilting angle, while the gap size of bulky molecule 8 is less sensitive to tilting. Agreement between $\tilde{R}_{exp}$ and corrected $\tilde{R}_{calc}$ for 0° and 22° tilts is similar to the full orientation-averaged case, with R^2 scores of 0.92 and 0.93, respectively. Ongoing measurements of a larger set of chemically diverse molecules will be able to validate the computational method for simulating NPOM systems further.

The existence of normal modes that are intensive in both absorption and Raman processes is verified by the comparison of powder measurements with simulations. Powder measurements are considerably easier to perform than solution or NPOM measurements, but comparison with simulations on single molecules is not expected to give very good agreement, as environmental effects and molecular orientation in crystal structures are not accounted for. Nevertheless, calculations for the thiol(-SH) forms provide a reasonable match with powder Raman spectra (mean distance 0.332) and integrated intensities ($R^2 = 0.63$). The agreement can be improved by simulating the crystal environment with molecular dimers and identifying the best-matching chemical forms (mean distance = 0.295, $R^2 = 0.68$; see Supplemental Material, Sec. VI. 5 [33]). For absorption spectra, the highly varying peak widths make it harder to compare with calculations, but mean distance is also reduced in this case, from 0.273 to 0.238. From DFT calculations, we identify normal modes that have high conversion intensity and confirm the presence of these peaks in the measured absorption and Raman spectra (see Supplemental Material, Sec. VI. 5 [33]).

VI. MACHINE-LEARNING RESULTS

The best training results achieved for the three target properties with EN and KRR models are compiled in Table I. The best performance is observed for A, followed by P and R. The mean absolute errors (MAE) for the test set are low enough to provide high-quality predictions for all three targets, considering that our QC database covers ranges of about $8-10\sigma$. The nonlinear KRR model does not seem to improve results of the linear EN model considerably.

Best candidates for THz detection are found around 6σ (Fig. 3), which corresponds to a 2-orders-of-magnitude increase in conversion intensity compared to random molecules (note that P is a logarithmic, standardized quantity). Results for the known SAM-forming molecules of the Sigma Aldrich database are also shown in Fig. 3. These are chemically much simpler molecules than those in the eMolecules and MolPort databases: predominantly aryl thiols like 1,1'-biphenyl-4-thiol [(BPT); see Fig. 3(b)] and alkyl thiols like butanethiol [Fig. 3(a)]. These types of molecules form stable and well-structured SAMs [29], and, thus, they are often used in SERS experiments. Regarding their applicability for THz detection, Fig. 3 shows that BPT has slightly above-average performance, and butanethiol is not suitable for the detector at all. In comparison, the best molecules from the QC database [Figs. 3(c)–3(f)] have about 150–340 times higher conversion intensity than average SAM materials and 70–150 times higher intensity than an average molecule from the 93 000 database, which would be used if computational screening were not performed.

Since the molecular fingerprint used as the feature vector provides only 2D information on the molecules, the ML models are not able to give separate predictions for conformers of the same molecule. Absorption and Raman spectra are, however, sensitive to conformational changes. Even though we limit the number of rotatable bonds to < 4, several conformers can exist, and the conformer used in DFT calculations might not match the dominant conformer in experiments. Analyzing calculations on two conformers of 467 molecules, we find that the difference in P is 0.8 ± 0.7 between conformers. This might result in significantly different ordering of top molecules. Determining the most stable conformers of candidate molecules and corresponding spectral properties in a nanocavity is subject to future work.

TABLE I. Performance of different ML models on the three target quantities, given as R^2 score and MAE (in units of σ) for the test set.

	A		R		P	
ML model	R^2	MAE	R^2	MAE	R^2	MAE
KRR	0.90	0.27	0.60	0.49	0.75	0.57
EN	0.88	0.29	0.60	0.49	0.73	0.60

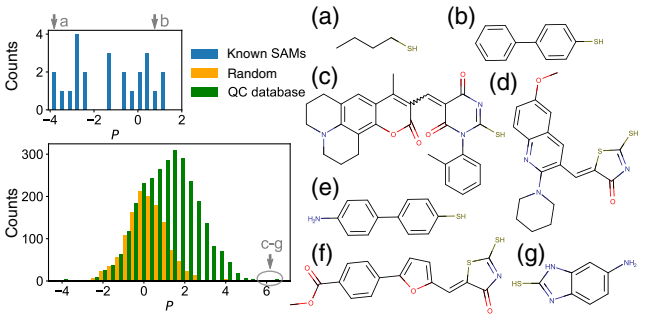

FIG. 3. Distribution of P (standardized logarithm of summed conversion intensity) among (blue) known self-assembly materials, (orange) randomly chosen molecules, and (green) the QC database. Some molecules with (a) low, (b) moderate, and (c)–(g) high potential for THz detection are also depicted.

Training linear EN models enables us to determine which molecular fragments influence the predicted properties the most by analyzing weights of features in the trained models (Fig. 4). We note that a large weight does not necessarily mean that the fragment in question is spectroscopically highly active, just that in our current database the presence of the fragment correlates with high intensity. The most prominent fragments for A and R are consistent with the nature of absorption and Raman processes: polar bonds for A and highly polarizable aromatic moieties for R. P shares some of its most important fragments with A and R. The optimal fingerprint radius is 1 for A, which means that only atoms within maximally one bond distance from the starting atom are considered; this again shows that absorption intensities are mainly bond specific, while it is 2 for R and P, showing that larger environments are needed to assess Raman intensities. It can be verified by looking at the vibrations of the top molecules, that, e.g., the amine group is among the most THz-conversion active groups. Our trained linear EN models, therefore, provide valuable information for designing new molecules to be synthesized,

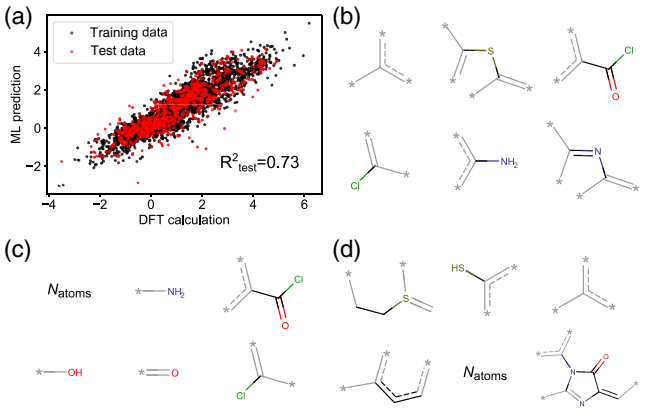

FIG. 4. (a) ML results for target P with the EN model. Features with largest weights in the trained EN model for (b) THz conversion P, (c) absorption A, and (d) Raman R targets.

TABLE II. Properties relevant in selecting molecules for THz detection, shown for the top five molecules from the DFT database [depicted in Figs. 3(c)–3(g)].

Mol	P	P_{max}	S (Å^2)	g (Å)	a (bohr^3)	s
(c)	6.18	8.34	129	17.3	446	0.28
(d)	5.95	8.16	127	14.6	373	0.32
(e)	5.60	7.84	31	13.7	315	0.85
(f)	5.41	7.62	54	17.4	540	0.41
(g)	5.30	7.47	27	11.1	271	0.68

as performance can potentially be enhanced by addition of the most active functional groups. This is highlighted by the example of BPT, whose conversion intensity can be increased 85-fold by a simple substitution to an amine group at position 4′ [Fig. 3(e)].

For selecting the best candidates, apart from the P value, we also have to consider geometrical features, that can influence N and E considerably. Table II compiles the most relevant properties of the top five molecules [molecules (c)–(g) in Fig. 3].

Geometrical properties S and g are calculated for 0° tilting angles, and thus are only crude estimations, but the table already shows that significant differences are expected in N and E between the top molecules. It can be more beneficial to prioritize molecules with low S and g, such as molecules (e) and (g), instead of choosing candidates solely based on the P value. Molecules with high anisotropy a are high risk, high reward, as they would need to be fixed in a specific orientation to take full advantage of their highest possible Raman cross sections, but might result in very small cross sections if the orientation is not right. The highest achievable P value for a specific orientation (P_{max}) also shows that, if the orientation can be optimized in experiments, even higher conversion efficiency can be achieved (see Supplemental Material, Sec. S7 [33], for orientation dependence of P). The ordering of P_{max} is consistent with that of P, showing that both quantities could be used to select top molecules. For molecules with a high s score [e.g., molecules (e) and (g)], SAM formation is likely, and preparation techniques can be learned from the literature of the most similar SAM materials (see Supplemental Material, Sec. S8 [33]). In contrast, a low s score does not imply that the molecule is not suitable for SAM formation, just that similar types of molecules have not been tested yet. Experimental proof of the superiority of the candidates selected by the screening method requires a device for testing THz conversion, which is currently under development.

At this exploratory phase of development, the first detector prototypes will operate at frequencies where the top molecules have highest conversion efficiency in order to maximize the effect. With the use of our screening methodology, however, it is also possible to choose the most suitable molecules for detection of specific THz light sources and in this way tailor the device for various fields of application. Here, we give examples for two types of THz application: narrow-band and broadband detection. For the narrow-band detection of atomic oxygen in planetary atmospheres, its atomic transition at 4.745 THz (185.3 cm^{-1}) has been used previously [44]. We train an ML model on a narrow frequency window around this transition (Supplemental Material, Sec. S9 [33]). When defining a target property for such a narrow range, it is more beneficial to integrate broadened conversion intensities for the frequency range [$\tilde{P}$, Eq. (S14) [33]] instead of summing discrete intensities. This is due to a large proportion of molecules having no transitions in this spectral range. The MAE of predictions is higher (0.68σ) than for the 30–1000-cm^{-1} frequency range used before, but the ML model can still differentiate between highly active and inactive molecules. We identify the top molecules for oxygen detection in Supplemental Material, Sec. S9 [33].

For broadband detection, not every spectral range is suitable, as the characteristic vibrational transitions of organic molecules do not cover the spectrum uniformly. We select two ranges—30–180 cm^{-1} (0.9–5.4 THz) and 600–800 cm^{-1} (18–24 THz), where transitions are more dense—and apply peak broadening to model experimental conditions. ML predictions for $\tilde{P}$ are of similar quality for the two ranges (MAE around 0.5σ), and separately trained ML models for predicting the spectral flatness give MAEs of about 0.1 (see Supplemental Material, Sec. S10 [33], for details). By screening for simultaneously high values of $\tilde{P}$ and spectral flatness, top candidates for broadband applications can be identified (see Supplemental Material, Sec. S10 [33]). One might need to compromise on a lower sensitivity of the device to achieve broadband detection compared to narrow-band detection, but the overall sensitivity will also be highly dependent on the specific construct of the device. We note that mixed SAMs or an ensemble of nanocavities with different molecules could also be used for broadband detection if they are carefully engineered to complement each other.

VII. FURTHER APPLICATIONS

Other than for THz detection, our database and screening method can potentially be used for other fields of applications. To mention a few promising examples, our method can be utilized to maximize absorption intensities at frequencies where high-intensity lasers are available and in this way enable reaching the vibrational strong coupling regime. In this regime, tuning the frequency of vibrational modes in order to control chemical reactions and rates [25,45] or creating a Raman-laser-based optical parametric oscillator producing coherent MIR beams [26] becomes possible. The latter would need vibrational modes that are active in both IR and Raman, very similar to our screening criteria for THz detection.

Investigation of collective effects in SERS could also be facilitated by molecular screening: The required laser intensity that induces these effects can be reduced by careful design, and one of the key aspects is to have a vibrational mode with high Raman intensity [46]; for this purpose, our database and ML model would be highly useful.

Once a detector prototype is available, this would also open up a new way for investigating the properties of molecules inside the nanogap. This would mean that properties influencing the conversion efficiency of the detector, e.g., molecular orientation, conformations, and vibrational lifetimes, could be assessed experimentally. A combination of IR, Raman, and SFG measurements would be highly useful for determining orientation distributions [28], and, therefore, carefully selected molecules could also function as local probes of surface structure.

VIII. CONCLUSIONS

We developed a machine-learning-based computational method for predicting the vibrational properties of molecules and selecting the best candidates for THz detection from a database of commercially available compounds. The combination of quantum-chemistry calculations and machine-learning methods provides accurate predictions and saves time and cost when assessing molecules of the database.

The quantum-chemistry method was validated for a range of compounds by powder, solution, and nanoparticle-on-mirror Raman measurements. It was shown that most spectral features and integrated Raman intensities are accurately predicted at the current level of modeling. Absorption measurements confirmed the presence of vibrational modes highly active in both absorption and Raman scattering. Trained machine-learning models have shown good accuracy of predictions for absorption, Raman scattering, and THz conversion target properties. Molecular screening of the database gives candidates with 2-orders-of-magnitude larger THz-to-Vis and NIR conversion intensity than molecules typically used in similar (surface-enhanced) experimental setups. Although trained for a specific database of compounds, the predictive power of the ML model can potentially be extended to other types of compounds, which needs to be verified by computations. Functional groups found to correlate with higher conversion efficiency within the database can potentially be used to enhance the intensity elicited by commonly used molecules, as we have shown on the example of BPT. We have also discussed how geometrical factors and other molecular properties can influence applicability of molecules for THz detection and how they can be exploited for the design of highly efficient devices.

We demonstrated the strength of our method to provide candidate molecules for narrow-band and broadband detection applications and discussed how the molecular screening method can be highly useful for investigating vibrational strong coupling and collective SERS effects or probing surface structures.

In light of the recent experimental demonstrations of the up-conversion effect [47,48], we believe that identifying the best molecules for THz ranges will significantly facilitate the next steps for realizing an efficient THz detector based on molecular optomechanics in the near future.

To inspire and facilitate the development of THz and other applications, we share our QC database along with interactive plotting and analysis tools as a web application; Molecular Vibration Explorer [49].

ACKNOWLEDGMENTS

We are grateful for stimulating discussions with Philippe Roelli and Christophe Galland. Z. K.-B. thanks Balint Koczor for technical help with analytic integration of orientation averages. We acknowledge funding from the European Research Council (ERC) under Horizon 2020 research and innovation program THOR (Grant Agreement No. 829067). Z. K.-B. and E. R. also acknowledge funding from EPSRC (EP/R013012/1, EP/L027151/1) and ERC Project No. 757850 BioNet. We are grateful to the United Kingdom Materials and Molecular Modelling Hub for computational resources, which is partially funded by EPSRC (EP/P020194/1). A. X., A. L. B., R. A., C. R., and J. J. B. acknowledge support from ERC under Horizon 2020 research and innovation programs POSEIDON (Grant Agreement No. 861950) and PICOFORCE (No. 883703). We also acknowledge funding from the EPSRC (Cambridge NanoDTC EP/L015978/1, EP/L027151/1, EP/S022953/1, EP/P029426/1, and EP/R020965/1). R. A. acknowledges funding from the Royal Society Te Aprangi-Rutherford Foundation and the Winton Programme for the Physics of Sustainability.

[1] M. Tonouchi, *Cutting-Edge Terahertz Technology*, Nat. Photonics **1**, 97 (2007).
[2] S. S. Dhillon *et al.*, *The 2017 Terahertz Science and Technology Roadmap*, J. Phys. D **50**, 043001 (2017).
[3] J.-S. Rieh, *Introduction to Terahertz Electronics* (Springer, New York, 2021).
[4] P. Roelli, C. Galland, N. Piro, and T. J. Kippenberg, *Molecular Cavity Optomechanics as a Theory of Plasmon-Enhanced Raman Scattering*, Nat. Nanotechnol. **11**, 164 (2016).
[5] P. Roelli, D. Martin-Cano, T. J. Kippenberg, and C. Galland, *Molecular Platform for Frequency Up-Conversion at the Single-Photon Level*, Phys. Rev. X **10**, 031057 (2020).
[6] A. Xomalis, X. Zheng, A. Demetriadou, A. Martinez, R. Chikkaraddy, and J. J. Baumberg, *Interfering Plasmons in Coupled Nanoresonators to Boost Light Localization and SERS*, Nano Lett. **21**, 2152 (2021).

[7] F. Neubrech, C. Huck, K. Weber, A. Pucci, and H. Giessen, *Surface-Enhanced Infrared Spectroscopy Using Resonant Nanoantennas*, Chem. Rev. **117**, 5110 (2017).

[8] P. L. Stiles, J. A. Dieringer, N. C. Shah, and R. P. Van Duyne, *Surface-Enhanced Raman Spectroscopy*, Annu. Rev. Anal. Chem. **1**, 601 (2008).

[9] C. Humbert, T. Noblet, L. Dalstein, B. Busson, and G. Barbillon, *Sum-Frequency Generation Spectroscopy of Plasmonic Nanomaterials: A Review*, Materials **12**, 836 (2019).

[10] V. Chechik and C. J. M. Stirling, in *PATAI'S Chemistry of Functional Groups*, edited by Z. Rappoport (Wiley, New York, 2009).

[11] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Machine Learning for Molecular and Materials Science*, Nature (London) **559**, 547 (2018).

[12] B. Sanchez-Lengeling and A. Aspuru-Guzik, *Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering*, Science **361**, 360 (2018).

[13] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, *Machine Learning for Molecular Simulation*, Annu. Rev. Phys. Chem. **71**, 361 (2020).

[14] R. Gómez-Bombarelli *et al.*, *Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach*, Nat. Mater. **15**, 1120 (2016).

[15] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, *Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules*, ACS Cent. Sci. **4**, 268 (2018).

[16] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, *Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space*, J. Phys. Chem. Lett. **6**, 2326 (2015).

[17] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, *Quantum-Chemical Insights from Deep Tensor Neural Networks*, Nat. Commun. **8**, 1 (2017).

[18] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, and M. Ceriotti, *Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning*, Proc. Natl. Acad. Sci. U.S.A. **116**, 3401 (2019).

[19] J. P. Janet and H. J. Kulik, *Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks*, Chem. Sci. **8**, 5137 (2017).

[20] R. Ramakrishnan, M. Hartmann, E. Tapavicza, and O. A. Von Lilienfeld, *Electronic Spectra from TDDFT and Machine Learning in Chemical Space*, J. Chem. Phys. **143**, 084111 (2015).

[21] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke, *Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra*, Adv. Sci. **6**, 1801367 (2019).

[22] M. Gastegger, J. Behler, and P. Marquetand, *Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra*, Chem. Sci. **8**, 6924 (2017).

[23] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, *Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields*, Nat. Commun. **9**, 3887 (2018).

[24] W. Hu, S. Ye, Y. Zhang, T. Li, G. Zhang, Y. Luo, S. Mukamel, and J. Jiang, *Machine Learning Protocol for Surface-Enhanced Raman Spectroscopy*, J. Phys. Chem. Lett. **10**, 6026 (2019).

[25] A. Shalabney, J. George, J. A. Hutchison, G. Pupillo, C. Genet, and T. W. Ebbesen, *Coherent Coupling of Molecular Resonators with a Microcavity Mode*, Nat. Commun. **6**, 5981 (2015).

[26] J. del Pino, F. J. Garcia-Vidal, and J. Feist, *Exploiting Vibrational Strong Coupling to Make an Optical Parametric Oscillator out of a Raman Laser*, Phys. Rev. Lett. **117**, 277401 (2016).

[27] V. Krishna and J. C. Tully, *Vibrational Lifetimes of Molecular Adsorbates on Metal Surfaces*, J. Chem. Phys. **125**, 054706 (2006).

[28] K.-K. Hung, U. Stege, and D. K. Hore, *IR Absorption, Raman Scattering, and IR-vis Sum-Frequency Generation Spectroscopy as Quantitative Probes of Surface Structure*, Appl. Spectrosc. Rev. **50**, 351 (2015).

[29] J. C. Love, L. A. Estroff, J. K. Kriebel, R. G. Nuzzo, and G. M. Whitesides, *Self-Assembled Monolayers of Thiolates on Metals as a Form of Nanotechnology*, Chem. Rev. **105**, 1103 (2005).

[30] C. Vericat, M. Vela, G. Benitez, P. Carro, and R. Salvarezza, *Self-Assembled Monolayers of Thiols and Dithiols on Gold: New Challenges for a Well-Known System*, Chem. Soc. Rev. **39**, 1805 (2010).

[31] G. Heimel, F. Rissner, and E. Zojer, *Modeling the Electronic Properties of π-Conjugated Self-Assembled Monolayers*, Adv. Mater. **22**, 2494 (2010).

[32] J. J. Baumberg, J. Aizpurua, M. H. Mikkelsen, and D. R. Smith, *Extreme Nanophotonics from Ultrathin Metallic Gaps*, Nat. Mater. **18**, 668 (2019).

[33] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.11.041035 for computational and experimental methodology, and detailed data analysis.

[34] eMolecules database, https://www.emolecules.com.

[35] MolPort database, https://www.molport.com.

[36] Sigma Aldrich, https://www.sigmaaldrich.com.

[37] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, *Quantum Chemistry Structures and Properties of 134 Kilo Molecules*, Sci. Data **1**, 1 (2014).

[38] D. Rogers and M. Hahn, *Extended-Connectivity Fingerprints*, J. Chem. Inf. Model. **50**, 742 (2010).

[39] S. E. J. Bell, G. Charron, E. Corts, J. Kneipp, M. L. de la Chapelle, J. Langer, M. Prochzka, V. Tran, and S. Schlcker, *Towards Reliable and Quantitative Surface-Enhanced Raman Scattering (SERS): From Key Parameters to Good Analytical Practice*, Angew. Chem. **59**, 5454 (2020).

[40] Y. Zhang, R. Esteban, R. A. Boto, M. Urbieta, X. Arrieta, C. Shan, S. Li, J. J. Baumberg, and J. Aizpurua, *Addressing Molecular Optomechanical Effects in Nanocavity-Enhanced Raman Scattering beyond the Single Plasmonic Mode*, Nanoscale **13**, 1938 (2021).

[41] C. G. T. Feugmo and V. Liégeois, *Analyzing the Vibrational Signatures of Thiophenol Adsorbed on Small Gold*

Clusters by DFT Calculations, Chem. Phys. Chem. **14**, 1633 (2013).

[42] A. T. Zayak, Y. S. Hu, H. Choo, J. Bokor, S. Cabrini, P. J. Schuck, and J. B. Neaton, *Chemical Raman Enhancement of Organic Adsorbates on Metal Surfaces*, Phys. Rev. Lett. **106**, 083003 (2011).

[43] S. Frey, V. Stadler, K. Heister, W. Eck, M. Zharnikov, M. Grunze, B. Zeysing, and A. Terfort, *Structure of Thioaromatic Self-Assembled Monolayers on Gold and Silver*, Langmuir **17**, 2408 (2001).

[44] H. Richter, M. Wienold, L. Schrottke, K. Biermann, H. T. Grahn, and H.-W. Hübers, *4.7-THz Local Oscillator for the Great Heterodyne Spectrometer on Sofia*, IEEE Trans. Terahertz Sci. Technol. **5**, 539 (2015).

[45] A. Thomas, L. Lethuillier-Karl, K. Nagarajan, R. M. Vergauwe, J. George, T. Chervy, A. Shalabney, E. Devaux, C. Genet, J. Moran *et al.*, *Tilting a Ground-State Reactivity Landscape by Vibrational Strong Coupling*, Science **363**, 615 (2019).

[46] Y. Zhang, J. Aizpurua, and R. Esteban, *Optomechanical Collective Effects in Surface-Enhanced Raman Scattering from Many Molecules*, ACS Photonics **7**, 1676 (2020).

[47] A. Xomalis, X. Zheng, tiR. Chikkaraddy, Z. Koczor-Benda, E. Miele, E. Rosta, G. A. E. Vandenbosch, A. Martnez, and J. J. Baumberg, *Detecting Mid-infrared Light by Molecular Frequency Upconversion with Dual-Wavelength Hybrid Nanoantennas*, arXiv:2107.02507.

[48] W. Chen, P. Roelli, H. Hu, S. Verlekar, S. P. Amirtharaj, A. I. Barreda, T. J. Kippenberg, M. Kovylina, E. Verhagen, A. Martnez, and C. Galland, *Continuous-Wave Frequency Upconversion with a Molecular Optomechanical Nanocavity*, arXiv:2107.03033.

[49] Z. Koczor-Benda, P. Roelli, C. Galland, and E. Rosta, *Molecular Vibration Explorer: An Online Database and Toolbox for Terahertz Detection, Surface-Enhanced Infrared and Raman Spectroscopy*, https://molecular-vibration-explorer.materialscloud.io/ (2021).